FRDC FINAL REPORT

DEVELOPING AN ANALYTICAL MODULE FOR LARGE-SCALE RECREATIONAL FISHERY DATA BASED ON PHONE-DIARY SURVEY METHODOLOGY

J.M. Lyle, S. Wotherspoon and K.E. Stark

May 2010

FRDC Project No. 2007/064

Australian Government
Fisheries Research and Development Corporation

tafi
Tasmanian Aquaculture and Fisheries Institute

ISBN 978-1-86295-566-0.

Developing an analytical module for large-scale recreational fishery data based on phone-diary survey methodology

Published by the Marine Research Laboratories – Tasmanian Aquaculture and Fisheries Institute, University of Tasmania, Private Bag 49, Hobart, Tasmania 7001.

E-mail: Jeremy.Lyle@utas.edu.au Ph. (03) 6227 7277 Fax: (03) 6227 8035

The Fisheries Research and Development Corporation plans, invests in and manages fisheries research and development throughout Australia. It is a statutory authority within the portfolio of the federal Minister for Agriculture, Fisheries and Forestry, jointly funded by the Australian Government and the fishing industry.

FRDC FINAL REPORT

DEVELOPING AN ANALYTICAL MODULE FOR LARGE-SCALE RECREATIONAL FISHERY DATA BASED ON PHONE-DIARY SURVEY METHODOLOGY

J.M. Lyle, S. Wotherspoon and K.E. Stark

May 2010

FRDC Project No. 2007/064

Tasmanian Aquaculture and Fisheries Institute

TABLE OF CONTENTS

2007/064 Developing an analytical module for large-scale recreational fishery data based on phone-diary survey methodology

PRINCIPAL INVESTIGATOR Dr Jeremy M. Lyle
ADDRESS University of Tasmania
Tasmanian Aquaculture and Fisheries Institute
Private Bag 49
Hobart TAS 7001
Telephone: 03 6227 7255 Fax: 03 6227 8035

OBJECTIVES

1 Review and document statistical procedures for analysing large-scale phone/diary recreational survey data

2 Develop an integrated and flexible data analysis module for phone-diary recreational survey data

3 Undertake a re-analysis of key NRIFS data outputs

4 Roll-out and demonstrate the analysis module to potential users

NON-TECHNICAL SUMMARY

The 2000-01 National Recreational Fishing Survey (NRFS) represented the first comprehensive assessment of recreational fishing in Australia, providing socio-demographic, fishing and fishing-related economic activity information for the resident population of Australia.

The NRFS employed a two-phase survey design, the first phase being a general population survey to establish participation rates and the second phase to collect detailed information on fishing and expenditure activity. The general population survey was conducted by telephone and the second phase was administered as a telephone-diary survey in which fishing and expenditure activity was monitored over a 12-month period. From a methodological perspective, exceptionally high response rates and a comprehensive approach to data quality and response bias issues are features that have been recognised internationally as a benchmark for large-scale, off-site surveys. However, the complexity of the NRFS design and sheer quantity of data collected meant that data analysis was not straightforward. The ability to independently replicate the original analyses and disaggregate data has been difficult and this has hindered further use of the NRFS dataset, especially given the need to incorporate statistical uncertainty with estimates.

While the likelihood of conducting a repeat of the national survey may be low, several jurisdictions have identified the need to provide state-wide and large-scale regional information about their recreational fisheries. Recognising the suitability of the NRFS

methodology to provide this information cost-effectively, it is critical that data analysis is undertaken efficiently and with appropriate statistical rigor.

Any survey represents an impost on respondents, and as such those with greater interest in the survey topic are more likely to co-operate. These differential response rates lead to non-response bias - the views or actions of those with an interest in the survey topic become over-represented in the sample and this results in bias. Auxiliary information can be used to reduce non-response bias by re-weighting the sample. In essence, respondents that are under-represented in the sample are weighted more heavily in the analysis, and those that are over-represented are weighted down. There are two major techniques for calculating weighting adjustment – response propensity modelling and calibration.

The original NRFS analysis used both response propensity modelling and calibration to reduce bias. Initial weights were calculated by calibrating to population census data and then a series of further response adjustments were performed to correct for perceived deficiencies in the sample. The primary disadvantage of such an approach is that a sequence of adjustments is performed, and successive adjustments may conflict.

Since the NRFS there have been several advances in the theory of calibration for multi-phase designs which have been applied here to develop a statistical package to analyse recreational survey data. The *RecSurvey* package has been implemented in the statistical computing language R and provides a flexible and transparent platform specifically designed for the phone-diary survey methodology. In addition to providing a step by step guide to analysis, with the capability for users to make decisions about what assumptions are applied in the calibration processes, the package documentation provides recommendations on database structure and queries necessary to prepare data for analysis.

Example analyses indicate the capability of the package to deliver disaggregated data outputs. For instance, catch and effort data can be readily disaggregated by fishing method, platform, region, target species, or combinations of these factors, and reported with associated uncertainty on the estimates.

Key NRFS data for Tasmania and South Australia were re-analysed using the *RecSurvey* package and compared with original estimates. Participation rates by region of residence and age group did not differ significantly between the original and re-analysed estimates. Furthermore, state-wide harvest estimates for the major species were not significantly different, indicating that the original analyses were generally robust. Detailed re-analysis of NRFS data for Tasmania was also reported to serve as an example of the type of outputs that can be achieved using the analytical package.

A workshop demonstration of the package was presented to researchers, managers and recreational stakeholders (October 2009) with several jurisdictions indicating interest in applying the package to re-analyse their NRFS data and for use in current and future recreational fishing surveys.

The key products arising from this project, namely the *RecSurvey* package (including functions and help files), an example database, worked data example and manual, will be available for download from the TAFI and FRDC websites.

OUTCOMES ACHIEVED

The primary outcome of this project is a foundation to support sustainable fisheries management through the inclusion of statistically robust information relating to the recreational sector.

The *RecSurvey* package represents a significant development in the analysis of complex survey data, providing an efficient, flexible and statistically robust tool that will benefit many jurisdictions as they seek to quantify and account of the impacts of recreational fishing activities.

KEYWORDS: Recreational fishing; large-scale off-site survey; data analysis; non-response adjustment; response propensity modelling; calibration procedures

ACKNOWLEDGEMENTS

We are especially indebted to Laurie West (Kewagama Research) who provided invaluable input in developing design specifications, consideration of non-response and unexpected fishing adjustments, and organising the ABS population benchmark data.

We also gratefully acknowledge the contributions of project steering committee - Rod Pearn (DPIPWE), Tony Burton (TARFish), Nick Crawford (ANSA), Brett Cleary (TARFish), Michael Hanek, and Ross Winstanley (Recfish Research) - in providing direction and encouragement for the project.

This study was funded by the Fisheries Research and Development Corporation (Project 2007/064) with contributions provided by the Tasmanian Department of Primary Industries, Parks and Environment and Primary Industries and Resources, South Australia.

BACKGROUND

The 2000-01 National Recreational and Indigenous Fishing Survey (NRIFS) yielded the first comprehensive assessment of non-commercial fishing in Australia (Henry and Lyle 2003). The NRIFS was comprised of several independent surveys, namely the National Recreational Fishing Survey (NRFS), the Indigenous Fishing Survey of Northern Australia and the Overseas Visitor Fishing Survey. The NRFS provided socio-demographic, fishing and fishing-related economic activity information for the Australian resident population.

Core components of the NRFS were a general population survey of households followed by a phone-diary survey in which fishing and expenditure activity was monitored over a 12-month period. From a methodological perspective, exceptionally high response rates and a comprehensive approach to data quality and response bias issues (Lyle *et al*. 2002) were features that have been recognised internationally as a benchmark for large-scale, off-site surveys (Pollock 2003). However, the complexity of the NRFS design and sheer quantity of data collected meant that data analysis was not straightforward and simple statistical approaches were not available. Weighting factors required to expand estimates to population totals were derived using a complex and step-wise process that took account of calibration against population benchmarks and various adjustments for non-response. The ability to independently replicate these analyses and undertake further disaggregation of the data has been difficult and this has hindered further use of the NRFS dataset, especially given the need to incorporate statistical uncertainty with estimates.

While the likelihood of conducting a repeat national survey may be low, several jurisdictions have identified the need to provide on-going state-wide and regional information about their recreational fisheries. State-wide surveys based on the NRFS design have been completed recently in Tasmania (Lyle *et al*., 2009) and South Australia (Jones, 2009); the Northern Territory implemented a survey during 2009 and Queensland is in the planning phase for a state-wide survey. Recognising that analyses need to be statistically robust, transparent and repeatable, the primary objective of the current project was to develop a flexible analytical framework that could be used to re-analyse existing NRFS data as well as providing an efficient statistical tool for use in the analysis of future recreational fishing surveys.

NEED

While the efficacy of the telephone-diary methodology in providing detailed and robust information about recreational fishing has been established, a need to further develop and refine the statistical tools necessary to do the analyses has remained. There is, therefore, a requirement to develop an analytical module that is robust, efficient and flexible, enabling further analysis to be conducted on existing and future datasets. The

development of such an analytical module specific to the NRFS methodology would represent a significant advancement in the provision of recreational data and will have immediate and on-going application in a number of jurisdictions as future surveys are completed. This project directly addresses Australian Fisheries Management Forum and national recreational R&D priorities relating to assessment of non-commercial fishery impacts.

OBJECTIVES

1 Review and document statistical procedures for analysing large-scale phone/diary recreational survey data
2 Develop an integrated and flexible data analysis module for phone/diary recreational survey data
3 Undertake a re-analysis of key NRIFS data outputs
4 Roll-out and demonstrate the analysis module to potential users

PART 1: ANALYTICAL MODULE

1. INTRODUCTION

In 2000-01 a national survey of recreational fishing in Australia was conducted (Henry and Lyle, 2003). The survey was a joint initiative of Commonwealth and State Governments, and the first comprehensive examination of the non-commercial components of Australian fisheries at the national level. A two-phase approach was used, involving a (telephone) screening survey of households to ascertain demographic and fishing characteristics, followed by a diary survey during which fishing and economic activity was monitored over a 12-month period. The National Recreational Fishing Survey (NRFS) provides invaluable baseline data about participation, catch and effort, and although such a large-scale national level survey may not be undertaken again, the need to provide on-going state-wide and regional information about recreational fisheries has been identified, and follow-up surveys have already been conducted or are planned in several states.

The complexity of the NRFS design and subsequent analysis highlighted the need to develop further, and more fully automate the process of data analysis, providing the motivation for this project. In particular, the need to simplify and integrate the analysis, and incorporate recent advances in the calibration of survey data (Särndal and Lundström, 2005; Estevao and Särndal, 2006) were identified. The analytical module presented here aims to ensure that a robust statistical approach is taken to the analysis of future recreational fishing survey data; that the methods are transparent and traceable, and the results repeatable.

1.1 Survey design

The analytical package, referred to here as *RecSurvey*, assumes that the design of the survey to be analysed mirrors that of the NRFS, with the optional inclusion of data (from call-backs or other sources) characterising non-responders and non-fishers.

More precisely, it is assumed that the survey is a two-phase design. An initial screening phase gathers background data from a general sample of the population. The data collected are then used to identify a more focused sub-sample which is then re-sampled in the second, more intensive phase of the survey, providing detailed catch and effort data.

The two phases may be conducted using any single stage cluster or element sampling design. In the NRFS, the first phase was conducted as stratified cluster samples in which households were the primary sampling unit and strata were defined geographically. While this structure is *not* mandatory, it will be used in this manual to illustrate the basic method of analysis. If the survey is not a two-phase design, the same

basic methods of analysis can be applied but the specialist features provided by this package are no longer required.

The package also allows for two optional stages of data collection aimed at bias reduction, i.e. non-response call-backs and non-intender call-backs.

Non-response call-backs are follow-up calls made to a sample of units that did not respond in the initial screening phase, and provide baseline data for any adjustment for non-response bias. It is recognised that due to legislative and ethical constraints, such data may not be available for future surveys and so provision has also been made for data from past surveys to be used in the analysis.

Non-intender call-backs are follow-up calls made to a sample of units that signalled no intention to fish during the second phase of the survey, and as such were deemed ineligible for the second phase. This provides baseline data on the uptake of fishing between the screening and intensive phase of sampling.

1.2 Non-response adjustment

Some level of non-response bias is inevitable in any survey. The *RecSurvey* package provides several mechanisms to reduce bias due to non-response.

Every survey is an impost, and as such those with greater interest in the survey topic are more likely to co-operate. These differential response rates lead to non-response bias - the views or actions of those with an interest in the survey topic become over-represented in the sample and this results in bias. In the recreational fishing context, more avid fishers may be more likely to cooperate than less avid fishers or non-fishers; the avid fishers would then become over-represented in the sample resulting in inflated estimates of catch and effort.

Auxiliary information can be used to reduce non-response bias by re-weighting the sample. In essence, respondents that are under-represented in the sample are weighted more heavily in the analysis, and those that are over-represented are weighted down. There are two major techniques for calculating weighting adjustment – response propensity modelling and calibration.

Response propensity modelling uses data characterising non-responders (from call-backs or partial responders) to construct a model for the probability that an arbitrary population member will respond. The weight adjustments are then chosen in inverse proportion to the predicted response probabilities of the sample members.
Calibration determines weights by forcing the survey estimates of key auxiliary quantities to exactly match known population totals. The rationale being that removing bias in estimates of known quantities will also reduce bias in other estimates. The more directly the calibrated quantities relate to the survey topic, the more valuable the calibration.

The *RecSurvey* package allows both response propensity modelling and calibration to be used to reduce bias. The original NRFS analysis also used both techniques. However, the methods implemented in this package are more robust than those used in the original NRFS analysis.

The original NRFS analysis used both response propensity modelling and calibration to reduce bias. Initial weights were calculated by calibrating to basic demographic data obtained from the ABS, and then a series of further response propensity adjustments were performed to correct for perceived deficiencies in the sample. The primary disadvantage such an approach is that a sequence of adjustments is performed, and successive adjustments may conflict. For example, although the initial weights calibrate the gender and age distribution of the sample to that of the Australian population, this may not be preserved by any subsequent adjustment.

Since the NRFS there have been a number of advances in the theory of calibration for multi-phase designs (Särndal and Lundström, 2005; Estevao and Särndal, 2006). The phases of a two-phase design can be viewed as distinct surveys. So not only can the first and second phases be calibrated to known population totals, but it is also possible to exploit information gained in the first phase to calibrate the second phase to estimates from the first. Moreover, all three forms of calibration can be performed as single integrated procedure, removing the possibility of conflict amongst competing adjustments.

1.3 Technical prerequisites

In this section we describe the prerequisites required to make use of the *RecSurvey* package.

1.3.1 R

The *RecSurvey* package is implemented in the statistical computing language R, and builds upon the *Survey* package developed by Thomas Lumley (Lumley, 2004, 2010) to provide facilities specifically tailored to the analysis of data from recreational fishing surveys conducted in the style of the NRFS.

R is a computing environment for statistical data analysis and graphics (R Development Core Team, 2008). R provides an enormous variety of built-in functions for statistical analysis, together with a high level scripting language that allows for almost unlimited extension of the base system. R is supported by a large developer community, and has hundreds of thousands of users world-wide, many of whom have contributed add-on packages to extend R's base functionality.

Both R and the *Survey* package are freely available and can be obtained from CRAN,

http://cran.r-project.org

a comprehensive archive of R resources.

While this manual aims to provide a tutorial in the use of the *RecSurvey* package, some familiarity with R will be essential to the effective use of the package. There are numerous texts that describe the use of R - in particular Dalgaard (2000) provides a basic introduction, while both Venables and Ripley (2002) and Crawley (2007) have a more statistical focus. Those intending to perform considerable manipulation of data prior to analysis may find Spector (2008) particularly relevant.

1.3.2 Data management

Any survey of the magnitude of the NRFS requires some form of database for data storage, and the precise structure of this database will be determined by the fine detail of the survey. The *RecSurvey* package itself takes no responsibility for management of the survey data. It is assumed that the user is capable of providing data in the format required for analysis.

R is capable of interfacing to a wide range of external databases (Lapsley and Ripley, 2008), and so rather than be prescriptive as to the form of data storage, the package assumes that the user can retrieve the data in the format required for analysis. Section 5 presents an example database structure and the associated queries necessary for exporting data to R for a survey similar in design to the NRFS.

1.3.3 Sampling concepts

This manual assumes the reader is familiar with the basic theory of survey sampling. In particular, it is assumed the reader is familiar with the following concepts: random, cluster and stratified sampling, sampling weights and finite population correction, non-response and non-response bias, call-backs, two-phase sampling, response adjustment and calibration. Lohr (1999) provides a general introduction to the theory of survey sampling, while Särndal and Lundström (2005) focuses on the problems of non-response. Two-phase sampling, non-response adjustment and calibration are of particular relevance.

2 PHASE ONE ANALYSIS

The screening phase of the survey may be considered a survey in its own right, and typically the data from the screening will be available significantly earlier than data from the second more intensive phase of data collection. In these circumstances it can be of considerable value to obtain early results from the screening data rather than wait for the completion of the full survey.

The screening data can be analysed using the tools provided by the *Survey* package (Lumley, 2004, 2010), and so any documentation or tutorials for that package are particularly relevant to this analysis.

The basic analysis consists of four main steps:

1. Import the relevant data from the survey database.
2. Describe the sampling design.
3. Optionally, calibrate to known population totals.
4. Compute population estimates.

Where data is available, a response propensity model may also be applied as an initial adjustment for non-response bias. This would be employed between steps 1 and 2, and is discussed in detail in Section 4.

2.1 Data requirements

The analysis of the screening phase requires the survey data from the screening phase, together with one or more tables of benchmark data which are used for determining initial sampling weights and for any calibration.

The user must provide the data in the form of a single dataframe, where each row corresponds to a sampling element. The first phase of the NRFS sampled individuals aged five years or older clustered by household, and so for the example dataframe each row corresponds to an individual aged five years or older.

Each column of the dataframe corresponds to a response or an element descriptor, and should include:

Cluster identifier: For clustered designs, an identifier that uniquely allocates elements to clusters. The NRFS was a sample of individuals clustered by household, so the household identifier is the cluster identifier.

Stratum identifier: For stratified designs, an identifier that uniquely allocates elements to strata. The NRFS was stratified by geographic district, so the district identifier is the stratum identifier.

Weights and Population sizes: To fully specify the sampling design, it is necessary to specify sampling weights or sampling fractions. For samples with large sampling fractions, computation of finite population corrections will also require (stratum) population sizes.

Auxiliary responses: Responses used to classify elements. For the NRFS, an individual's age, gender, region of residence, and the composition of their household (number of residents) are auxiliary responses.

Responses: Quantities about which inferences are to be made. For the screening phase of the NRFS, these include whether an individual fished in the preceding 12 months, whether they intended to fish during the second phase of the survey, and whether the household owned a boat.

Benchmark data are known population totals from external sources (e.g. census data) and are used to calculate sampling weights, the finite population correction, and for calibrating against to reduce any non-response bias. The minimum requirement is for population totals at the sampling unit level (i.e. households in the NRFS) by stratum (for stratified designs such as the NRFS). For use in calibration, the population totals need to be broken down by some other auxiliary variable such as Household Type. Benchmark data can also be provided at the sampling element level for calibration, e.g. total number of persons in population by sex and/or age class. Any benchmark dataframe should therefore include columns for stratum identifier, one or more auxiliary responses, and the known population total.

In general, categorical responses should be represented as factors in R, but binary indicator responses (Yes/No, True/False) can be represented as zero/one numeric data (e.g. Yes=1, No=0).

Section 5 presents an example database design for a survey similar in structure to the NRFS, together with sample queries that extract the data required by the various analyses. This design will be used throughout the remainder of the current section to illustrate the basic steps of the analysis.

2.1.1 RODBC

The analysis of the screening phase requires the data from the screening phase, together with any tables of benchmark data which are needed in the calibration step and for determining initial sampling weights. Typically these tables will be imported into R though an ODBC connection to the Survey database.

As a simple example, the following query extracts a subset of the screening data available in the example database (a more detailed example is given in Section 5)

```
> qry <- "SELECT
+   HouseholdID, PersonID, Stratum, HType,
+   Sex, AgeGp, Age, HFishedL12M, NPersons,
+   IIf(UsablePersons.PFishedStateL12M Is Not Null,
+     UsablePersons.PFishedStateL12M,
+     HFishedL12M) AS PFishedStateL12M,
+   IIf(UsablePersons.DaysFishedL12M Is Not Null,
+     UsablePersons.DaysFishedL12M,'0') AS DaysFishedL12M
+ FROM UsablePersons
+ WHERE AgeGp<>'[0,5)';"
```

This query can be executed directly from R through an ODBC connection with the RODBC library to extract the screening data. Benchmark data can be extracted at the same time

```
> library(RODBC)
> ch <- odbcConnectAccess("Example.mdb")
> d.screen <- sqlQuery(ch, qry)
> d.house <- sqlQuery(ch, "SELECT * FROM HouseholdBenchmarks;")
> d.person <- sqlQuery(ch, "SELECT * FROM PersonBenchmarks WHERE
AgeGp<>'[0,5)';")
> odbcClose(ch)
```

This executes the three queries for extracting the screening data, the household benchmark data and the person benchmark data, and stores the results as dataframes. The precise syntax for the ODBC connection will depend upon the database type and configuration.

The screening data takes the form

```
> head(d.screen)
```

```
HouseholdID PersonID Stratum HType Sex AgeGp Age HFishedL12M
1      H10001 H10001P798      42   5PP   F [45,60)  46    Y
2      H10001 H10001P799      42   5PP   M [45,60)  47    Y
3      H10001 H10001P800      42   5PP   M  [5,15)  13    Y
4      H10001 H10001P801      42   5PP   M  [5,15)  11    Y
5      H10001 H10001P802      42   5PP   F [15,30)  16    Y
6      H10002 H10002P913      42    3P   M [15,30)  25    Y

NPersons PFishedStateL12M DaysFishedL12M
1        5                Y          [1,5)
2        5                Y          [1,5)
3        5                Y          [1,5)
4        5                Y          [1,5)
5        5                Y          [1,5)
6        3                Y            20+
```

The RODBC library will automatically recognise any textual columns as categorical and convert them to factors upon import, but any columns of categorical data that are represented with numeric labels must be converted manually. For example, in the example database the stratum identifiers are all numeric and will not be recognized as categorical labels. They need to be converted within R

```
> d.screen$Stratum <- factor(d.screen$Stratum)
> d.house$Stratum <- factor(d.house$Stratum)
> d.person$Stratum <- factor(d.person$Stratum)
```

2.1.2 Sampling weights

The sampling weights and stratum totals for finite population correction must be calculated using the benchmark data and then added to the basic screening data. If population estimates are required for quantities recorded at the cluster level, it may also be necessary to construct additional weighted variables (see below Section 2.1.3).

The sampling weights are the reciprocals of the sampling fractions. The example survey is a stratified cluster sample of households so the weight for a particular stratum is the total number of households within that stratum divided by the number of households sampled within that stratum. For the finite population correction R requires the total number of households within each stratum. The number of households sampled can be determined by counting the number of unique household identifiers in each stratum, and the total number of households is determined by aggregating the household benchmark data over household type

```
> (hsampled <- tapply(d.screen$HouseholdID, d.screen$Stratum,
+     function(x) length(unique(x))))


 42  43  44  45
898 249 753 599


> (htotal <- tapply(d.house$N, d.house$Stratum, sum))


   42    43    44    45
82904 14414 55913 44066
```

The sampling weights and stratum totals for finite population correction must be added to the screening data so that each individual is allocated the weight and total for the stratum from which they are drawn

```
> d.screen$weight <-
htotal[d.screen$Stratum]/hsampled[d.screen$Stratum]
> d.screen$fpc <- htotal[d.screen$Stratum]
```

Alternatively, the sampling weights and finite population correction may be calculated within the database, and included in the table of screening data read in using the ODBC connection.

2.1.3 Cluster level estimates

Care must be taken when estimating population totals for quantities recorded at the cluster level. In the example data, the `HFishedL12M` response records whether an individual resides in a household that fished in the previous year. Estimating population totals for this response directly will estimate the number of *people* that live in households that fished in the previous year. To estimate the number of *households* that fished in the previous year a weighted variable must be constructed that allocates the fishing to a single member of the household

```
> d.screen$W <- ifelse(!duplicated(d.screen$HouseholdID),
+     1, 0)
> d.screen$HFishedL12MW <- ifelse(d.screen$HFishedL12M ==
+     "Y", d.screen$W, 0)
```

2.2 Specifying the design

After reading in the screening survey and benchmark data, the second step is to load the *Survey* package and describe the sampling design.

The sampling design is specified with the `svydesign` function. This creates a survey design object that contains the survey data together with the information required to define the sampling design. In particular, the user must specify:

`ids`: a formula specifying the cluster identifiers or `0` for non-clustered designs.

`strata`: a formula specifying stratum identifiers, or `NULL` for un-stratified designs.

`weight`: a formula specifying the sampling weights

`fpc`: a formula specifying the stratum population totals. This argument is optional and may be omitted if the sampling fractions are sufficiently small that any finite population correction is negligible.

`data`: a data frame containing the survey responses, any relevant auxiliary responses, and the aforementioned cluster and strata identifiers, sampling weights and population sizes.

For the example database, the design is specified as

```
> library(survey)
> s <- svydesign(ids=~HouseholdID,
+                strata=~Stratum,
+                weight=~weight,
+                fpc=~fpc,
+                data=d.screen)
```

2.3 Estimation

If no calibration is performed, the final stage of the process is estimation. Calibration will be discussed in detail in the next section (Section 2.4)

The functions `svymean`, `svytotal` and `svytable` are used to compute estimates from the survey design object. For numeric variables, the `svytotal` and `svymean` calculate population totals and means, but for categorical variables these functions compute total counts and proportions for each category. The `svytable` can be used to produce complex cross-tabulations.

The screening data in the example database records no numeric variables of any real interest. But for the sake of example, the mean number of persons per household can be estimated as

```
> svymean(~NPersons, s)


           mean     SE
NPersons 3.1646 0.0408
```

The mean number of persons per household is estimated to be 3.16, with a standard error of 0.04.

Similarly, participation in the State fishery can be estimated by computing the proportion of individuals that have fished within the state in the preceding twelve months

```
> svymean(~PFishedStateL12M, s)


                      mean     SE
PFishedStateL12MN 0.76026 0.0087
PFishedStateL12MY 0.23974 0.0087
```

The participation rate amongst the population aged five years or older is estimated as 23.97% and correspondingly the non-participation rate as 76.03%, both with standard error 0.87%. Note that as `PFishedStateL12M` is categorical, R computes the proportions for every level of the factor, and so computes both the participation and non-participation rates. Total numbers of fishers and non-fishers can be determined with `svytotal`

```
> svytotal(~PFishedStateL12M, s)


                    total     SE
PFishedStateL12MN 357568 5139.7
PFishedStateL12MY 112758 4387.2
```

The `svytable` function provides the same results but without standard errors

```
> svytable(~PFishedStateL12M, s)


PFishedStateL12M
       N        Y
357567.7 112757.8
```

More detailed information on the level of participation can be determined from `DaysFishedL12M`

```
> svymean(~DaysFishedL12M, s)


                           mean     SE
DaysFishedL12M[1,5)   0.079628 0.0052
DaysFishedL12M[10,15) 0.042381 0.0038
DaysFishedL12M[15,20) 0.021989 0.0027
DaysFishedL12M[5,10)  0.049616 0.0040
DaysFishedL12M0       0.756555 0.0087
DaysFishedL12M20+     0.049831 0.0039
```

Note that the proportion of the population that fished zero days in the last 12 months (0.7566) is less than the proportion that didn't fish (0.7603), this is because in the NRFS `DaysFishedL12M` refered to fishing in *any* State, while `PfishedStateL12M` is only for fishing in the State of interest.

To compute participation by gender, `svytable` can be used to calculate totals cross-tabulated by participation and gender

```
> tab <- svytable(~Sex + PFishedStateL12M, s)
> tab


   PFishedStateL12M
Sex         N         Y
  F 198790.36  37233.25
  M 158777.32  75524.52
```

In turn these results can be used to compute participation rates separately for the two genders,

```
> tab/rowSums(tab)


   PFishedStateL12M
Sex         N         Y
  F 0.8422478 0.1577522
  M 0.6776614 0.3223386
```

or for the contribution of each gender to overall participation

```
> tab/sum(rowSums(tab))


   PFishedStateL12M
Sex          N          Y
  F 0.42266554 0.07916487
  M 0.33759032 0.16057927
```

Similar results can be obtained by computing the mean for the interaction of `Sex` with `PFishedStateL12M`, providing standard errors as well

```
> svymean(~I(Sex:PFishedStateL12M), s)

                                   mean     SE
I(Sex:PFishedStateL12M)F:N 0.422666 0.0057
I(Sex:PFishedStateL12M)F:Y 0.079165 0.0044
I(Sex:PFishedStateL12M)M:N 0.337590 0.0062
I(Sex:PFishedStateL12M)M:Y 0.160579 0.0057
```

Estimates for subgroups can also be obtained with the `svyby` function. This function requires a formula that specifies what is to be calculated, a second formula that defines the subgroups, and the function used to compute the estimates within the subgroups. So to compute participation rates separately for the two genders

```
> svyby(~PFishedStateL12M, ~Sex, s, svymean)


  Sex PFishedStateL12MN PFishedStateL12MY se.PFishedStateL12MN
F   F         0.8422478         0.1577522          0.008645912
M   M         0.6776614         0.3223386          0.011039201
  se.PFishedStateL12MY
F          0.008645912
M          0.011039201
```

Using these basic techniques arbitrarily complex cross-tabulations can be constructed. To compute participation within gender and stratum,

```
> tab <- svytable(~Stratum + Sex + PFishedStateL12M, s)
> sweep(tab, 1:2, apply(tab, 1:2, sum), "/")


, , PFishedStateL12M = N

       Sex
Stratum         F         M
     42 0.8378871 0.6858491
     43 0.7580071 0.6038961
     44 0.8511838 0.6722783
     45 0.8655462 0.6942496

, , PFishedStateL12M = Y

       Sex
Stratum         F         M
     42 0.1621129 0.3141509
     43 0.2419929 0.3961039
     44 0.1488162 0.3277217
     45 0.1344538 0.3057504
```

or alternatively we can use `svyby` to obtain standard errors around the estimates

```
> svyby(~PFishedStateL12M, ~Stratum + Sex, s, svymean)


     Stratum Sex PFishedStateL12MN PFishedStateL12MY
42.F      42   F         0.8378871         0.1621129
43.F      43   F         0.7580071         0.2419929
44.F      44   F         0.8511838         0.1488162
45.F      45   F         0.8655462         0.1344538
42.M      42   M         0.6858491         0.3141509
43.M      43   M         0.6038961         0.3961039
44.M      44   M         0.6722783         0.3277217
45.M      45   M         0.6942496         0.3057504
     se.PFishedStateL12MN se.PFishedStateL12MY
42.F           0.01466354           0.01466354
43.F           0.03176721           0.03176721
44.F           0.01470933           0.01470933
45.F           0.01649165           0.01649165
42.M           0.01847891           0.01847891
43.M           0.03502743           0.03502743
44.M           0.01963670           0.01963670
45.M           0.02209024           0.02209024
```

As noted previously, considerable care must be taken when estimating cluster level quantities. Computing

```
> svytotal(~HFishedL12M, s)


               total     SE
HFishedL12MN 311064 5638.4
HFishedL12MY 159262 5659.5
```

does not provide an estimate of the number of *households* that fished in the previous year, but rather the number of *individuals* aged five years or older that reside in an household that fished in the previous year. To estimate the number of households that fished in the previous year the weighted variable `HFishedL12MW` constructed in Section 2.1.3 must be used

```
> svytotal(~HFishedL12MW, s)


              total     SE
HFishedL12MW 56768 1788.2
```

But the fraction of households that fished in the previous year cannot be calculated with

```
> svymean(~HFishedL12MW, s)


               mean     SE
HFishedL12MW 0.1207 0.0037
```

as this calculates the number of households that fished as a fraction of the total number of *individuals* in the population. To correctly compute the fraction of households that fished, it is necessary to first compute the number of households,

```
> sum(d.house$N)


[1] 197297
```

and scale both the estimated number of fishing households and its standard error by this total. To do this it is necessary to convert the results returned by `svytotal` to a dataframe

```
> as.data.frame(svytotal(~HFishedL12MW, s))/sum(d.house$N)


                  total          SE
HFishedL12MW 0.2877289 0.009063293
```

2.4 Calibration

Calibration forces the survey estimates of key auxiliary quantities to exactly match known population totals. The rationale being that removing bias in estimates of known quantities will also reduce bias in other estimates.

Given the auxiliary data present in the sample, it is possible to estimate many demographic quantities unrelated to the purpose of the survey. If the true population values for these quantities are known, these values can be used to calibrate the survey. Calibration is the process of adjusting the sampling weights so that survey estimates exactly reproduce known values of chosen auxiliary quantities. Population totals used for calibration may be at the sampling unit level (e.g. household) and/or the sample element level (e.g. person).

Calibration is performed with the `calibrate` function. The user must supply a survey design object to calibrate, a formula specifying the variables to calibrate, and a vector of the known population totals for the quantities to be calibrated (in the benchmark dataframes, see Section 2.1). For cluster samples, the `aggregate.stage` argument should be set to force the weight adjustments to be constant within clusters.

The specification of the known population values for the calibration is not straight forward. The `calibrate` function generates the design matrix corresponding to a calibration formula, and then adjusts the sampling weights so that the weighted column sums of this design matrix match the given values. In general, *the values required for calibration are the column sums that would be obtained from the design matrix constructed for the entire population*. To use `calibrate` in its most general form requires a detailed understanding of the construction of design matrices in R.

2.4.1 Calibrating to element totals

In the simplest case the values to be calibrated to are all quantities recorded at the element level (i.e. person level in NRFS example).

If the responses used in the calibration are all categorical then the calibration is equivalent to a complex form of post-stratification. In this case the appropriate population totals are easily constructed with the `model.matrix` function.

In the example database, the PersonsBenchmarks table records the total number of individuals of each gender in six age groups across four geographic strata

```
> head(d.person)


  Stratum Sex    AgeGp     N
1      42   F  [15,30) 20679
2      42   F  [30,45) 21254
3      42   F  [45,60) 22287
4      42   F   [5,15) 12856
5      42   F [60,100) 20870
6      42   M  [15,30) 20652
```

Calibrating to this data is equivalent to post-stratifying the design by stratum, gender and age group.

The population values required for the calibration can be computed with the `model.matrix` function. This function is used to construct a design matrix from the person benchmark data; the matrix is weighted by the category totals and the required totals are the column sums of this weighted matrix.

```
> totals <- colSums(d.person$N * model.matrix(~Stratum *
+     Sex * AgeGp, data = d.person))
> s1 <- calibrate(s, ~Stratum * Sex * AgeGp, totals, aggregate.stage =
1)
```

Setting `aggregate.stage=1` forces the adjustments to the sampling weights to be consistent with the first stage of clustering, in this case, within households.

After calibration the survey estimates for the calibrated quantities should match given values exactly, and this provides a simple check of the calibration. In this case, estimating the number of individuals of each gender within each age group and stratum should reproduce the benchmark data shown above (albeit in a different order)

```
> svyby(~Sex, ~Stratum + AgeGp, s1, svytotal)


          Stratum   AgeGp     SexF   SexM    se.SexF     se.SexM
42.[15,30)      42  [15,30) 20679 20652 1.494611e-12 1.094475e-12
43.[15,30)      43  [15,30)  2468  2683 4.122188e-13 2.133839e-13
44.[15,30)      44  [15,30) 12725 13022 8.390421e-13 4.671811e-13
45.[15,30)      45  [15,30)  9376  9440 8.085818e-13 6.047329e-13
42.[30,45)      42  [30,45) 21254 19680 8.453559e-13 5.210471e-13
43.[30,45)      43  [30,45)  3543  3487 1.736155e-13 1.625710e-13
44.[30,45)      44  [30,45) 13777 13252 3.399934e-13 2.102969e-13
45.[30,45)      45  [30,45) 10827 10471 3.517026e-13 2.470235e-13
42.[45,60)      42  [45,60) 22287 20857 9.296858e-13 6.301673e-13
43.[45,60)      43  [45,60)  4344  4611 1.911388e-13 1.647807e-13
44.[45,60)      44  [45,60) 14699 14591 3.392446e-13 1.989743e-13
45.[45,60)      45  [45,60) 11658 11535 4.295449e-13 2.459468e-13
42.[5,15)       42   [5,15) 12856 13365 7.051750e-13 7.249259e-13
43.[5,15)       43   [5,15)  2401  2540 2.072784e-13 1.280595e-13
44.[5,15)       44   [5,15)  9168  9866 4.080319e-13 2.922118e-13
45.[5,15)       45   [5,15)  7430  7954 3.594718e-13 2.914994e-13
42.[60,100)     42 [60,100) 20870 17622 8.779996e-13 4.621355e-13
43.[60,100)     43 [60,100)  3397  4040 2.079219e-13 1.592201e-13
44.[60,100)     44 [60,100) 14686 13085 4.860701e-13 6.049535e-13
45.[60,100)     45 [60,100) 11730 11015 3.639654e-13 3.747035e-13
```

Note that the standard errors are (effectively) zero, reflecting the fact that the estimates for calibrated quantities are exact.

Once the design is calibrated, estimation proceeds as before

```
> svymean(~PFishedStateL12M, s1)


                      mean     SE
PFishedStateL12MN 0.75075 0.0091
PFishedStateL12MY 0.24925 0.0091


> svymean(~DaysFishedL12M, s1)


                           mean     SE
DaysFishedL12M[1,5)    0.082075 0.0055
DaysFishedL12M[10,15) 0.044867 0.0041
DaysFishedL12M[15,20) 0.022692 0.0028
DaysFishedL12M[5,10)   0.051148 0.0041
DaysFishedL12M0        0.747160 0.0091
DaysFishedL12M20+      0.052058 0.0043
```

showing that in this case, the estimate of participation rate is slightly increased by the calibration.

The calibration formula determines the quantities to which the survey is calibrated. This provides considerable flexibility. Specifying the formula `~Stratum+Sex+AgeGp` calibrates to the total number of individuals within in each stratum, the total number of individuals within each gender and the total number of individuals within each age group. Specifying `~Stratum+Sex*AgeGp` calibrates to the total number of individuals in each stratum, and the total number of individuals in each gender/age group combination. Similarly the formula `~Stratum*Sex+AgeGp` calibrates the total number of individuals in each stratum/gender combination, and each age group, while the formula `~Stratum*Sex*AgeGp` calibrates to the total number of individuals in each stratum/gender/age group combination.

2.4.2 Calibrating to element *and* cluster totals

The survey may also be calibrated against known values for quantities recorded at the cluster level; this was effectively the basis for the integrated weighting procedure used in the original NRFS analysis (Henry and Lyle, 2003). However, the same caution must be exercised as when estimating quantities recorded at the cluster level (Section 2.3).

In the example database the HouseholdBenchmarks table records the total number of households by stratum and household type - a categorical variable based on the number of people in each household.

Unfortunately, determining the appropriate calibration formula and values for a mixture of person and household level quantities is considerably more difficult. The task is complicated by both the need to construct weighted variables for household level quantities and the fact that by default R is free to re-order the terms of a formula. Because of the need to weight household level quantities, it is not easy to calibrate to household type directly. The simplest approach is to create indicator variables that correspond to the dummy variables R would create in constructing a design matrix

```
> d.screen$HType1P <- ifelse(d.screen$HType=="1P",d.screen$W,0)
> d.screen$HType2P <- ifelse(d.screen$HType=="2P",d.screen$W,0)
> d.screen$HType3P <- ifelse(d.screen$HType=="3P",d.screen$W,0)
> d.screen$HType4P <- ifelse(d.screen$HType=="4P",d.screen$W,0)
> d.screen$HType5PP <- ifelse(d.screen$HType=="5PP",d.screen$W,0)
> s <- svydesign(ids=~HouseholdID,
+                strata=~Stratum,
+                weight=~weight,
+                fpc=~fpc,
+                data=d.screen)
```

The person level values are calculated as before, except that it is necessary to suppress the reordering of terms in the formula

```
> p.formula <- terms(~Stratum * Sex * AgeGp, keep.order = T)
> p.totals <- colSums(d.person$N * model.matrix(p.formula,
+     data = d.person))
```

To calibrate to the household composition within each stratum, the totals are calculated from the interaction of stratum and household type, but for cluster level quantities it is necessary to omit any intercept term

```
> h.formula <- terms(~-1 + Stratum:HType, keep.order = T)
> h.totals <- colSums(d.house$N * model.matrix(h.formula,
+     d.house))
> h.totals

Stratum42:HType1P  Stratum43:HType1P  Stratum44:HType1P
            22631               3474              14873
 Stratum45:HType1P  Stratum42:HType2P  Stratum43:HType2P
            11380              28376               5537
 Stratum44:HType2P  Stratum45:HType2P  Stratum42:HType3P
            20026              16086              13367
 Stratum43:HType3P  Stratum44:HType3P  Stratum45:HType3P
             2130               8610               6743
 Stratum42:HType4P  Stratum43:HType4P  Stratum44:HType4P
            11856               1913               7592
 Stratum45:HType4P Stratum42:HType5PP Stratum43:HType5PP
             6019               6674               1360
Stratum44:HType5PP Stratum45:HType5PP
             4812               3838
```

The calibration formula is the concatenation of the person level formula, and the household level formula re-expressed in terms of the weighted indicator variables

```
> formula <- terms(~Stratum * Sex * AgeGp + Stratum:(HType1P +
+     HType2P + HType3P + HType4P + HType5PP), keep.order = T)
> s1 <- calibrate(s, formula, c(p.totals, h.totals), aggregate.stage =
1)
```

The survey estimates now reproduce the person level benchmarks, *and* the household compositions within each stratum

```
> svyby(~HType1P + HType2P + HType3P + HType4P + HType5PP,
+     ~Stratum, s1, svytotal)


   Stratum HType1P HType2P HType3P HType4P HType5PP   se.HType1P
42      42   22631   28376   13367   11856     6674 3.773999e-13
43      43    3474    5537    2130    1913     1360 5.916314e-14
44      44   14873   20026    8610    7592     4812 2.619444e-13
45      45   11380   16086    6743    6019     3838 1.809818e-13
     se.HType2P   se.HType3P   se.HType4P  se.HType5PP
42 9.043168e-13 5.220820e-13 4.833667e-13 6.466153e-13
43 1.729613e-13 1.900163e-13 1.459183e-13 1.857109e-13
44 4.927171e-13 4.103533e-13 2.288228e-13 2.411389e-13
45 5.238984e-13 3.016192e-13 3.433039e-13 3.219435e-13
```

Note that it is necessary to calculate with the weighted indicator variables to obtain the number of households of each composition.

Following the calibration estimation proceeds as for the uncalibrated case (Section 2.3)

```
> svymean(~PFishedStateL12M, s1)


                      mean     SE
PFishedStateL12MN 0.75343 0.0087
PFishedStateL12MY 0.24657 0.0087


> svymean(~DaysFishedL12M, s1)


                           mean     SE
DaysFishedL12M[1,5)    0.080418 0.0052
DaysFishedL12M[10,15) 0.043564 0.0039
DaysFishedL12M[15,20) 0.023266 0.0028
DaysFishedL12M[5,10)   0.051171 0.0042
DaysFishedL12M0        0.749622 0.0088
DaysFishedL12M20+      0.051959 0.0042
```

2.4.3 Assessing weights

Calibration works by making adjustments to the sampling weights such that units that appear over-represented in the sample are weighted down, and those that appear under-represented are weighted up. This re-weighting is achieved by forcing the survey estimates to reproduce exactly certain known population characteristics.

Ideally the weight adjustments are small, but it is possible for the weights of some units to grow to the point where those units dominate the sample and exert an unhealthy influence on survey estimates.

The re-weighting selected by the calibration can be assessed by plotting the calibrated weights against the initial sampling weights. Figure 2.1 shows boxplots by strata of the ratio of the adjusted weights and original weights for the screening data calibrated to both person and household level benchmarks.

```
> plot(weights(s1)/weight ~ Stratum, data = d.screen, ylab =
"Adjustment")
```

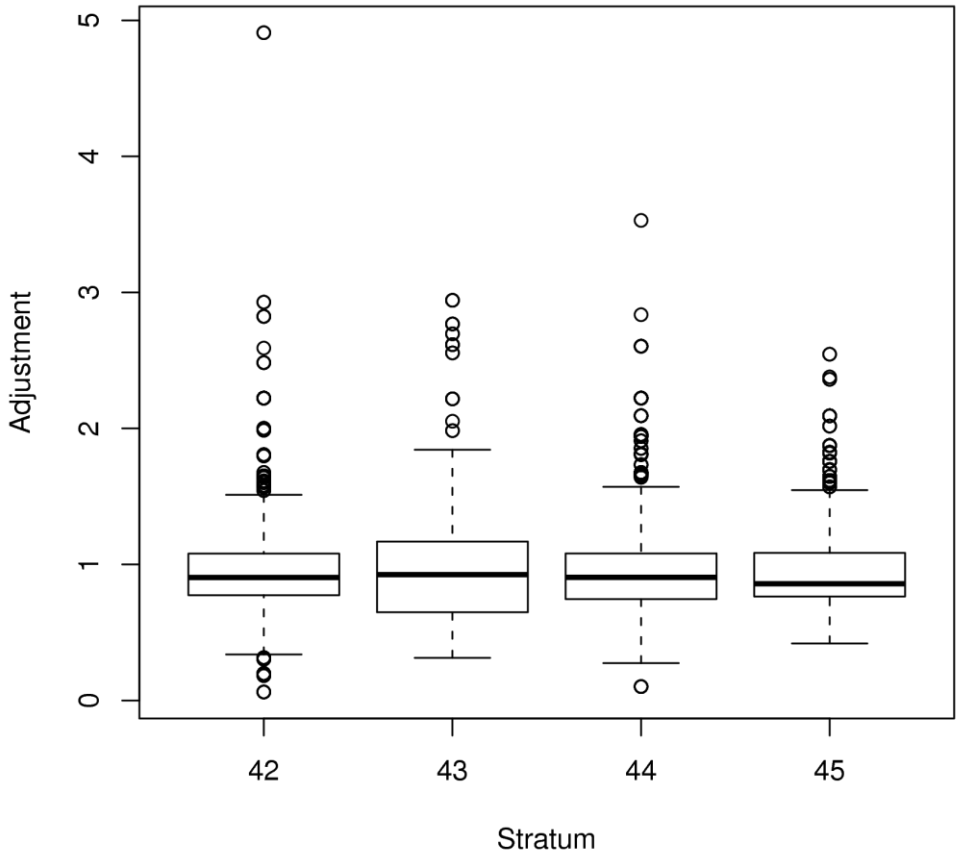

Fig. 2.1: Boxplots by strata of the ratio of the calibrated weights to the initial sampling weights for the screening data.

The re-weighting process may result in some weights becoming negative, or extremely large. While in general they are not necessarily a problem, they are clearly undesirable. Extreme weights can be prevented by placing *bounds* on the ratio of the adjusted and original weights. The `calibrate` function allows this through the optional `bounds` parameter – so for example, to bound the ratios of the adjusted (calibrated) and original weights to a specified interval, e.g. [0.1,20]

```
> s1 <- calibrate(s, formula, c(p.totals, h.totals), aggregate.stage =
1,
+     bounds = c(0.1, 20))
```

However, if the imposed bounds are too restrictive it may not be possible to achieve calibration

```
> try(calibrate(s, formula, c(p.totals, h.totals), aggregate.stage =
1,
bounds = c(0.9, 1.2)))


Error in calibrate.survey.design2(s, formula, c(p.totals, h.totals), :
  Calibration failed
In addition: Warning message:
In grake(mm, ww, calfun, bounds = bounds, population = population,  :
  Failed to converge: eps=0.266734083524727 in 51 iterations
```

3 PHASE TWO ANALYSIS

The second phase of the survey focuses on the primary intent of the survey – estimating recreational catch and effort.

Following the screening, those sampling units that indicate an intention to fish are invited to join the second phase of the survey. No further data are collected from sampling units that do not intend to fish in the duration of the second phase. Instead, it is assumed that those that indicate they do not intend to fish will actually not fish. While this approach is extremely efficient as sampling effort is only expended on those likely to fish, several difficulties arise.

Firstly, this second phase of sampling introduces further potential for non-response bias. Sampling units must agree to join the second phase sample, but more avid fishers may be more likely to agree and consequently become over-represented in the sample, resulting in bias.

Selecting the second phase sample based on intention to fish introduces another potential bias. Of those units that intend to fish, some fraction will not actually fish in the period of the survey, and conversely, of those that do not intend to fish, some fraction may fish during the survey period. Fishing by those not intending to fish represents an 'influx' of effort, and the failure to fish by intending fishers an 'outflux'. Outflux is represented in the sample, but any influx goes undetected as no further data is collected from non-intending fishers in the second phase. This results in bias, as a potentially large component of fishing effort is never sampled, resulting in a tendency to under-estimate the true catch and effort.

Both forms of bias may be adjusted for in the analysis. Non-response bias in both the screening phase and the second phase samples may be corrected through calibration. For bias resulting from an influx or outflux, there are several possible approaches:

1. **Make no adjustment:** The data are analysed directly, as is, and no adjustment is made for the potential bias. Although potentially biased, this approach does provide a meaningful lower bound – the true catch and effort are likely to be at least as large as the estimates obtained in this way.
2. **Omit fishers that fail to fish:** Analyse the data with those units that intended to fish but did not omitted from the second phase sample. If the population is in equilibrium in the sense that influx is exactly balanced by the outflux, deleting the outflux from the second phase sample will result in unbiased estimates. However there seems to be little basis for expecting influx to balance outflux.
3. **Calibrate to estimated measures of participation.** Adjust for any imbalance between influx and outflux by calibrating to external estimates of participation. Influx and outflux effectively distort participation – calibrating to participation upweights those in the sample that actually fish. Participation can be estimated either by making call-backs to the non-intending units from the screening sample, or by assuming the participation rate is constant over the entirety of the survey,

and using the participation estimates from the screening phase as estimates of participation during the second phase.

For options 1 and 2, it may still be necessary to calibrate to adjust for non-response bias, but in option 3, calibration is used to adjust for both non-response bias and the imbalance between influx and outflux. Which of these methods is most appropriate will be determined by the levels of influx and outflux and reliability of external estimates of participation.

3.1 Two-phase analysis

First consider the simpler form of two- phase analysis where no calibration is performed. This forms the basis for the more complex calibrated analyses to be considered in Section 3.3.

3.1.1 Data requirements

The two-phase analysis has the same basic data requirements as the screening. The user must provide the data in the form of a single dataframe, where the rows correspond to the elements sampled in the first phase sample. In addition to the identifiers, weights and stratum sizes required for the screening analysis, the user must also include the catch and effort data required by the analysis, and a column that specifies which elements were also sampled in the second phase of sampling.

The relevant catch and effort data is determined by the responses to the second phase sample. Respondents can be divided into three groups, with catch and effort determined as follows:

- individuals from households that did not intend to fish are assumed to have not fished and are assumed to have zero catch and effort;
- individuals from households that intended to fish and participated in the second phase have a catch and effort determined by the second phase sample;
- individuals from households that intended to fish but did not participate in the second phase are assumed to have unknown catch and effort (coded as `NA`).

The data required to specify and calibrate the design is very similar to the data required for the screening analysis. In the following example the query described in Section 5.3.5 has been stored as view named ‘Phase2’ and provides all required variables described above, except for the second phase responses (i.e. catch and/or effort)

```
> library(RODBC)
> ch <- odbcConnectAccess("Example.mdb")
> d.phase2 <- sqlQuery(ch, "SELECT * FROM Phase2;")
> d.person <- sqlQuery(ch, "SELECT * FROM PersonBenchmarks WHERE
AgeGp<>'[0,5)';")
> d.house <- sqlQuery(ch, "SELECT * FROM HouseholdBenchmarks;")
> odbcClose(ch)
> d.phase2$Stratum <- factor(d.phase2$Stratum)
> d.house$Stratum <- factor(d.house$Stratum)
```

The field `EligiblePhase2` indicates whether an individual intended to fish during the period of the second phase sample and is hence eligible for the intensive sampling phase, and the field `UsablePhase2` indicates whether an individual's data is sufficiently complete to be used in the analysis.

The sampling weights and stratum totals for the first phase sample are calculated as for the screening

```
> hsampled <- tapply(d.phase2$HouseholdID, d.phase2$Stratum,
+     function(x) length(unique(x)))
> htotal <- tapply(d.house$N, d.house$Stratum, sum)
> d.phase2$weight <-
      htotal[d.phase2$Stratum]/hsampled[d.phase2$Stratum]
> d.phase2$fpc <- htotal[d.phase2$Stratum]
```

This comprises all that is required to specify and calibrate the design. The second phase responses of interest must be adjoined to this basic data for analysis.

The following code imports the kept and released catches of sand flathead, then crosstabulates by person to compute a total kept and released catch for each person (even those that did not fish), and then merges the results into a single data frame

```
> ch <- odbcConnectAccess("Example.mdb")
> d.catch <- sqlQuery(ch, "SELECT * FROM CatchBySpecies
      WHERE CommonName='Flathead - sand'")
> odbcClose(ch)
> levels(d.catch$PersonID) <- levels(d.phase2$PersonID)
> d.catch <- merge(as.data.frame(xtabs(Kept ~ PersonID,
+     data = d.catch), responseName = "KeptFlathead"),
+     as.data.frame(xtabs(Released ~ PersonID, data = d.catch),
+     responseName = "ReleasedFlathead"))
> d.catch$TotalFlathead <- d.catch$KeptFlathead
      + d.catch$ReleasedFlathead
```

The result is a data frame with a row for person and columns for total, kept and released flathead catch

```
> head(d.catch)


    PersonID KeptFlathead ReleasedFlathead TotalFlathead
1 H10001P798          0.0             0.25          0.25
2 H10001P799          0.0             0.25          0.25
3 H10001P800          0.0             0.25          0.25
4 H10001P801          0.0             0.25          0.25
5 H10001P802          1.0             0.00          1.00
6 H10002P913         15.5            15.50         31.00
```

This is then merged with the basic phase two data

```
> d.phase2 <- merge(d.phase2, d.catch)
```

3.1.2 Specifying the design

The primary difference between a basic two-phase analysis and the analysis presented for the screening data is in the specification of the design.

A two-phase design is specified with the `twophase` function. This creates a survey design object that contains the survey data together with the information required to define the two phases of the sampling design. As for the screening, the user must specify

`id` : a list of two formulae specifying the cluster identifiers for the two phases of the design, or `0` for non-clustered designs. A bug in the current implementation of `twophase` requires that these be numeric, not factors.

`strata` : a list of two formulae specifying stratum identifiers for the two phases of the design, or `NULL` for un-stratified designs.

`weights` : a list of two formulae specifying the sampling weights for the two phases of the design.

`fpc` : a formula specifying the stratum population totals for the two phases of the design.

`subset` : a formula indicating the second phase sample as a subset of the screening phase.

`data` : a data frame containing the survey responses, any relevant auxiliary responses, and the aforementioned cluster and strata identifiers, sampling weights and population sizes.

Note that the screening sample forms the population for the second phase sample, and so generally R can calculate the sampling weights and stratum population totals for the second phase sample, so these may be omitted from the design specification.

To meet the requirement that the cluster identifiers be numeric, the factor `HouseholdID` is converted to a sequence of numeric values with `unclass`

```
> d.phase2$hid <- unclass(d.phase2$HouseholdID)
```

For the unadjusted analysis the subset of the screening sample that forms the second phase sample is:

- those units that did not intend to fish and are assumed to have zero catch and effort, together with
- those units that agreed to participate and provided sufficient data to have been deemed usable.

These are simply the units for which `UsablePhase2` is “Y” as all non-intending fishers are automatically deemed usable,

```
> d.phase2$Subset <- d.phase2$UsablePhase2 == "Y"
```

The weights, strata and finite population corrections for the first phase sample are exactly those specified for the screening analysis. The second phase sample is a sample of the first phase sample, stratified by intention to fish, and so `EligiblePhase2` provides the stratum identifiers for the second phase. The weights and finite population correction totals for the second phase can be computed from the first phase data and can be omitted. Finally, `Subset` defines the subset of the first phase sample that forms the second phase sample, and so the sampling design is specified as

```
> library(survey)
> s <- twophase(id=list(~hid,~hid),
+               strata=list(~Stratum,~EligiblePhase2),
+               weights=list(~weight,NULL),
+               fpc=list(~fpc,NULL),
+               subset=~Subset,
+               method="approx",
+               data=d.phase2)
```

Specifying `method="approx"` forces R to use an approximate method for computing standard errors. The approximations used are more than adequate accuracy for most purposes. R can compute exact standard errors for two-phase designs but the required calculations are substantially more resource intensive and are only feasible for smaller surveys.

If no calibration is done, the second phase participation levels as days fished are

```
> svymean(~DaysFished, s)


                       mean     SE
DaysFished[1,5)   0.1248801 0.0065
DaysFished[10,15) 0.0122118 0.0018
DaysFished[15,20) 0.0072149 0.0013
DaysFished[5,10)  0.0482674 0.0038
DaysFished0       0.7986358 0.0081
DaysFished20+     0.0087900 0.0015
```

and the estimated total, kept and released flathead catches during the period of the second phase of the survey are

```
> svytotal(~TotalFlathead + KeptFlathead + ReleasedFlathead,
+     s)


                   total     SE
TotalFlathead    1181195 133781
KeptFlathead      717751  82274
ReleasedFlathead  463444  56154
```

These estimates account for outflow but not influx to the fishery, and should be viewed as lower bounds on the catch.

Note that although the estimated total catch is the sum of the estimated kept and released catches, the standard error for the total cannot be derived from the standard errors of the kept and released estimates because the kept and released catches are correlated.

To perform the analysis omitting the intending households that failed to fish, we first need to create a new variable called HFished ("Y" or "N")

```
> d.phase2$HFished <-
+   ifelse(ave(d.phase2$DaysFished!="0",d.phase2$HouseholdID,FUN=any),
+    "Y","N")
```

The subset of the screening sample that forms the second phase sample then contains those households with usable data, except those that were eligible to fish but did not

```
> d.phase2$Subset <- d.phase2$UsablePhase2=="Y" &
+     !(d.phase2$EligiblePhase2=="Y" & d.phase2$HFished=="N")
```

The sampling design is then specified as before

```
> library(survey)
> s <- twophase(id=list(~hid,~hid),
+               strata=list(~Stratum,~EligiblePhase2),
+               weights=list(~weight,NULL),
+               fpc=list(~fpc,NULL),
+               subset=~Subset,
+               method="approx",
+               data=d.phase2)
```

Now participation levels are estimated as

```
> svymean(~DaysFished, s)


                       mean     SE
DaysFished[1,5)   0.1580880 0.0076
DaysFished[10,15) 0.0154591 0.0023
DaysFished[15,20) 0.0091335 0.0016
DaysFished[5,10)  0.0611026 0.0046
DaysFished0       0.7450895 0.0088
DaysFished20+     0.0111274 0.0019
```

The estimated total, kept and released flathead catches are

```
> svytotal(~TotalFlathead + KeptFlathead + ReleasedFlathead,
+     s)


                   total     SE
TotalFlathead    1184823 149232
KeptFlathead      725451  92408
ReleasedFlathead  459373  61407
```

Comparing with the unadjusted estimates above, it can be seen that assuming influx balances outflux increases the estimated participation and inflates catch estimates. Unfortunately, unless call-backs are made to the non-intending units, there is no real way of assessing the validity of this assumption, and hence the validity of these estimates. In particular, the standard errors reported here are conditional on the assumption of balance – that is, they are only a valid measure of variability in the estimates if the assumption of balance is met.

3.2 Call-backs

Two-phase analysis also provides a natural approach to the treatment of call-back data (Lohr, 1999). A call-back occurs when a sampling unit is re-interviewed after the main survey for the purpose of obtaining additional information.

For example, to estimate the true level of participation adjusted for persons entering (influx) and leaving (outflux) the fishery in the second phase, call-backs are made to households that had not intended to fish to determine if any actually did fish. Again the second phase sample consists of two strata, those households intended to fish and those that did not, except in this case the effort of a sample of non-intending fishers is measured through call-backs, rather than simply assumed to be zero.

The data required is the number of days fished for each person with usable phase one data. For the intending households sampled in the second phase with usable data, and the non-intending households sampled in the call-backs the days fished is determined from the sampled data, but takes missing values for households that did not receive a call-back or did not have usable phase two data.

The sampling weights and stratum totals for the first phase sample are calculated as for the screening

```
> library(RODBC)
> ch <- odbcConnectAccess("Example.mdb")
> d.callback <- sqlQuery(ch, "SELECT * FROM NICData;")
> d.house <- sqlQuery(ch, "SELECT * FROM HouseholdBenchmarks;")
> odbcClose(ch)
> d.callback$Stratum <- factor(d.callback$Stratum)
> d.house$Stratum <- factor(d.house$Stratum)
> d.callback$hid <- unclass(d.callback$HouseholdID)
> hsampled <- tapply(d.callback$HouseholdID, d.callback$Stratum,
```

```
+     function(x) length(unique(x)))
> htotal <- tapply(d.house$N, d.house$Stratum, sum)
> d.callback$weight <-
htotal[d.callback$Stratum]/hsampled[d.callback$Stratum]
> d.callback$fpc <- htotal[d.callback$Stratum]
```

The subset of units included in the second phase are all those for which the number of days fished is known - i.e. those that were contacted in the call-back or those that were eligible for the second phase sampling and produced usable data

```
> d.callback$Subset <- d.callback$NICallback == "Y" |
(d.callback$EligiblePhase2 ==
+     "Y" & d.callback$UsablePhase2 == "Y")
```

The second phase is again stratified by intention to fish as indicated by EligiblePhase2. The design is therefore specified as

```
> library(survey)
> s <- twophase(id=list(~hid,~hid),
+               strata=list(~Stratum,~EligiblePhase2),
+               weights=list(~weight,NULL),
+               fpc=list(~fpc,NULL),
+               subset=~Subset,
+               method="approx",
+               data=d.callback)
```

and the participation levels are estimated to be slightly higher than those estimated assuming influx balanced outflux (Section 3.1.2).

```
> svymean(~DaysFished, s)


                      mean     SE
DaysFished[1,5)   0.159712 0.0084
DaysFished[10,15) 0.020258 0.0036
DaysFished[15,20) 0.011940 0.0026
DaysFished[5,10)  0.057230 0.0044
DaysFished0       0.739002 0.0102
DaysFished20+     0.011858 0.0021
```

3.3 Calibration

Calibration for a two-phase design is identical in concept to calibration for a single-phase design – there is simply more scope for calibration to occur.

In a two-phase design, all estimates are based on the second phase sample, and any calibration is applied to the second phase sample. However, there is still valuable information contained in the first phase sample. So for a two-phase design, it is possible to:

1. Calibrate the second phase to known population totals. This is completely analogous to the calibration of the screening analysis, and adjusts for bias in the overall selection of the second phase sample.
2. Calibrate the second phase sample to totals estimated from the first phase sample. This process has no analogue in the screening analysis, and adjusts for bias in the selection of the second phase sample from the first phase sample. Calibrating the second phase to the first in this way ensures that the results of the two-phase analysis remain consistent with the results of the screening analysis.

Särndal and Lundström (2005) describe calibration for two- phase designs in detail. In practice, two-phase calibration proceeds by first calibrating the first phase sample to known population totals. The calibrated first phase is then used to estimate selected population totals, and both the known population totals and the estimated population totals are used to calibrate the second phase as if it were a single phase sample from the population.

In the current implementation, approximate standard errors are calculated by treating the two-phase design as an equivalent single-phase design.

Calibration is performed with the `calibrate2` function. The user must supply:

`design` : A survey design object to calibrate.

`formula1, population` : A formula specifying the variables to calibrate to known population totals, and the vector of respective totals.

`formula2` : a formula specifying any additional variables used to calibrate the second phase to the first. Estimated population totals for this component of the calibration are computed automatically from the first phase sample.

`formula3, estimates` : Optionally, a third formula and the respective population totals may be specified to calibrate to `estimated` population totals (see Section 3.3.2).

`aggregate.stage` : For cluster samples, the `aggregate.stage` argument should be set to force the weight adjustments to be constant within clusters.

3.3.1 Call-backs

Performing a calibrated analysis of the call-back data essentially requires combining the steps from Sections 2.4 and 3.2.

The basic data requirements are the same, and the sampling weights and stratum totals for the first phase sample, and the second phase stratum identifiers and subset are calculated as in Section 3.2

```
> library(RODBC)
> ch <- odbcConnectAccess("Example.mdb")
> d.callback <- sqlQuery(ch, "SELECT * FROM NICData;")
> d.house <- sqlQuery(ch, "SELECT * FROM HouseholdBenchmarks;")
> d.person <- sqlQuery(ch, "SELECT * FROM PersonBenchmarks WHERE
AgeGp<>'[0,5)';")
```

```
> odbcClose(ch)
> d.callback$Stratum <- factor(d.callback$Stratum)
> d.house$Stratum <- factor(d.house$Stratum)
> d.person$Stratum <- factor(d.person$Stratum)
> d.callback$hid <- unclass(d.callback$HouseholdID)
> hsampled <- tapply(d.callback$HouseholdID, d.callback$Stratum,
+     function(x) length(unique(x)))
> htotal <- tapply(d.house$N, d.house$Stratum, sum)
> d.callback$weight <-
htotal[d.callback$Stratum]/hsampled[d.callback$Stratum]
> d.callback$fpc <- htotal[d.callback$Stratum]
> d.callback$Subset <- d.callback$NICallback == "Y" |
(d.callback$EligiblePhase2 ==
+     "Y" & d.callback$UsablePhase2 == "Y")
```

To calibrate to the household benchmark data it is again necessary to construct indicator variables that correspond to the dummy variables R would create in constructing a design matrix as described in Section 2.4.2

```
> d.callback$W <- ifelse(!duplicated(d.callback$HouseholdID),
+     1, 0)
> d.callback$HType1P <- ifelse(d.callback$HType == "1P",
+     d.callback$W, 0)
> d.callback$HType2P <- ifelse(d.callback$HType == "2P",
+     d.callback$W, 0)
> d.callback$HType3P <- ifelse(d.callback$HType == "3P",
+     d.callback$W, 0)
> d.callback$HType4P <- ifelse(d.callback$HType == "4P",
+     d.callback$W, 0)
> d.callback$HType5PP <- ifelse(d.callback$HType == "5PP",
+     d.callback$W, 0)
```

The design is specified as in Section 3.2

```
> library(survey)
> s <- twophase(id=list(~hid,~hid),
+               strata=list(~Stratum,~EligiblePhase2),
+               weights=list(~weight,NULL),
+               fpc=list(~fpc,NULL),
+               subset=~Subset,
+               method="approx",
+               data=d.callback)
```

and population totals for the calibration are calculated as in Section 2.4.2

```
> p.formula <- terms(~Stratum * Sex * AgeGp, keep.order = T)
> p.totals <- colSums(d.person$N * model.matrix(p.formula,
+     data = d.person))
> h.formula <- terms(~-1 + Stratum:HType, keep.order = T)
> h.totals <- colSums(d.house$N * model.matrix(h.formula,
+     d.house))
```

To calibrate to the household and person level benchmark data, the first formula is specified as in Section 2.4.2, and the second formula consists of just an intercept

```
> source("calibrate2.r")
> formula <- terms(~Stratum*Sex*AgeGp+
+                  Stratum:(HType1P+HType2P+HType3P+HType4P+HType5PP),
+                  keep.order=T)
> s1 <- calibrate2(s,
+                  formula1=formula,
+                  c(p.totals,h.totals),
+                  formula2=~1,
+                  aggregate.stage=1)
```

This calibration slightly inflates the estimated participation levels for the survey period,

```
> svymean(~DaysFished, s1)


                      mean     SE
DaysFished[1,5)   0.169666 0.0095
DaysFished[10,15) 0.019872 0.0035
DaysFished[15,20) 0.012603 0.0026
DaysFished[5,10)  0.062352 0.0050
DaysFished0       0.722703 0.0114
DaysFished20+     0.012806 0.0021
```

but produces estimated participation levels for the previous 12 months that are inconsistent with the analysis of the screening data presented Section 2.4.2.

```
> svymean(~DaysFishedL12M, s1)


                          mean     SE
DaysFishedL12M[1,5)   0.087822 0.0068
DaysFishedL12M[10,15) 0.050094 0.0052
DaysFishedL12M[15,20) 0.023958 0.0034
DaysFishedL12M[5,10)  0.052831 0.0051
DaysFishedL12M0       0.731135 0.0121
DaysFishedL12M20+     0.054160 0.0050
```

Bias in the selection of the second phase sample can be adjusted for by calibrating the second phase to the first. If the second phase is calibrated using `DaysFishedL12M` totals estimated from the first phase,

```
> s1 <- calibrate2(s,
+                  formula1=formula,
+                  c(p.totals,h.totals),
+                  formula2=~DaysFishedL12M,
+                  aggregate.stage=1)
```

the analysis now exactly reproduces the results of the screening analysis,

```
> svymean(~DaysFishedL12M, s1)


                          mean     SE
DaysFishedL12M[1,5)   0.080418 0.0064
DaysFishedL12M[10,15) 0.043564 0.0047
DaysFishedL12M[15,20) 0.023266 0.0033
DaysFishedL12M[5,10)  0.051171 0.0050
DaysFishedL12M0       0.749622 0.0119
DaysFishedL12M20+     0.051959 0.0049
```

and the estimated participation levels for the period of the survey are also affected

```
> svymean(~DaysFished, s1)


                       mean     SE
DaysFished[1,5)   0.163614 0.0094
DaysFished[10,15) 0.019453 0.0036
DaysFished[15,20) 0.012430 0.0027
DaysFished[5,10)  0.059800 0.0049
DaysFished0       0.732325 0.0113
DaysFished20+     0.012378 0.0020
```

For later reference (Section 3.3.2), the estimated total participation by effort class is

```
> svytable(~DaysFished, s1)


DaysFished
     [1,5)    [10,15)    [15,20)     [5,10)          0        20+
 74271.204   8830.731   5642.482  27145.883 332433.903   5618.797
```

3.3.2 Catch and effort

Calibration for the catch and effort data proceeds in the same way as for the call-back data, with one important difference - the design can also be calibrated to participation levels estimated from the call-back data to adjust for influx and outflux from the fishery.

Consider for example the unadjusted analysis presented in Section 3.1. Again the bulk of the calculation proceeds as before

```
> library(RODBC)
> ch <- odbcConnectAccess("Example.mdb")
> d.phase2 <- sqlQuery(ch, "SELECT * FROM Phase2;")
> d.house <- sqlQuery(ch, "SELECT * FROM HouseholdBenchmarks;")
> d.person <- sqlQuery(ch, "SELECT * FROM PersonBenchmarks
+     WHERE AgeGp<>'[0,5)';")
> d.catch <- sqlQuery(ch, "SELECT * FROM Catch
+     WHERE CommonName='Flathead - sand'")
> odbcClose(ch)
```

```
> levels(d.catch$PersonID) <- levels(d.phase2$PersonID)
> d.catch <- merge(as.data.frame(xtabs(Kept ~ PersonID,
+     data = d.catch), responseName = "KeptFlathead"),
+     as.data.frame(xtabs(Released ~ PersonID, data = d.catch),
+     responseName = "ReleasedFlathead"))
> d.catch$TotalFlathead <- d.catch$KeptFlathead +
+     d.catch$ReleasedFlathead
> d.phase2 <- merge(d.phase2, d.catch)
> d.phase2$Stratum <- factor(d.phase2$Stratum)
> d.house$Stratum <- factor(d.house$Stratum)
> d.person$Stratum <- factor(d.person$Stratum)
> hsampled <- tapply(d.phase2$HouseholdID, d.phase2$Stratum,
+     function(x) length(unique(x)))
> htotal <- tapply(d.house$N, d.house$Stratum, sum)
> d.phase2$weight <-
      htotal[d.phase2$Stratum]/hsampled[d.phase2$Stratum]
> d.phase2$fpc <- htotal[d.phase2$Stratum]
> d.phase2$hid <- unclass(d.phase2$HouseholdID)
> d.phase2$Subset <- d.phase2$UsablePhase2 == "Y"
```

and indicator variables are constructed as described in Section 2.4.2

```
> d.phase2$W <- ifelse(!duplicated(d.phase2$HouseholdID), 1, 0)
> d.phase2$HType1P <- ifelse(d.phase2$HType == "1P", d.phase2$W, 0)
> d.phase2$HType2P <- ifelse(d.phase2$HType == "2P", d.phase2$W, 0)
> d.phase2$HType3P <- ifelse(d.phase2$HType == "3P", d.phase2$W, 0)
> d.phase2$HType4P <- ifelse(d.phase2$HType == "4P", d.phase2$W, 0)
> d.phase2$HType5PP <- ifelse(d.phase2$HType == "5PP", d.phase2$W, 0)
```

To calibrate to the household and person benchmark data the design is specified as before

```
> library(survey)
> s <- twophase(id=list(~hid,~hid),
+               strata=list(~Stratum,~EligiblePhase2),
+               weights=list(~weight,NULL),
+               fpc=list(~fpc,NULL),
+               subset=~Subset,
+               method="approx",
+               data=d.phase2)
```

and population totals for the calibration of the first phase are calculated as in Section 2.4.2.

```
> p.formula <- terms(~Stratum * Sex * AgeGp, keep.order = T)
> p.totals <- colSums(d.person$N * model.matrix(p.formula,
+     data = d.person))
> h.formula <- terms(~-1 + Stratum:HType, keep.order = T)
> h.totals <- colSums(d.house$N * model.matrix(h.formula,
+     d.house))
```

Calibrating to just the benchmark data

```
> source("calibrate2.r")
> formula <- terms(~Stratum*Sex*AgeGp+
+                  Stratum:(HType1P+HType2P+HType3P+HType4P+HType5PP),
+                  keep.order=T)
> s1 <- calibrate2(s,
+                  formula1=formula,
+                  c(p.totals,h.totals),
+                  formula2=~1,
+                  aggregate.stage=1)
```

yields estimates of catch and effort that are depressed in comparison with the unadjusted analysis presented in Section 3.1

```
> svymean(~DaysFished, s1)


                       mean     SE
DaysFished[1,5)   0.1253436 0.0067
DaysFished[10,15) 0.0121972 0.0018
DaysFished[15,20) 0.0071503 0.0013
DaysFished[5,10)  0.0505077 0.0040
DaysFished0       0.7959577 0.0085
DaysFished20+     0.0088435 0.0015


> svytotal(~TotalFlathead + KeptFlathead + ReleasedFlathead,
+     s1)


                   total     SE
TotalFlathead    1139473 123546
KeptFlathead      692196  75983
ReleasedFlathead  447277  52705
```

and produces estimates of participation levels for the previous 12 months that are in good agreement with the screening analysis.

```
> svymean(~DaysFishedL12M, s1)


                          mean     SE
DaysFishedL12M[1,5)   0.082167 0.0057
DaysFishedL12M[10,15) 0.046352 0.0045
DaysFishedL12M[15,20) 0.023420 0.0031
DaysFishedL12M[5,10)  0.051499 0.0045
DaysFishedL12M0       0.744707 0.0096
DaysFishedL12M20+     0.051855 0.0045
```

Given the estimated participation levels for the previous 12 months are already in good agreement with the screening analysis, if the second phase is calibrated to the `DaysFishedL12M` totals estimated from the first phase,

```
> s1 <- calibrate2(s,
+                  formula1=formula,
+                  c(p.totals,h.totals),
+                  formula2=~DaysFishedL12M,
+                  aggregate.stage=1)
```

then the screening estimates of participation levels for the previous 12 months are reproduced exactly

```
> svymean(~DaysFishedL12M, s1)


                          mean     SE
DaysFishedL12M[1,5)   0.080418 0.0056
DaysFishedL12M[10,15) 0.043564 0.0042
DaysFishedL12M[15,20) 0.023266 0.0031
DaysFishedL12M[5,10)  0.051171 0.0045
DaysFishedL12M0       0.749622 0.0094
DaysFishedL12M20+     0.051959 0.0045
```

but estimates of catch and effort are largely unaffected.

```
> svymean(~DaysFished, s1)

                       mean     SE
DaysFished[1,5)   0.1232259 0.0066
DaysFished[10,15) 0.0120356 0.0018
DaysFished[15,20) 0.0070990 0.0013
DaysFished[5,10)  0.0497781 0.0040
DaysFished0       0.7990672 0.0084
DaysFished20+     0.0087943 0.0015

> svytotal(~TotalFlathead + KeptFlathead + ReleasedFlathead,
+     s1)

                   total     SE
TotalFlathead    1138322 123370
KeptFlathead      691646  75850
ReleasedFlathead  446675  52645
```

Although this calibrates the second phase to the first, it still makes no direct adjustment for influx or outflux of effort.

It is possible to adjust for influx and outflux directly by calibrating the design to the participation levels estimated from the call-back data. In essence, the calibration matches (based on effort) fishers from the call-back sample for which there is no detailed catch data with fishers that were sampled in the second phase.

The vector of estimated population totals is determined from the calibrated call-back analysis (Section 3.3.1)

```
> est <- c(`DaysFished[1,5)` = 74271.2, `DaysFished[10,15)` = 8830.7,
+     `DaysFished[15,20)` = 5642.5, `DaysFished[5,10)` = 27145.9,
+     DaysFished0 = 332433.9, `DaysFished20+` = 5618.8)
> sum(est)


[1] 453943
```

It is essential that these figures are consistent with the population benchmarks, and the numbers may need to be rounded to ensure the total population matches the person benchmark data. As these are only estimates, as opposed to population totals, they are specified through the optional third formula

```
> s1 <- calibrate2(s,
+                  formula1=formula,
+                  c(p.totals,h.totals),
+                  formula2=~DaysFishedL12M,
+                  formula3=~DaysFished-1,
+                  estimates=est,
+                  aggregate.stage=1)
```

Now the analysis reproduces the estimated participation levels from both the screening and the call-back analysis

```
> svymean(~DaysFishedL12M, s1)

                          mean     SE
DaysFishedL12M[1,5)   0.080418 0.0063
DaysFishedL12M[10,15) 0.043564 0.0047
DaysFishedL12M[15,20) 0.023266 0.0033
DaysFishedL12M[5,10)  0.051171 0.0048
DaysFishedL12M0       0.749622 0.0100
DaysFishedL12M20+     0.051959 0.0049


> svymean(~DaysFished, s1)

                      mean     SE
DaysFished[1,5)   0.163613 0.0089
DaysFished[10,15) 0.019453 0.0031
DaysFished[15,20) 0.012430 0.0024
DaysFished[5,10)  0.059800 0.0049
DaysFished0       0.732325 0.0110
DaysFished20+     0.012378 0.0021
```

while the catch estimates adjusted for influx and outflux are

```
> svytotal(~TotalFlathead + KeptFlathead + ReleasedFlathead,
+     s1)

                   total     SE
TotalFlathead    1032919 108946
KeptFlathead      628409  69110
ReleasedFlathead  404510  44884
```

3.3.3 Effort measures

Influx and outflux to the fishery are adjusted for by calibrating to some measure of effort estimated from the call-back data. However, there is considerable flexibility in the choice of the measure of effort used in the calibration.

In the example presented in the previous section, the design was calibrated to the estimated days fished during the period of the second phase. For the intending fishers the days fished is determined directly from the intensive second phase of sampling. But for non-intending fishers the days fished is determined from a single call-back, and so may be more prone to recall bias.

There are three possible strategies for dealing with this potential problem:

1. Construct effort profiles for both fishers and non-fishers based on a measure of effort recorded in the screening. For example, individuals can be classified by whether or not they have fished during the period of the second phase sample and the number of days they reported fishing in the previous 12 months. Since the days fished in the previous 12 months is recorded in the phase one sample, both the intending and non-intending subgroups are equally prone to recall bias.

2. Assume that the data collected through the call-back is equally reliable as data collected in the second phase sample.

3. Adjust the call-back data to allow for any potential bias. For example, the number of days fished reported in the call-back might be revised down to allow for recall bias.

In the example analyses presented here, option 3 has been used.

3.3.4 Assessing weights

Just as for the screening analysis, the calibration weights for a two-phase design can be assessed by plotting the ratios of the calibration and initial weights (Figure 3.1).

The only minor complication is that for a two-phase design, weights are only defined for the subset in the second phase sample

```
> ratio <- weights(s1)/weights(s)
> stratum <- with(d.phase2, Stratum[Subset])
> plot(ratio ~ stratum, data = d.screen, ylab = "Adjustment")
```

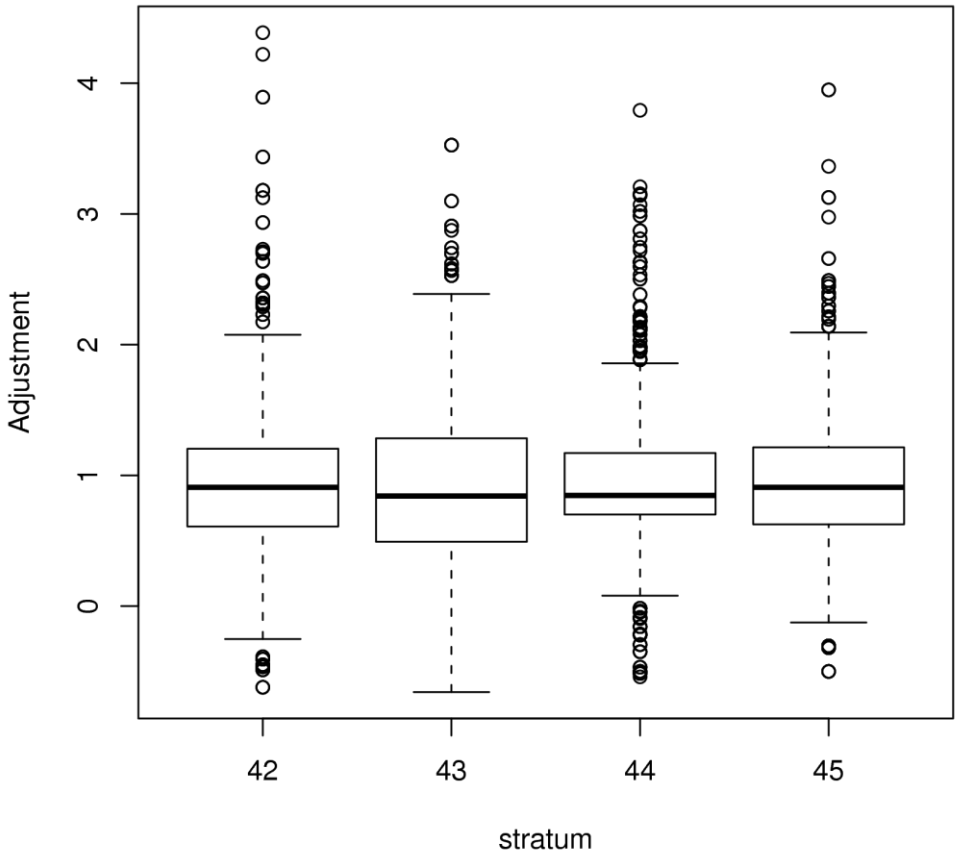

Fig. 3.1: Boxplots by strata of the ratio of the calibrated weights to the initial sampling weights for the second phase data.

Again, negative weights can be prevented by bounding,

```
> s2 <- calibrate2(s,
+                  formula1=formula,
+                  c(p.totals,h.totals),
+                  formula2=~DaysFishedL12M,
+                  formula3=~DaysFished-1,
+                  estimates=est,
+                  aggregate.stage=1,
+                  bounds=c(0.1,20))
```

having minimal impact on the survey estimates in this case

```
> svytotal(~TotalFlathead + KeptFlathead + ReleasedFlathead,
+     s2)


                   total     SE
TotalFlathead    1028828 107792
KeptFlathead      624816  68440
ReleasedFlathead  404012  44354
```

3.4 Large analyses

The analyses presented in this manual have been deliberately kept small for the purposes of illustration. But in practice it will be necessary to analyse larger more complex data sets.

It is strongly recommended that large complex analyses be broken down into smaller more manageable components. So for example, where the example analyses have focused on estimating flathead catch, a more realistic analysis might estimate catch for each major species group or even each major species. Other quantities of interest (for example catch for each fishing method) would then be estimated in separate, distinct analyses.

For analyses that involve no calibration, it is simplest to extract the relevant data from the database and rebuild the design object from scratch for each new analysis. But calibration is computationally expensive so for calibrated analyses it is most efficient to construct the design object and calibrate only once, then reuse the calibrated object for subsequent analyses.

Survey design objects contain the survey data together with the information required to define the sampling design. Once a design object is calibrated, to perform further analyses with the same calibration it suffices to replace just the data component of the calibrated object.

So a more typical workflow would repeat the steps of Section 3.3.2 to construct the calibrated survey object, but without including any catch or effort data

```
> library(RODBC)
> ch <- odbcConnectAccess("Example.mdb")
> d.phase2 <- sqlQuery(ch, "SELECT * FROM Phase2;")
> d.house <- sqlQuery(ch, "SELECT * FROM HouseholdBenchmarks;")
> d.person <- sqlQuery(ch, "SELECT * FROM PersonBenchmarks WHERE
AgeGp<>'[0,5)';")
> odbcClose(ch)
> d.phase2$Stratum <- factor(d.phase2$Stratum)
> d.house$Stratum <- factor(d.house$Stratum)
> d.person$Stratum <- factor(d.person$Stratum)
> hsampled <- tapply(d.phase2$HouseholdID, d.phase2$Stratum,
+     function(x) length(unique(x)))
```

```
> htotal <- tapply(d.house$N, d.house$Stratum, sum)
> d.phase2$weight <-
htotal[d.phase2$Stratum]/hsampled[d.phase2$Stratum]
> d.phase2$fpc <- htotal[d.phase2$Stratum]
> d.phase2$hid <- unclass(d.phase2$HouseholdID)
> d.phase2$Subset <- d.phase2$UsablePhase2 == "Y"
> d.phase2$W <- ifelse(!duplicated(d.phase2$HouseholdID),
+     1, 0)
> d.phase2$HType1P <- ifelse(d.phase2$HType == "1P", d.phase2$W,
+     0)
> d.phase2$HType2P <- ifelse(d.phase2$HType == "2P", d.phase2$W,
+     0)
> d.phase2$HType3P <- ifelse(d.phase2$HType == "3P", d.phase2$W,
+     0)
> d.phase2$HType4P <- ifelse(d.phase2$HType == "4P", d.phase2$W,
+     0)
> d.phase2$HType5PP <- ifelse(d.phase2$HType == "5PP",
+     d.phase2$W, 0)
> library(survey)
> s <- twophase(id = list(~hid, ~hid), strata = list(~Stratum,
+     ~EligiblePhase2), weights = list(~weight, NULL),
+     fpc = list(~fpc, NULL), subset = ~Subset, method = "approx",
+     data = d.phase2)
> p.formula <- terms(~Stratum * Sex * AgeGp, keep.order = T)
> p.totals <- colSums(d.person$N * model.matrix(p.formula,
+     data = d.person))
> h.formula <- terms(~-1 + Stratum:HType, keep.order = T)
> h.totals <- colSums(d.house$N * model.matrix(h.formula,
+     d.house))
> source("calibrate2.r")
> formula <- terms(~Stratum * Sex * AgeGp + Stratum:(HType1P +
+     HType2P + HType3P + HType4P + HType5PP), keep.order = T)
> est <- c(`DaysFished[1,5)` = 77714.2, `DaysFished[10,15)` = 11248.5,
+     `DaysFished[15,20)` = 5303.2, `DaysFished[5,10)` = 29982.4,
+     DaysFished0 = 323340.1, `DaysFished20+` = 6354.6)
> s <- calibrate2(s, formula1 = formula, c(p.totals, h.totals),
+     formula2 = ~DaysFishedL12M, formula3 = ~DaysFished -
+         1, estimates = est, aggregate.stage = 1)
```

Then each separate analysis extracts appropriate catch and effort data,

```
> library(RODBC)
> ch <- odbcConnectAccess("Example.mdb")
> d.catch <- sqlQuery(ch, "SELECT * FROM Catch WHERE
CommonName='Flathead - sand'")
> odbcClose(ch)
> levels(d.catch$PersonID) <- levels(s2$variables$PersonID)
> d.catch <- merge(as.data.frame(xtabs(Kept ~ PersonID,
+     data = d.catch), responseName = "KeptFlathead"),
+     as.data.frame(xtabs(Released ~ PersonID, data = d.catch),
+         responseName = "ReleasedFlathead"))
> d.catch$TotalFlathead <- d.catch$KeptFlathead +
d.catch$ReleasedFlathead
```

which is then adjoined to a copy of the calibrated design object

```
> s1 <- s
> s1$variables <- merge(s1$variables, d.catch)
> svytotal(~TotalFlathead + KeptFlathead + ReleasedFlathead,
+     s1)


                   total     SE
TotalFlathead    1008620 109139
KeptFlathead      612977  69546
ReleasedFlathead  395642  44757
```

4. RESPONSE PROPENSITY MODELS

Response propensity modelling provides an alternative to calibration for adjusting for non-response bias when data is available that characterises non-responding units.

Responding to a recreational fishing survey is voluntary, and inevitably, some sampling units will refuse to participate in the survey from the outset. This leads to non-response bias – the views and actions of the non-responding sub-group become under-represented in the sample and bias results. Calibration provides some protection from non-response bias by reweighting demographic sub-groups that are under or over-represented in the sample. But more direct adjustment is possible when there is data is available that specifically characterises non-responding units.

Typically, data characterizing the non-responding units is obtained through call-backs. The ethics of this practice are dubious as it requires re-interviewing units that have already expressed a preference not to participate, and in some jurisdictions it is expressly prohibited by privacy laws.

Alternately, non-responding units can be characterised by data obtained from partially responding units. Often a unit will give only a partial response as they are content to answer some very basic queries but are not prepared to participate in the full survey. If it can be assumed that the partial responders are typical of all non-responding units, the partial responses can be used in place of call-back data.

Response propensity modelling provides a flexible approach to the analysis of non-response data. Response propensity modelling combines data from non-responding units and data from responding units to model the probability that a unit will respond. This model is then applied to adjust the sampling weights to account for any differences in response propensity across sampled units.

4.1 Data requirements

At heart, a response propensity model is simply a Binary Generalised Linear Model (McCullagh and Nelder, 1989; Collett, 1991) that models whether a unit responds or not, weighted by the sampling fraction. The model may use any covariates that are common to data characterising the non-responders and the first phase of the survey.

The data required for fitting the model comprises the data available for non-responding units (obtained from either partial responders or call-backs) and the corresponding data from units that (fully) responded to the first phase survey. The user must provide the data in the form of a single dataframe, where each row corresponds to a primary sampling unit. In addition to any stratum identifiers, the user must supply:

Covariates: the covariates used in fitting the model,

Response: a binary variable indicating whether the unit fully responded to the survey or not, coded as 1 for response or 0 for non-response, and

Weights: the sampling weights – these are the reciprocals of the sampling fractions for the data used to fit the model viewed as a sample of the *originally intended sample*. That is, for fully responding units the sampling fraction is 1, and for non-responding units the sampling fraction is the fraction of non-responding units for which data is available. So for data obtained by non-response call-backs the weight is the total number of non-responding units in that stratum divided by the number of units that received a call-back in that stratum, while for data obtained from partial responses, the weights are the number of fully non-responding units (full refusal) in that stratum divided by the number of partially responding units (partial refusal) in that stratum.

Depending upon how the data for the non-responding units is obtained, the data used to fit the response propensity model may or may not be collected with the main survey data. For this reason no provision has been made for the storage of non-response call-back data in the example database structure presented in Section 5. The example below simply assumes that appropriate data is available in the table `RPMData`

```
> ch <- odbcConnectAccess("Example.mdb")
> d.rpm <- sqlFetch(ch, "RPMData")
> odbcClose(ch)
> d.rpm$Stratum <- factor(d.rpm$Stratum)
> d.rpm$Response <- ifelse(d.rpm$Response == "Y", 1, 0)
> d.rpm$RPMWeight <- 1/(d.rpm$F)
> head(d.rpm)
```

```
  HouseholdID Stratum State Urban HFishedL12M Response Responsecode F
1      H10257      42   TAS     Y           Y        1             1 1
2      H10258      44   TAS     N           N        1             1 1
3      H10259      44   TAS     N           N        1             1 1
4      H10260      44   TAS     N           N        1             1 1
5      H10261      45   TAS     N           Y        1             1 1
6      H10262      42   TAS     Y           N        1             1 1
  RPMWeight
1         1
2         1
3         1
4         1
5         1
6         1
```

4.2 Fitting the model

A response propensity model is simply a Binomial generalised linear model (Collett, 1991) and can be fitted with `glm`, the standard R function for fitting generalised linear models (see for example Venables and Ripley, 2002). The user must specify

`formula` : A formula specifying the covariates to be included in the fitted model.

`weights` : The weights.

`data` : A dataframe containing the response and the covariates used in fitting the model.

More generally, model selection can be conducted using the standard `add1`, `drop1` and `step` functions (see for example Venables and Ripley, 2002).

For the example data, the probability a household will respond is modelled in terms of the stratum in which it lies and whether the household fished in the previous year, but the effect of fishing history is allowed to vary depending upon whether or not the household lies in an urban area

```
> fit <- glm(Response ~ Stratum + Urban * HFishedL12M,
+     weights = RPMWeight, data = d.rpm)
> summary(fit)


Call:
glm(formula = Response ~ Stratum + Urban * HFishedL12M, data = d.rpm,
    weights = RPMWeight)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3655   0.1180   0.1441   0.1446   0.1465

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.8554121  0.0126279  67.740   <2e-16 ***
Stratum43            0.0067733  0.0249818   0.271   0.7863
Stratum44           -0.0019404  0.0180648  -0.107   0.9145
Stratum45            0.0004665  0.0188935   0.025   0.9803
UrbanY                     NA         NA      NA       NA
HFishedL12MY         0.0629661  0.0180675   3.485   0.0005 ***
UrbanY:HFishedL12MY -0.0363421  0.0302402  -1.202   0.2296
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1209859)

    Null deviance: 326.09  on 2686  degrees of freedom
Residual deviance: 324.36  on 2681  degrees of freedom
AIC: 1828.6

Number of Fisher Scoring iterations: 2
```

4.3 Non-response adjustment

Once a response propensity model has been fitted, a survey is adjusted with the `reweight` function. The user must specify

`weights` : The sampling weights for the design.

`fit` : The fitted response propensity model.

`data` : A dataframe containing the covariates used in the model.

`cluster.id` : Optionally, a cluster identifier. If specified, the reweighting is forced to be constant within the clusters identified by this argument.

`preserve.totals` : Optionally, a stratum identifier. If specified the reweighting is forced to maintain weight totals within the strata specified by this argument.

To repeat the simplest screening analysis presented in Section 2 but adjust for non-response with a response propensity model, the data are extracted as before but it is necessary to extract some additional covariate data to match the covariates used to fit the response propensity model

```
> qry <- "SELECT
+   HouseholdID, PersonID, Stratum, Urban,  HType,
+   Sex, AgeGp, Age, HFishedL12M,
+   IIf(UsablePersons.PFishedStateL12M Is Not Null,
+     UsablePersons.PFishedStateL12M,
+     HFishedL12M) AS PFishedStateL12M,
+   IIf(UsablePersons.DaysFishedL12M Is Not Null,
+     UsablePersons.DaysFishedL12M,'0') AS DaysFishedL12M
+ FROM UsablePersons
+ WHERE AgeGp<>'[0,5)';"
> library(RODBC)
> ch <- odbcConnectAccess("Example.mdb")
> d.screen <- sqlQuery(ch,qry)
> d.house <- sqlQuery(ch,"SELECT * FROM HouseholdBenchmarks;")
> d.person <- sqlQuery(ch,"SELECT * FROM PersonBenchmarks WHERE
AgeGp<>'[0,5)';")
> odbcClose(ch)
> d.screen$Stratum <- factor(d.screen$Stratum)
> d.house$Stratum <- factor(d.house$Stratum)
> d.person$Stratum <- factor(d.person$Stratum)
```

It is essential that the covariates match the data used to fit the response propensity model, even down to the exact levels of factors.

The initial sampling weights and stratum totals are calculated as before

```
> hsampled <- tapply(d.screen$HouseholdID, d.screen$Stratum,
+     function(x) length(unique(x)))
> htotal <- tapply(d.house$N, d.house$Stratum, sum)
> d.screen$weight <-
htotal[d.screen$Stratum]/hsampled[d.screen$Stratum]
> d.screen$fpc <- htotal[d.screen$Stratum]
```

but once calculated, the weights are then adjusted from the response propensity model

```
> source("reweight.r")
> d.screen$weight1 <- reweight(weight,fit,
+                              data=d.screen,
+                              cluster.id=HouseholdID,
+                              preserve.totals=Stratum)
```

Here the `cluster.id` argument ensures individuals from the same household receive the same weight, and the `preserve.totals` argument ensures the adjustment preserves the sums of the weights within each stratum. Re-weighting aims to up or down weight under or over-represented units in the sample; preserving the sums of weights within strata ensures no stratum becomes under or over-represented in the sample.

From this point on the analysis proceeds as before. First the design is specified using the new adjusted weights

```
> library(survey)
> s <- svydesign(ids=~HouseholdID,
+                strata=~Stratum,
+                weight=~weight1,
+                fpc=~fpc,
+                data=d.screen)
```

The adjusted estimates of participation and participation levels are

```
> svymean(~PFishedStateL12M, s)


                      mean     SE
PFishedStateL12MN 0.76866 0.0085
PFishedStateL12MY 0.23134 0.0085

> svymean(~DaysFishedL12M, s)


                           mean     SE
DaysFishedL12M[1,5)    0.076933 0.0051
DaysFishedL12M[10,15) 0.040820 0.0036
DaysFishedL12M[15,20) 0.021236 0.0026
DaysFishedL12M[5,10)   0.047888 0.0039
DaysFishedL12M0        0.765125 0.0085
DaysFishedL12M20+      0.047999 0.0038
```

slightly lower than the unadjusted estimates in Section 2.

After the response propensity adjustment has been applied, the design can then be calibrated just as before

```
> d.screen$W <- ifelse(!duplicated(d.screen$HouseholdID),1,0)
> d.screen$HType1P <- ifelse(d.screen$HType=="1P",d.screen$W,0)
> d.screen$HType2P <- ifelse(d.screen$HType=="2P",d.screen$W,0)
> d.screen$HType3P <- ifelse(d.screen$HType=="3P",d.screen$W,0)
> d.screen$HType4P <- ifelse(d.screen$HType=="4P",d.screen$W,0)
> d.screen$HType5PP <- ifelse(d.screen$HType=="5PP",d.screen$W,0)
> s <- svydesign(ids=~HouseholdID,
+                strata=~Stratum,
+                weight=~weight1,
+                fpc=~fpc,
+                data=d.screen)
```

```
> p.formula <- terms(~Stratum*Sex*AgeGp,keep.order=T)
> p.totals <-
colSums(d.person$N*model.matrix(p.formula,data=d.person))
> h.formula <- terms(~-1+Stratum:HType,keep.order=T)
> h.totals <- colSums(d.house$N*model.matrix(h.formula,d.house))
> formula <-
terms(~Stratum*Sex*AgeGp+Stratum:(HType1P+HType2P+HType3P+HType4P+HTyp
e5PP),keep.order=T)
> s1 <- calibrate(s,formula,c(p.totals,h.totals),aggregate.stage=1)
> svymean(~PFishedStateL12M,s1)


                      mean     SE
PFishedStateL12MN 0.76133 0.0086
PFishedStateL12MY 0.23867 0.0086


> svymean(~DaysFishedL12M,s1)


                           mean     SE
DaysFishedL12M[1,5)   0.077918 0.0051
DaysFishedL12M[10,15) 0.042115 0.0038
DaysFishedL12M[15,20) 0.022549 0.0028
DaysFishedL12M[5,10)  0.049585 0.0041
DaysFishedL12M0       0.757668 0.0086
DaysFishedL12M20+     0.050166 0.0041
```

5 DATABASE STRUCTURE

Any survey of appreciable magnitude will require some form of database for data management and storage. The precise structure of this database will be dictated by the fine detail of the survey and the many practical issues associated with data collection.

As a guide to the design of a survey database, this section outlines an example survey with a similar structure to the NRFS, and presents the core design of a relational database for the management of the survey data, together with sample queries to extract data in the format required by the *RecSurvey* package. The table structure presented here focuses only on the data required for the survey analysis. In practice this structure will need to be augmented to incorporate the practical details required to administer the survey, but this additional structure is easily built around the core structure suggested here.

Table definitions and queries are presented in both standard SQL92, and the dialect of SQL recognised by Microsoft Access. While SQL92 standardises syntax, the data types available for data storage are highly dependent on the database management system. For portability the structures below use only integer, text and timestamp data types, but more efficient representations of the data may be available in the particular database management system.

5.1 Example survey design

The example survey is based on the NRFS (Henry and Lyle, 2003), and is a two-phase design. The first phase comprises a large-scale screening survey, where the demographic profile, recent recreational fishing involvement and likelihood (expectation) of future recreational fishing activity are determined. Within the screening survey, those households that have fished previously or intend to fish within the next 12-months are asked extra questions, providing more detail at the individual level.

Households containing any person with a positive expectation of fishing in the upcoming 12-month period are invited to participate in the second phase, which in the NRFS was a phone-diary survey. The subset of households from the first phase that are eligible and agree to participate in the second phase are surveyed in more detail regarding their fishing practises. All fishing events within the second phase period are recorded, with detail on the location, method used, effort expended, catch composition and quantity.

5.2 Database design

This section outlines the core design of a relational database corresponding to the example survey described in the previous section.

5.2.1 Households table

The Households table records data specific to the sampled households; its structure is described in Table 5.1.

Note that the table is not perfectly normalized. In principle, the state of residence, whether the household is in an urban district, SD and SSD can all be determined from the SLA. Similarly Npersons could be computed from the information in the Persons table described below, but is retained as this question is directly asked in the survey. The HUsablePhase1 and HUsablePhase2 columns indicate whether the household has provided enough information to be used in the analysis of the two phases, and typically will be filled only upon completion of the respective phase of sampling.

5.2.2 Person table

The Persons table records individual details for the members of all sampled households. The table structure is presented in Table 5.2

At minimum the survey aims to determine the age class of every respondent, but in some cases respondents volunteer their precise age and this is recorded separately.

5.2.3 PersonsDetail table

Additional questions are asked of the members of households that have either fished in the last 12 months or intend to fish during the period of the second phase sampling. This additional detail is recorded in the PersonsDetail table, as described in Table 5.3.

5.2.4 FishingEvents table

The FishingEvents table records Phase2 data. A fishing event is a continuous period of fishing with a single gear type at a single location by one or more fishers. If a group of fishers simultaneously fish using several methods or gear types, this constitutes multiple events. Similarly, if a group of fishers change fishing method or location during a continuous period of fishing this also constitutes multiple events.

The structure of the FishingEvents table is described in Table 5.4. Note that StartDate and EndDate are timestamps, they contain both the date and time that the event starts and ends.

5.2.5 CatchEvents table

A catch event is the total catch of a single species from a single fishing event. If several species are caught during a single fishing event, then these constitute multiple catch events. Catch events are related to fishing events and hence fishing parties, not individual fishers, and catch recorded is the total catch of the entire fishing party, not a specific individual. In practice, however, data may be provided at the level of the individual fisher within a fishing party.

The structure of the CatchEvents table is described in Table 5.5

5.2.6 FishingPartys table

The FishingPartys table links fishing events with the fishers participating in that event. The structure of the FishingPartys table is described in Table 5.6. This is a standard tabular representation of the many to many relationship between fishers and fishing events.

5.2.7 Lookup tables

The Species table and Fishing methods table summarize information about the fish species and the fishing methods. These tables are described in Table 5.7.

The Regions and Subregions tables summarize information about the fishing locations. These tables are described in Table 5.8.

5.2.8 Benchmark tables

Certain known population totals are required for the calibration process. The HouseholdBenchmarks and PersonBenchmarks tables store data on the number of households of a particular composition and the number of persons of a particular age and gender within each sampling stratum. Typically this data is not collected during the survey but is obtained from an external agency such as the ABS.

The structure of these tables is described in Table 5.9. Typically these tables are a direct input required by the analysis and they will be selected in their entirety with a query of the form "SELECT * FROM tablename".

5.2.9 Non-Intending fisher call-backs

To gauge true participation rates, it is desirable to make call-backs to households that had not intended to fish to determine if any actually did fish during the second phase of the survey. The NICHouseholds and NICPersons tables (Table 5.10) store data on the fishing activity by non-intending households and household members obtained from call-backs.

Table 5.1: Structure of the Households table.

Households Table	
HouseholdID	Text code that identifies the household.
Stratum	Sampling stratum.
State	State of residence.
SD,SSD,SLA	Fine grain descriptors of household location as defined by the ABS - the statistical division (SD), statistical sub-division (SSD) and statistical local area (SLA).
Urban	Is the household in an urban district (Y/N)?
NPersons	Number of persons in the household.
HFishedL12M	Has any household member fished in the last 12 months (Y/N)?
HOwnBoat	Does the household own a boat (Y/N)?
HMemberClub	Is any household member currently a member of a recreational fishing club (Y/N)?
HLikelihood	Likelihood that any household member will fish in the second phase of the survey (1– very likely, 2– likely, 3– unlikely, 4– very unlikely, 5– unknown).
UsablePhase1	Is the first phase data from this household sufficiently complete to be used in the analysis (Y/N)?
EligiblePhase2	Is this household eligible for the second phase sample (Y/N)?
UsablePhase2	Is the second phase data from this household sufficiently complete to be used in the analysis (Y/N)?

```
CREATE TABLE Households (
   HouseholdID TEXT PRIMARY KEY,
   Stratum INTEGER NOT NULL,
   State TEXT NOT NULL,
   SD TEXT NOT NULL,
   SSD TEXT NOT NULL,
   SLA TEXT NOT NULL,
   Urban TEXT NOT NULL,
   NPersons INTEGER,
   HFishedL12M TEXT,
   HOwnBoat TEXT,
   HMemberClub TEXT,
   HLikelihood INTEGER,
   UsablePhase1 TEXT,
   EligiblePhase2 TEXT,
   UsablePhase2 TEXT);
```

Table 5.2: Structure of the Persons table.

Persons Table	
PersonID	Text code that identifies the person.
HouseholdID	Text code that identifies the person's household.
Sex	Person's gender (M/F).
AgeGp	Age class ([0,5), [5,15), [15,30), [30,45), [45,60), [60,100)).
Age	Age at last birthday if known.

```
CREATE TABLE Persons (
   PersonID TEXT PRIMARY KEY,
   HouseholdID TEXT NOT NULL,
   Sex TEXT,
   AgeGp TEXT,
   Age INTEGER,
   FOREIGN KEY (HouseholdID)
   REFERENCES
   Households(HouseholdID));
```

Table 5.3: Structure of the PersonsDetail table.

PersonsDetail Table	
PersonID	Text code that identifies the person.
PFishedStateL12M	Has person fished recreationally in this State in the last 12 months (Y/N)?
PFishedOtherStateL12M	Has person fished recreationally in another State in the last 12 months (Y/N)?
FreshOrSalt	Was fishing predominantly in freshwater, saltwater or both (F,S,B)?
DaysFishedL12M	Code indicating number of days fished ([1,5), [5,10), [10,15), [15,20), 20+).
PLikelihood	Likelihood person will fish recreationally next 12 months (1– very likely, 2– likely, 3– unlikely, 4– very unlikely, 5– unknown).
PMemberClub	Is person currently a member of a recreational fishing club (Y/N)?

```
CREATE TABLE PersonsDetail (
   PersonID TEXT PRIMARY KEY,
   PFishedStateL12M TEXT,
   PFishedOtherStateL12M TEXT,
   FreshOrSalt TEXT,
   DaysFishedL12M TEXT,
   PLikelihood INTEGER,
   PMemberClub TEXT,
   FOREIGN KEY (PersonID)
   REFERENCES
   Persons(PersonID));
```

Table 5.4: Structure of the FishingEvents table

FishingEvents Table	
FishingEventID	Integer code identifying the event.
StartDate, EndDate	The start and end date and time of the event.
MinutesBreak	Time spent not fishing (minutes).
RegionCode, SubregionCode	Codes identifying the location fished.
PrimaryTargetCode, SecondaryTargetCode	Species targeted in fishing.
MethodCode	Code describing the fishing method.
NGear, NHauls	When applicable, the number of hauls/gear used.
NPersons	Number of fishers involved in event.
Platform	Fishing platform - Boat, Shore or Both.
Boat	Where applicable boat type – Private or Charter.
Shore	Where applicable, integer code categorizing the shore type.

```
CREATE TABLE FishingEvents (
   FishingEventID INTEGER
     PRIMARY KEY AUTOINCREMENT,
   StartDate TIMESTAMP,
   EndDate TIMESTAMP,
   MinutesBreak INTEGER,
   RegionCode INTEGER,
   SubregionCode INTEGER,
   PrimaryTargetCode TEXT,
   SecondaryTargetCode TEXT,
   MethodCode INTEGER,
   NGear INTEGER,
   NHauls INTEGER,
   NPersons INTEGER,
   Platform TEXT,
   Boat TEXT,
   Shore INTEGER,
   FOREIGN KEY (MethodCode)
   REFERENCES
   FishingMethods(MethodCode));
```

Table 5.5: Structure of the CatchEvents table

CatchEvents Table	
CatchEventID	Integer code indentifying the catch event.
FishingEventID	Integer code indentifying the fishing event.
SpeciesID	Text code that identifies the species caught.
NKept	The number kept.
NReleased	The number released.
NTooSmall,	Number released by reason for release.
NUnderSize,	
NTooMany,	
NOverLimit,	
NCatchRelease,	
NOther	

```
CREATE TABLE CatchEvents (
   CatchEventID INTEGER
     PRIMARY KEY AUTOINCREMENT,
   FishingEventID INTEGER
   NOT NULL,
   SpeciesID INTEGER,
   NKept INTEGER,
   NReleased INTEGER,
   NTooSmall INTEGER,
   NUnderSize INTEGER,
   NTooMany INTEGER,
   NOverLimit INTEGER,
   NCatchRelease INTEGER,
   NOther INTEGER);
```

Table 5.6: Structure of the FishingPartys table.

FishingPartys Table	
FishingEventID	Integer code that identifies the fishing event.
PersonID	Text code that identifies the person fishing.

```
CREATE TABLE FishingPartys (
   FishingEventID INTEGER NOT NULL,
   PersonID TEXT NOT NULL,
   PRIMARY KEY(FishingEventID,PersonID),
   FOREIGN KEY (FishingEventID)
   REFERENCES
   FishingEvents(FishingEventID),
   FOREIGN KEY(PersonID)
   REFERENCES
   Persons(PersonID));
```

Table 5.7: Structure of the Species and Fishing Methods tables

Species Table	
SpeciesID	Species identification code.
CommonName	The common name of the species.
CAABCode	CSIRO CAAB (Codes for Australian Aquatic Biota).
CAABFamilyCode	CSIRO CAAB family.
StandardName	The standard name of the species.
ScientificName	The scientific name of the species.
GeneralName	The general name of the species.
SppGroup	The species 'group' for reporting purposes.

```
CREATE TABLE Species (
   SpeciesID INTEGER
     PRIMARY KEY,
   CommonName TEXT,
   CAABCode INTEGER UNIQUE,
   CAABFamilyCode INTEGER,
   StandardName TEXT,
   ScientificName TEXT,
   GeneralName TEXT,
   SppGroup TEXT);
```

FishingMethods Table	
MethodCode	Integer code identifying the fishing method.
MethodName	Description of the fishing method.
MethodGroup	Fishing method 'group', for reporting purposes.

```
CREATE TABLE FishingMethods (
   MethodCode INTEGER
     PRIMARY KEY,
   MethodName TEXT NOT NULL,
   MethodGroup TEXT NOT NULL);
```

Table 5.8: Structure of the Regions and Subregions tables.

Regions Table	
RegionCode	Integer code identifying Region.
Region	Description of the region.
RegionGroup	The region 'group', for reporting purposes.

```
CREATE TABLE Regions (
   RegionCode INTEGER
     PRIMARY KEY,
   Region TEXT,
   RegionGroup TEXT);
```

Subregions Table	
SubregionCode	Integer code identifying the fishing subregion.
Subregion	Description of the subregion.
SubregionGroup	The subregion 'group', for reporting purposes.

```
CREATE TABLE Subregions (
   SubregionCode INTEGER
     PRIMARY KEY,
   Subregion TEXT,
   SubregionGroup TEXT);
```

Table 5.9: Structure of the Household and Person Benchmark tables.

HouseholdBenchmarks Table	
Stratum	Sampling stratum.
Urban	Is the household in an urban district (Y/N)?
HType	Household composition code (1P- one person household; 2P- 2 person household, ..., 5PP- 5 or more person household).
N	Number of households with this composition in the stratum.

```
CREATE TABLE HouseholdBenchmarks (
  Stratum INTEGER NOT NULL,
  Urban TEXT NOT NULL,
  HType TEXT NOT NULL,
  N INTEGER NOT NULL,
  PRIMARY KEY(Stratum,HType));
```

PersonBenchmarks Table	
Stratum	Sampling stratum.
Sex	Gender (M/F).
AgeGp	Age class ([0,5), [5,15), [15,30), [30,45), [45,60), [60,100)).
N	The number of persons of this age class & gender in the stratum.

```
CREATE TABLE PersonBenchmarks (
  Stratum INTEGER NOT NULL,
  Sex TEXT NOT NULL,
  AgeGp TEXT NOT NULL,
  N INTEGER NOT NULL,
  PRIMARY KEY(Stratum,Sex,AgeGp));
```

Table 5.10: Structure of the Household and Person tables for call-backs to households that did not intend to fish in the second phase of the survey.

NICHouseholds Table	
HouseholdID	Text code that identifies the household.
HFished	Did any household member fish during the phase 2 sample period (Y/N)?

```
CREATE TABLE NICHouseholds (
   HouseholdID TEXT PRIMARY KEY,
   HFished TEXT NOT NULL,
   FOREIGN KEY (HouseholdID) REFERENCES
Households(HouseholdID));
```

NICPersons Table	
PersonID	Text code that identifies the person.
HouseholdID	Text code that identifies the person's household.
PFishedState	Did the person fish in the State during the phase 2 sample period (Y/N)?
PFishedOtherState	Did the person fish in another State during the phase 2 sample period (Y/N)?
DaysFished	Code indicating number of days fished during the phase 2 sample period.
FreshOrSalt	Predominantly fresh water, salt water or both (F,S,B).

```
CREATE TABLE NICPersons (
   PersonID TEXT PRIMARY KEY,
   HouseholdID TEXT NOT NULL,
   PFishedState TEXT,
   PFishedOtherState TEXT,
   DaysFished TEXT,
   FreshOrSalt TEXT,
   FOREIGN KEY(HouseholdID)
   REFERENCES NICHouseholds(HouseholdID),
   FOREIGN KEY(PersonID)
   REFERENCES Persons(PersonID));
```

5.3 Views

In addition to the tables described in the previous section, several sub-queries are instantiated as views to simplify subsequent queries. Microsoft Access does not distinguish views and queries, in Access a view is simply a query. But because these are referenced by other queries, the queries presented here must be stored within Access itself.

5.3.1 Household structure

The household composition view calculates features of the household composition directly from the Persons table. The view computes the total number of persons residing in each household, the number aged five years or older, and a code classifying the household in terms of its composition (Tables 5.11 and 5.12).

Table 5.11: HouseholdStructure view (SQL92).

```
CREATE VIEW HouseholdStructure AS
SELECT
  HouseholdID, NPersons, NPersons5P,
  CASE
    WHEN NPersons=1 THEN '1P'
    WHEN NPersons=2 THEN '2P'
    WHEN NPersons=3 THEN '3P'
    WHEN NPersons=4 THEN '4P'
    WHEN NPersons>4 THEN '5PP'
  END AS HType
FROM Households LEFT JOIN
    (SELECT
       HouseholdID,
       SUM(CASE
         WHEN AgeGp='[0,5)' THEN 0
         ELSE 1
       END) AS NPersons5P
     FROM Persons
     GROUP BY HouseholdID) AS Counts
     USING(HouseholdID);
```

Table 5.12: HouseholdStructure view (ACCESS).

```
SELECT
  Households.HouseholdID, Households.NPersons, Counts.NPersons5P,
  Switch(NPersons=1,'1P',
  NPersons=2,'2P',
  NPersons=3,'3P',
  NPersons=4,'4P',
  NPersons>4,'5PP') AS HType
FROM Households LEFT JOIN
    (SELECT
       HouseholdID,
       SUM(Switch(AgeGp='[0,5)',0,TRUE,1)) AS NPersons5P
     FROM Persons GROUP BY HouseholdID) AS Counts
     ON Households.HouseholdID=Counts.HouseholdID;
```

5.3.2 UsablePersons

The UsablePersons view forms the join of the Households, Persons and PersonsDetail tables, and retains only those rows corresponding to individuals from households that have responded with sufficient completeness to be deemed usable for the first phase of the survey (Tables 5.13 and 5.14).

Table 5.13: UsablePersons view (SQL92).

```
CREATE VIEW UsablePersons AS
SELECT
  Households.HouseholdID AS HouseholdID,
  PersonID, Stratum, State, SD, SSD, SLA, Urban,
  Households.NPersons, NPersons5P, HType,
  Sex, AgeGp, Age, HFishedL12M,
  HOwnBoat, HMemberClub, HLikelihood,
  PFishedStateL12M, PFishedOtherStateL12M,
  FreshOrSalt, DaysFishedL12M, PLikelihood, PMemberClub,
  EligiblePhase2, UsablePhase2
FROM Households
  INNER JOIN HouseholdStructure USING(HouseholdID)
  INNER JOIN Persons USING(HouseholdID)
  LEFT JOIN PersonsDetail USING(PersonID)
WHERE UsablePhase1='Y';
```

Table 5.14: UsablePersons query (ACCESS).

```
SELECT
  Households.HouseholdID, Persons.PersonID,
  Stratum, State, SD, SSD, SLA, Urban,
  Households.NPersons, NPersons5P, HType,
  Sex, AgeGp, Age, HFishedL12M,
  HOwnBoat, HMemberClub, HLikelihood,
  PFishedStateL12M, PFishedOtherStateL12M,
  FreshOrSalt, DaysFishedL12M, PLikelihood, PMemberClub,
  EligiblePhase2, UsablePhase2
FROM ((Households
  INNER JOIN HouseholdStructure
  ON Households.HouseholdID=HouseholdStructure.HouseholdID)
  INNER JOIN Persons
  ON HouseholdStructure.HouseholdID=Persons.HouseholdID)
  LEFT JOIN PersonsDetail
  ON Persons.PersonID=PersonsDetail.PersonID
WHERE UsablePhase1='Y';
```

5.3.3 NICallbacks

The NICallbacks view unifies the information from the NICHouseholds and NICPersons tables (Tables 5.15 and 5,16).

Table 5.15: NICallbacks view (SQL92).

```
CREATE VIEW NICallbacks AS
SELECT
  NICHouseholds.HouseholdID,
  Persons.PersonID,
  NICHouseholds.HFished,
  COALESCE(NICPersons.DaysFished,'0') AS DaysFished,
  CASE
    WHEN NICPersons.PersonID Is Null THEN 'N'
    WHEN NICPersons.PFishedState='Y' OR
         NICPersons.PFishedOtherState='Y' THEN 'Y'
    ELSE 'N'
  END AS PFished,
  NICPersons.FreshOrSalt
FROM NICHouseholds
     INNER JOIN Persons USING(HouseholdID)
     LEFT JOIN NICPersons USING(PersonID);
```

Table 5.16: NICallbacks view (ACCESS).

```
SELECT
  NICHouseholds.HouseholdID,
  Persons.PersonID,
  NICHouseholds.HFished,
  IIf(NICPersons.DaysFished Is Null,'0',NICPersons.DaysFished)
    AS DaysFished,
  IIf(NICPersons.PersonID Is Null,'N',IIf(NICPersons.PFishedState='Y'
    Or NICPersons.PFishedOtherState='Y','Y','N')) AS PFished,
  NICPersons.FreshOrSalt
FROM (NICHouseholds
       INNER JOIN Persons ON
       NICHouseholds.HouseholdID = Persons.HouseholdID)
       LEFT JOIN NICPersons ON Persons.PersonID = NICPersons.PersonID;
```

5.3.4 Phase one data

The first or screening phase of the survey can be viewed as a survey in its own right, and the data from this phase of the survey can be analysed before the data from the second phase are available. The queries presented in Tables 5.17 and 5.18 extract the data required for the analysis of the first phase, as described in Section 2.

The Phase1 query is very similar to the UsablePersons view. Some questions are only asked of individuals that are members of households that have fished in the previous year or intend to fish in the second phase sampling period. For these questions the individual responses are used where available, but where the question was not asked of an individual, the equivalent household level response is used instead. Furthermore, some households express a complete non-interest in fishing at an early stage of the interview. For these households the interview is terminated early and it is assumed that the members of the household undertake no fishing related activities. For example, members of fishing households are asked the likelihood they will fish in the period of the second phase sampling. For members of non-fishing and non-intending fisher households this information is not available at the individual level, and the likelihood that any member of the household will fish is used instead.

Table 5.17: Query for extracting the phase 1 data (SQL92).

```
CREATE VIEW Phase1 AS
SELECT
  HouseholdID, PersonID,
  Stratum, State, SD, SSD, SLA, Urban,
  NPersons, NPersons5P, HType,
  Sex, AgeGp, Age, HFishedL12M, HOwnBoat,
  COALESCE(PMemberClub,'N') AS PMemberClub,
  COALESCE(PFishedStateL12M, 'N') AS PFishedStateL12M,
  COALESCE(PFishedOtherStateL12M, 'N') AS PFishedOtherStateL12M,
  COALESCE(DaysFishedL12M,'0') AS DaysFishedL12M,
  COALESCE(PLikelihood,HLikelihood) AS Likelihood
FROM UsablePersons
WHERE AgeGp<>'[0,5)';
```

Table 5.18: Query for extracting the phase 1 data (ACCESS).

```
SELECT
  HouseholdID, PersonID,
  Stratum, State, SD, SSD, SLA, Urban,
  NPersons, NPersons5P, HType,
  Sex, AgeGp, Age, HFishedL12M, HOwnBoat,
  IIf(UsablePersons.PMemberClub Is Not Null,
   UsablePersons.PMemberClub,
   'N') AS PMemberClub,
  IIf(UsablePersons.PFishedStateL12M Is Not Null,
   UsablePersons.PFishedStateL12M,
   'N') AS PFishedStateL12M,
  IIf(UsablePersons.PFishedOtherStateL12M Is Not Null,
   UsablePersons.PFishedOtherStateL12M,
   'N') AS PFishedOtherStateL12M,
  IIf(UsablePersons.DaysFishedL12M Is Not Null,
   UsablePersons.DaysFishedL12M,'0') AS DaysFishedL12M,
  IIf(PLikelihood Is Null,
   HLikelihood, PLikelihood) AS Likelihood
FROM UsablePersons
WHERE AgeGp<>'[0,5)';
```

5.3.5 Phase two data

The queries presented in Tables 5.19 and 5.20 compute the number of days fished by each fisher during the second phase, where the number of days fished is defined as the number of distinct days on which a fishing event ends.

Table 5.19: Query for computing number of days fished per fisher (SQL92).

```
CREATE VIEW DaysFished AS
SELECT
  PersonID,
  Days,
  CASE
    WHEN Days=0 THEN '0'
    WHEN Days<5 THEN '[1,5)'
    WHEN Days<10 THEN '[5,10)'
    WHEN Days<15 THEN '[10,15)'
    WHEN Days<20 THEN '[15,20)'
    WHEN Days>=20 THEN '20+'
    END AS DaysFished
FROM (SELECT
  PersonID,
  COUNT(DISTINCT DATE(EndDate)) AS Days
FROM UsablePersons
  LEFT JOIN FishingPartys USING(PersonID)
  LEFT JOIN FishingEvents USING(FishingEventID)
GROUP BY PersonID)
```

Table 5.20: Query for computing number of days fished per fisher (ACCESS).

```
SELECT
  UsablePersons.PersonID,
  COUNT(Day) AS Days,
  Switch(
    Days=0,'0',
    Days<5,'[1,5)',
    Days<10,'[5,10)',
    Days<15,'[10,15)',
    Days<20,'[15,20)',
    Days>=20,'20+') AS DaysFished
FROM UsablePersons
  LEFT JOIN (SELECT DISTINCT
    PersonID,
    DateValue(EndDate) AS Day
  FROM FishingPartys
    INNER JOIN FishingEvents
    ON FishingPartys.FishingEventID=FishingEvents.FishingEventID) AS
Events
  ON UsablePersons.PersonID = Events.PersonID
GROUP BY UsablePersons.PersonID
```

The analysis of the second phase requires this effort data (DaysFished; Tables 2.19 and 5.20), together with the auxiliary data to specify and calibrate the design. This auxiliary data is effectively the same data required for the analysis of the first phase sample, and can be extracted with the queries presented in Tables 5.21. and 5.22. These queries are almost identical to their phase one counterparts, with the addition of fields describing days fished in the period of the second phase, along with identifiers for which individuals were eligible for the intensive sampling phase and whether the data collected was sufficiently complete to be deemed usable.

Table 5.21: Query for extracting the phase 2 data (SQL92).

```
CREATE VIEW Phase2 AS
SELECT
  HouseholdID, PersonID,
  Stratum, State, SD, SSD, SLA, Urban,
  NPersons, NPersons5P, HType,
  Sex, AgeGp, Age, HFishedL12M, HOwnBoat,
  COALESCE(PMemberClub,'N') AS PMemberClub,
  COALESCE(PFishedStateL12M, 'N') AS PFishedStateL12M,
  COALESCE(PFishedOtherStateL12M, 'N') AS PFishedOtherStateL12M,
  COALESCE(DaysFishedL12M,'0') AS DaysFishedL12M,
  COALESCE(PLikelihood,HLikelihood) AS Likelihood,
  DaysFished,
  EligiblePhase2,
  UsablePhase2
FROM UsablePersons
  INNER JOIN DaysFished ON Using(PersonID)
WHERE AgeGp<>'[0,5)';
```

Table 5.22: Query for extracting the phase 2 data (ACCESS).

```
SELECT
  HouseholdID, UsablePersons.PersonID,
  Stratum, State, SD, SSD, SLA, Urban,
  NPersons, NPersons5P, HType,
  Sex, AgeGp, Age, HFishedL12M, HOwnBoat,
 IIf(UsablePersons.PMemberClub Is Not Null,
  UsablePersons.PMemberClub, 'N') AS PMemberClub,
 IIf(UsablePersons.PFishedStateL12M Is Not Null,
  UsablePersons.PFishedStateL12M,
  'N') AS PFishedStateL12M,
 IIf(UsablePersons.PFishedOtherStateL12M Is Not Null,
  UsablePersons.PFishedOtherStateL12M,
  'N') AS PFishedOtherStateL12M,
 IIf(UsablePersons.DaysFishedL12M Is Not Null,
  UsablePersons.DaysFishedL12M,'0') AS DaysFishedL12M,
 IIf(PLikelihood Is Not Null,
  PLikelihood,HLikelihood) AS Likelihood,
 DaysFished.DaysFished,
 UsablePersons.EligiblePhase2,
 UsablePersons.UsablePhase2
 FROM UsablePersons
  INNER JOIN DaysFished ON UsablePersons.PersonID =
  DaysFished.PersonID
 WHERE (((UsablePersons.AgeGp)<>'[0,5)'));
```

5.3.6 NICallback data

The NICData queries presented in Tables 5.23 and 5.24 are similar to those for extracting Phase2 data (Section 5.3.5), except information on days fished for non-intending fishers is filled in using NICallbacks (Section 5.3.3).

Table 5.23: Query for extracting the NICallback data (SQL92).

```
SELECT
  UsablePersons.HouseholdID,
  UsablePersons.PersonID,
  Urban, Stratum, NPersons5P, HType, Sex, AgeGp, HFishedL12M,
  EligiblePhase2, UsablePhase2,
  CASE
    WHEN NICallbacks.HouseholdID Is Null THEN 'N'
    ELSE 'Y'
  END AS NICallback,
  COALESCE(UsablePersons.DaysFishedL12M,'0') AS DaysFishedL12M,
  CASE
    WHEN NICallbacks.HouseholdID Is Null AND EligiblePhase2='Y' THEN
DaysFished.DaysFished
    WHEN NICallbacks.HouseholdID Is Null THEN NULL
    ELSE NICallbacks.DaysFished
  END AS DaysFished
 FROM UsablePersons
  INNER JOIN DaysFished USING(PersonID)
  LEFT JOIN NICallbacks USING(PersonID)
 WHERE UsablePersons.AgeGp<>'[0,5)';
```

Table 5.24: Query for extracting the NICallback data (ACCESS).

```
SELECT
  UsablePersons.HouseholdID,
  UsablePersons.PersonID,
  Urban, Stratum, NPersons5P, HType, Sex, AgeGp, HFishedL12M,
  EligiblePhase2, UsablePhase2,
 IIf([NICallbacks].[HouseholdID] Is Null,'N','Y') AS NICallback,
 IIf([UsablePersons].[DaysFishedL12M] Is
  Null,'0',[UsablePersons].[DaysFishedL12M]) AS DaysFishedL12M,
 IIf([NICallbacks].[HouseholdID] Is
  Null,IIf([EligiblePhase2]='Y',[DaysFished].[DaysFished]),
  [NICallbacks].[DaysFished]) AS DaysFished
 FROM (UsablePersons
  INNER JOIN DaysFished ON UsablePersons.PersonID =
  DaysFished.PersonID)
  LEFT JOIN NICallbacks ON UsablePersons.PersonID =
  NICallbacks.PersonID
 WHERE (((UsablePersons.AgeGp)<>'[0,5)'));
```

5.4 Sample queries

In addition to the data described in the previous section, the analysis of the second phase also requires basic catch and effort data. This section presents a number of sample queries for extracting this data for analysis in the format required by the *RecSurvey* package. The queries are presented in both SQL92 and Access dialects, and would be executed through an ODBC connection to the database from within R.

5.4.1 Catch by species

The queries presented in Tables 5.25 and 5.26 join FishingPartys, FishingEvents, CatchEvents and Species tables to compute the total catch of each species recorded by each individual. The CatchEvents table records only the total catch for the fishing party, so the catch must be evenly divided amongst the members of the party. More specific catch details can be extracted by restricting the query with an appropriate `WHERE` clause.

Table 5.25: Query for computing total catch by species (SQL92).

```
SELECT
  PersonID,
  SpeciesID,
  CommonName,
  SUM(NKept/NPersons) AS Kept,
  SUM(NReleased/NPersons) AS Released
FROM FishingPartys
     INNER JOIN FishingEvents USING(FishingEventID)
     INNER JOIN CatchEvents USING(FishingEventID)
     INNER JOIN Species USING(SpeciesID)
GROUP BY PersonID, SpeciesID, CommonName
```

Table 5.26: Query for computing total catch by species (ACCESS).

```
SELECT
 FishingPartys.PersonID,
 Species.SpeciesID,
 Species.CommonName,
  SUM(NKept/NPersons) AS Kept,
  SUM(NReleased/NPersons) AS Released
FROM ((FishingPartys
     INNER JOIN FishingEvents ON FishingPartys.FishingEventID =
FishingEvents.FishingEventID)
     INNER JOIN CatchEvents ON FishingEvents.FishingEventID =
CatchEvents.FishingEventID)
     INNER JOIN Species ON CatchEvents.SpeciesID = Species.SpeciesID
GROUP BY PersonID, Species.SpeciesID, Species.CommonName
```

5.4.2 Catch by fishing method

The queries presented in Tables 5.27 and 5.28 compute the total catch by fisher and by a grouping of fishing methods defined by the column MethodGroup in the FishingMethods table.

Table 5.27: Query for computing total catch by fisher and grouped fishing method (SQL92).

```
SELECT
  PersonID,
  MethodGroup,
  Sum(NKept/NPersons) AS Kept,
  Sum(NReleased/NPersons) AS Released
FROM (FishingPartys
      INNER JOIN FishingEvents USING(FishingEventID)
      INNER JOIN CatchEvents USING(FishingEventID)
      INNER JOIN Species USING(SpeciesID))
     INNER JOIN FishingMethods USING(MethodCode)
GROUP BY PersonID, MethodGroup;
```

Table 5.28: Query for computing total catch by fisher and grouped fishing method (ACCESS).

```
SELECT
  FishingPartys.PersonID,
  FishingMethods.MethodGroup,
  Sum(NKept/NPersons) AS Kept,
  Sum(NReleased/NPersons) AS Released
FROM Species
  INNER JOIN (FishingMethods
  INNER JOIN ((FishingEvents
  INNER JOIN CatchEvents ON FishingEvents.FishingEventID =
CatchEvents.FishingEventID)
  INNER JOIN FishingPartys ON FishingEvents.FishingEventID =
FishingPartys.FishingEventID)
  ON FishingMethods.MethodCode = FishingEvents.MethodCode)
  ON Species.SpeciesID = CatchEvents.SpeciesID
GROUP BY FishingPartys.PersonID, FishingMethods.MethodGroup;
```

5.4.3 Days fished by region

The queries presented in Tables 5.29 and 5.30 compute the number of days fished by each fisher by Region grouping, where the number of days fished is defined as the number of distinct days on which a fishing event ends.

Table 5.29: Query for computing number of days fished per fisher (SQL92).

```
SELECT
  PersonID,
  RegionGroup AS RegionGp,
  Count(Day) AS DaysFished
FROM
  (SELECT DISTINCT
     PersonID,
     RegionGroup,
     DateValue(EndDate) AS Day
   FROM FishingPartys
        INNER JOIN FishingEvents USING(FishingEventID)
        INNER JOIN Regions USING(RegionCode)) AS Counts
GROUP BY PersonID, RegionGroup;
```

Table 5.30: Query for computing number of days fished per fisher by grouped Regions (ACCESS)

```
SELECT
  Counts.PersonID,
  Counts.RegionGroup AS RegionGp,
  Count(Counts.Day) AS DaysFished
FROM
  (SELECT DISTINCT
     PersonID,
     RegionGroup,
     DateValue(EndDate) AS Day
   FROM Regions
 INNER JOIN (FishingEvents
 INNER JOIN FishingPartys
 ON FishingEvents.FishingEventID=FishingPartys.FishingEventID)
 ON Regions.RegionCode=FishingEvents.RegionCode) AS Counts
GROUP BY Counts.PersonID, Counts.RegionGroup;
```

5.4.4 Effort (hours) by line fishing method

The queries presented in Tables 5.31 and 5.32 compute the total line fishing effort (in hours) by fisher.

Table 5.31: Query for computing total effort (hours) by fisher and fishing method (SQL92).

```
SELECT
  PersonID,
  MethodGroup,
  Sum((EndDate-StartDate)*24-COALESCE(MinutesBreak,0)/60) AS Hours
FROM FishingMethods
     INNER JOIN FishingEvents USING(MethodCode)
     INNER JOIN FishingPartys USING(FishingEventID)
WHERE MethodGroup='Line'
GROUP BY PersonID, MethodGroup;
```

Table 5.32: Query for computing total line fishing effort (hours) by fisher (ACCESS).

```
SELECT
      FishingPartys.PersonID,
      FishingMethods.MethodGroup,
      Sum((FishingEvents.EndDate-FishingEvents.StartDate)*24-
      IIf(FishingEvents.MinutesBreak Is Not
      Null,FishingEvents.MinutesBreak/60,0)) AS Hours
FROM FishingMethods
      INNER JOIN (FishingEvents
      INNER JOIN FishingPartys ON FishingEvents.FishingEventID =
FishingPartys.FishingEventID)
      ON FishingMethods.MethodCode = FishingEvents.MethodCode
WHERE FishingMethods.MethodGroup='Line'
GROUP BY FishingPartys.PersonID, FishingMethods.MethodGroup;
```

PART 2: RE-ANALYSIS OF KEY NRFS DATA

6.1 INTRODUCTION

The National Recreational Fishing Survey (NRFS) represented a component of the National Recreational and Indigenous Fishing Survey (Henry and Lyle, 2003) that was conducted in 2000-01, the focus being the recreational fishing activities of Australian residents. The purpose of this section is twofold; first to compare original and re-analysed NRFS estimates for key parameters, and second to present example outputs from the re-analysed Tasmanian component of the NRFS using the *RecSurvey* package. The comparison between original and re-analysed estimates is based on data for Tasmania and South Australia and provides insight into the implications of re-analysis on previously reported estimates (Henry and Lyle, 2003). In addition to the present report, re-analysed NRFS data have been reported and compared with findings from more recent state-wide recreational fishing surveys in Tasmania (Lyle *et al*., 2009) and in South Australia (Jones, 2009).

6.2 COMPARISON WITH ORIGINAL ESTIMATES

6.2.1 Estimation procedures

In order to assess the impacts of re-analysis, key phase one (screening) and phase two (diary) data have been compared with original estimates reported in Henry and Lyle (2003) for Tasmania and South Australia.

The original analysis procedure described in Henry and Lyle (2003) differed from the re-analysis approach described in Part 1 in several ways, these differences related mainly to the way adjustments for non-response were implemented. Specifically, in the original analysis, phase one non-response adjustment was based on non-response call-backs and was applied *after* calibration to ABS census data. Adjustments were made sensitive to demographic characteristics (age group, gender and stratum) and avidity (reported days fished), and were applied at the person level, often resulting in different weighting factors applied to individuals within a household. By contrast, for the re-analysis, non-response adjustment was applied *prior* to calibration and was based on household fishing propensity determined for partial responders (refer Section 4) rather than call-backs. The decision to use partial responders (part-refusals) rather than call-backs recognised that, for ethical reasons, future surveys would not be able to conduct non-response call-backs in the manner of the NRFS.

Non-response adjustment for the second phase was individually based in the original analysis but household based for the re-analysis; in both instances avidity (individual/household) and stratum were factored into the adjustments.

Participation in fishing is dynamic, with individuals entering (influx) and leaving (outflux) the fishery through time. This dynamic should ideally be measured and adjusted for in the analysis (refer Section 3). For the NRFS, a non-intending fisher call-back survey was designed to provide an adjustment for influx, however, the sample size of the call-back component was deemed inadequate and the adjustment considered unreliable (Henry and Lyle, 2003). As a consequence, equilibrium was assumed, an assumption that has been typically applied in other diary surveys (e.g. Bradford, 1998; Higgs, 1999; 2001). To ensure comparability between estimation procedures, equilibrium in terms of influx and outflux was assumed for the re-analysis (option 1 in Section 3.3.3).

6.2.2 Participation rates

Participation rates based on region of residence and age class are presented for Tasmania (Figure 6.1) and South Australia (Figure 6.2). There was strong alignment between original and re-analysed estimates of participation by stratum, indicating that the impact of the re-analysis on overall participation at state-wide and regional scales was minimal. Similarly, when data were disaggregated by age class, the re-analysed values were very similar to the original values (which were provided without standard error). These findings indicate that despite differences in adjustment and weighting procedures, estimates of the key demographic parameters did not differ significantly.

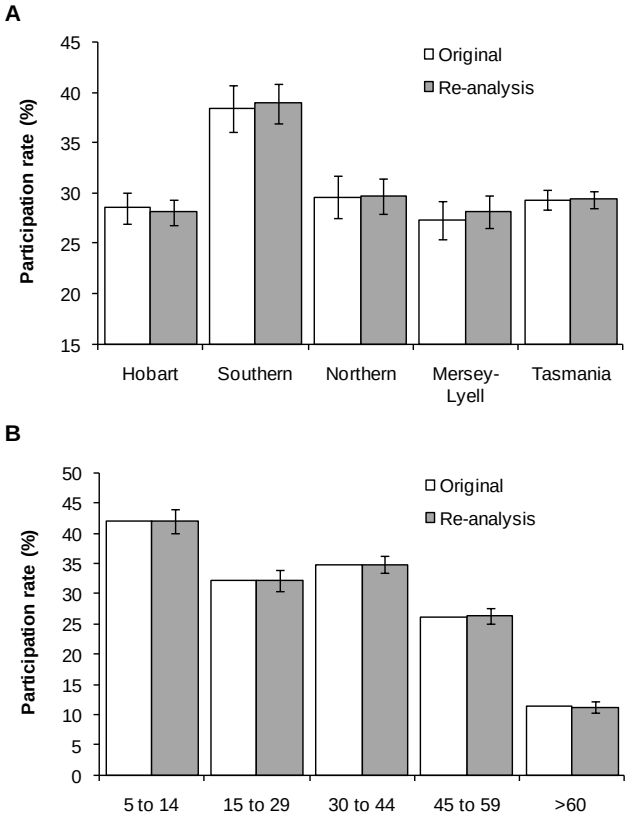

Fig. 6.1 Recreational fishing participation rates in 2000 by (A) region of residence (statistical division) and (B) age group for Tasmanian residents aged five years or older. Error bars represent one standard error.

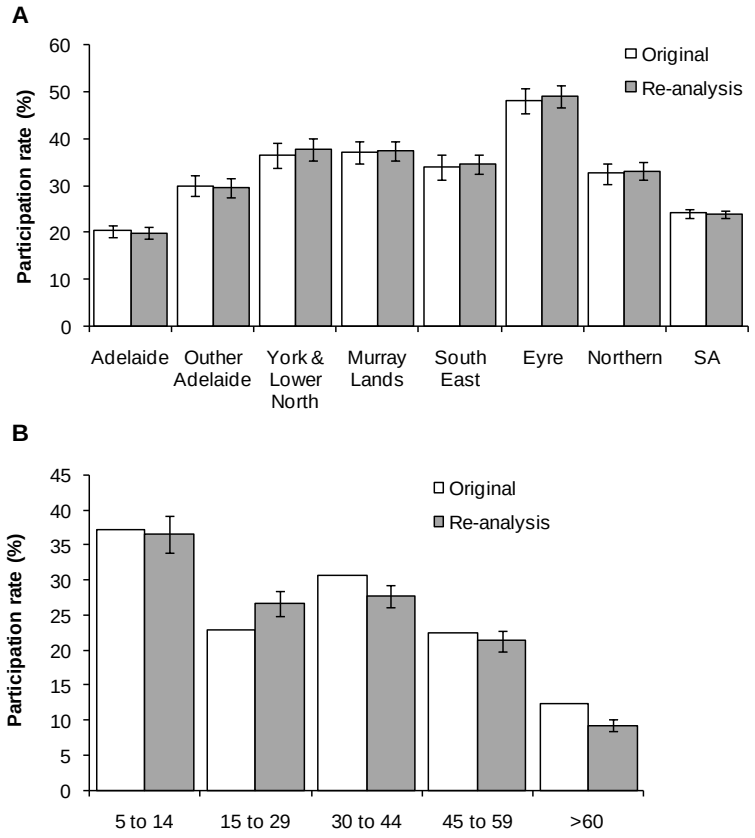

Fig. 6.2 Recreational fishing participation rates in 2000 by (A) region of residence (statistical division) and (B) age group for South Australian residents aged five years or older. Error bars represent one standard error.

6.2.3 Harvest estimates

State-wide harvest estimates derived from the re-analysis were based on catches taken by residents only, whereas estimates reported by Henry and Lyle (2003) include harvest taken by resident as well as non-resident fishers. While fishing effort and catches taken by non-residents represented a very minor component of the state-wide totals for Tasmania and South Australia (Henry and Lyle, 2003), for comparative purposes harvest taken by non-residents has been excluded from the original totals.

Harvest estimates for the major species taken in Tasmania and South Australia during 2000-01 are presented in Figure 6.3 and Figure 6.4, respectively. Flathead catches, the dominant species taken by recreational fishers in Tasmania, have been excluded from Figure 6.3 for clarity. Catches of this species were estimated to be 1,376,951 (SE 153,963) in the original analysis and compare with 1,158,110 (SE 140,348) for the re-analysis.

While some variation is to be expected between estimation procedures, differences in harvest estimates were not significant. In instances where estimates differed markedly, for instance jack mackerel (Tasmania), standard errors tended to be very large and overlapping.

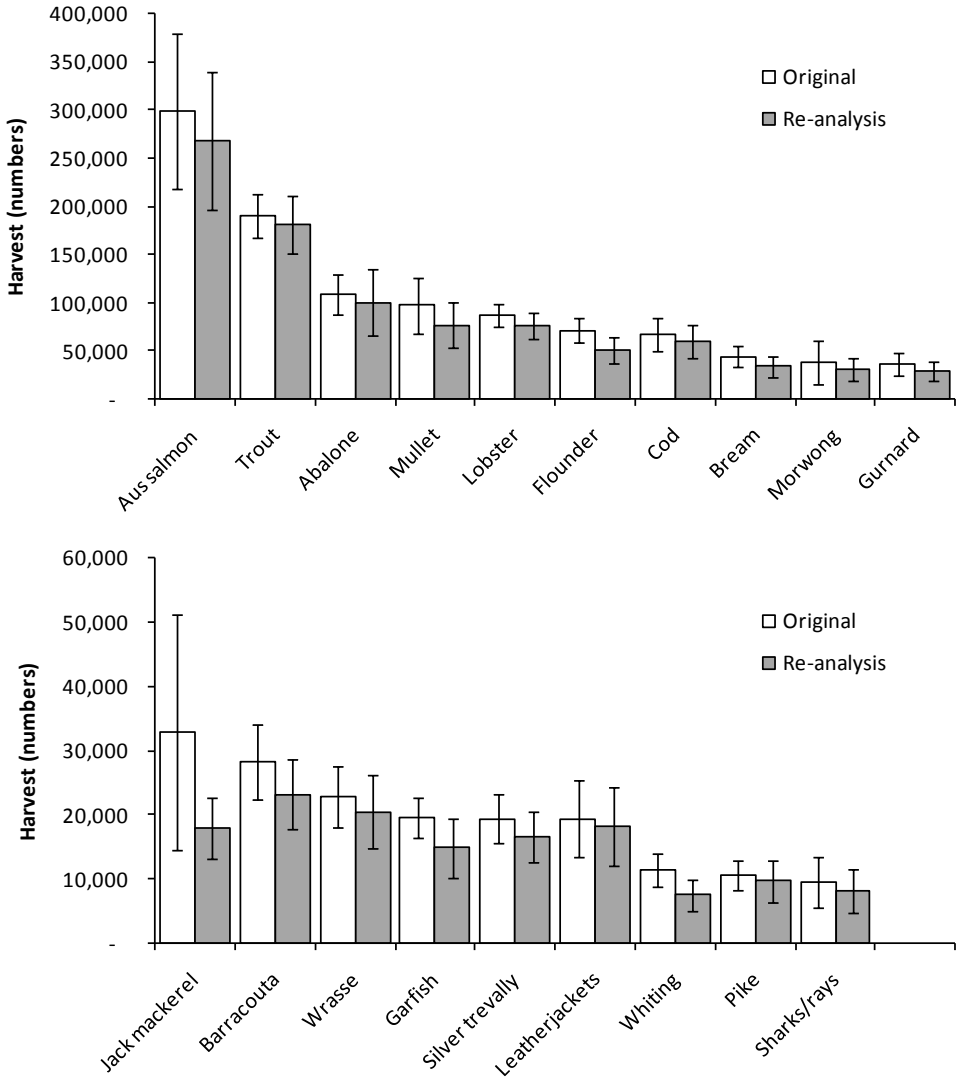

Fig 6.3. Key species harvest estimates for Tasmanian residents aged five years or older during 2000-01; comparison of original and re-analysed estimates. Error bars represent one standard error.

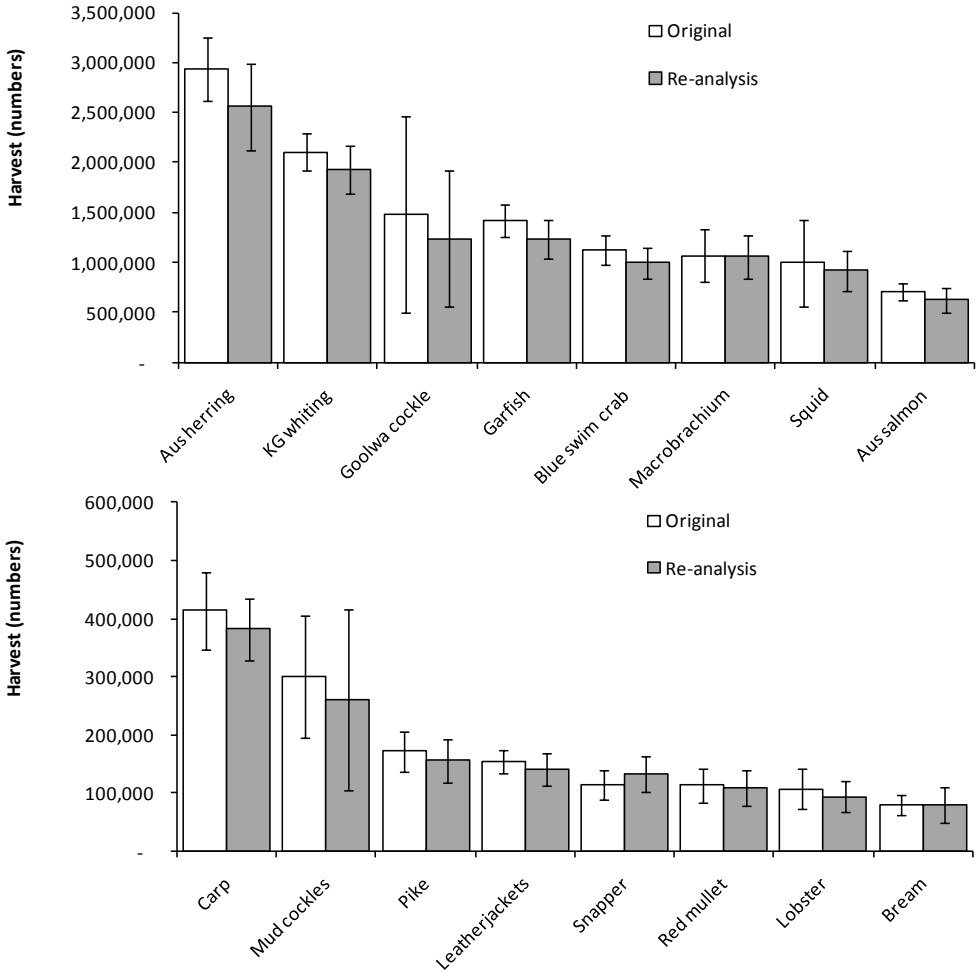

Fig 6.4. Key species harvest estimates for South Australian residents aged five years or older during 2000-01; comparison of original and re-analysed estimates. Error bars represent one standard error.

6.2.4 Conclusion

Overall the impact of re-analysis on key NRFS parameter estimates was statistically non-significant, indicating that the outputs from the original analysis were generally robust. Notwithstanding this observation, re-analysis of NRFS for all jurisdictions is recommended.

6.3 RE-ANALYSIS OF KEY TASMANIAN DATA

Lyle (2005) provided a detailed analysis of NRFS data as it pertained to Tasmania. However, owing to the complexity in calculating errors associated with the original analysis procedure, data disaggregated below state level were reported without associated errors. Since the relative size of the statistical error or uncertainty associated with parameter estimates increases with disaggregation, it is highly desirable to take this uncertainly into account when determining the reliability of individual estimates.

The primary aim of the present section is to demonstrate the capability of the *RecSurvey* package in providing disaggregated outputs of the type that have relevance to management and fisheries assessment. Key NRFS data for Tasmania are provided for this purpose and as a consequence interpretation of the implications of results is limited.

6.3.1 Regions

Selection of the initial household sample was based on regional stratification based on the four Australian Bureau of Statistics statistical divisions: Greater Hobart, Southern, Northern, and Mersey-Lyell (Figure 6.5). In describing household and population characteristics data were analysed at stratum and state levels.

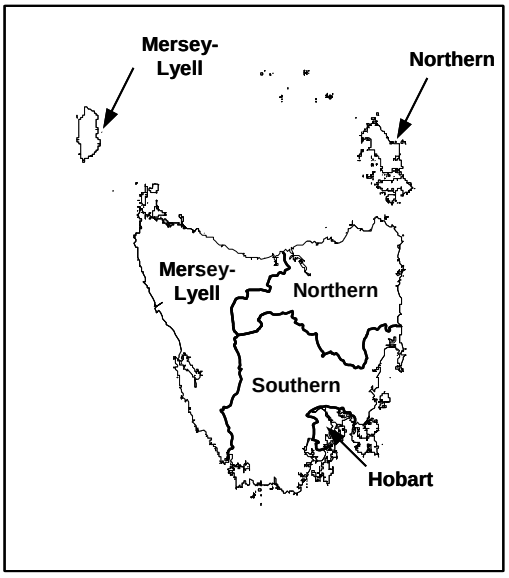

Fig. 6.5 Map of Tasmania showing ABS Statistical Divisions used for sample stratification.

During the diary survey, interviewers classified the location of each fishing activity (event) into one of 23 fishing regions. For reporting purposes it was necessary to collapse some regions to ensure that a minimum of 250 fishing events (i.e. raw unexpanded data) occurred in each reporting region. The fishing regions used for

reporting included inland, selected estuarine (Tamar and Derwent estuaries), and coastal regions as indicated in Figure 6.6.

Other fishing location information was also collected in the diary survey in terms of water-body type: marine waters greater than or less than 5kms from the coastline; estuarine waters; freshwater rivers; and freshwater lakes/dams, public or private.

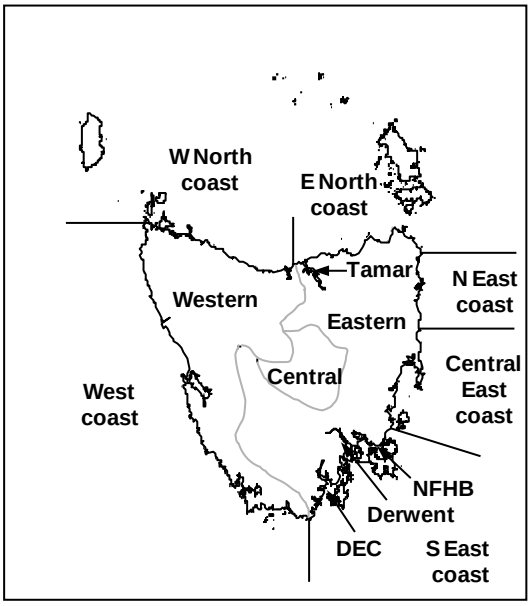

Fig. 6.6 Map of Tasmania showing analysis regions used for reporting fishing activities. Fishing regions - NFHB Norfolk and Frederick Henry bays; DEC D'Entrecasteaux Channel.

6.3.2 Participation

A total of 125,000 Tasmanian residents aged five years or older were estimated to have fished at least once in the 12 months prior to May 2000, representing an overall participation rate of just over 29% (Table 6.1). Participation was also disaggregated in terms of age, gender and stratum (statistical division). This analysis revealed that substantially more males than females fished and that participation rates were varied greatly with age group and stratum (Table 6.2).

Table 6.1: Estimated number of persons and percentage of the Tasmanian resident population aged five years or older who fished recreationally in the 12 months prior to May 2000.

Statistical division	Population number	Recreational fishers Number	SE	Participation rate (%)	SE
Greater Hobart	176,120	49,448	2,280	28.1	1.3
Southern	31,617	12,307	626	38.9	2.0
Northern	119,985	35,629	2,129	29.7	1.8
Mersey-Lyell	98,071	27,635	1,607	28.2	1.6
Total	**425,793**	**125,018**	**3,565**	**29.4**	**0.8**

Table 6.2: Estimated number of persons and percentage of the Tasmanian resident population aged five years or older by age, gender and statistical division, who fished recreationally in the 12 months prior to May 2000.

Statistical division	Age class	Males			Part'n rate (%)	Females			Part'n rate (%)
		Pop'n	Fishers	SE		Pop'n	Fishers	SE	
Greater Hobart									
	5 to 14	13,808	7,071	606	51.2	13,422	4,289	544	32.0
	15 to 29	19,915	7,116	638	35.7	20,172	4,852	588	24.1
	30 to 44	19,840	8,830	623	44.5	21,662	4,326	518	20.0
	45 to 59	17,742	7,042	549	39.7	18,287	2,939	390	16.1
	60+	13,894	2,417	336	17.4	17,378	565	178	3.3
Southern									
	5 to 14	2,821	1,809	159	64.1	2,572	1,059	146	41.2
	15 to 29	2,954	1,533	173	51.9	2,708	688	145	25.4
	30 to 44	3,833	2,163	161	56.4	3,853	1,310	141	34.0
	45 to 59	3,636	1,717	147	47.2	3,369	1,047	128	31.1
	60+	3,027	765	112	25.3	2,844	216	65	7.6
Northern									
	5 to 14	9,619	5,055	491	52.6	9,113	3,068	493	33.7
	15 to 29	12,881	5,593	648	43.4	13,094	4,007	610	30.6
	30 to 44	13,780	6,849	566	49.7	14,175	3,317	460	23.4
	45 to 59	12,116	3,753	471	31.0	12,241	1,367	316	11.2
	60+	10,597	2,175	353	20.5	12,369	443	165	3.6
Mersey Lyell -									
	5 to 14	8,474	4,051	454	47.8	7,932	2,145	319	27.0
	15 to 29	10,027	4,033	483	40.2	10,028	1,799	362	17.9
	30 to 44	11,440	5,619	429	49.1	11,858	2,579	344	21.7
	45 to 59	10,120	3,949	407	39.0	10,013	1,239	258	12.4
	60+	8,444	2,040	295	24.2	9,735	179	104	1.8
Tasmania									
	5 to 14	34,722	17,986	917	51.8	33,039	10,562	814	32.0
	15 to 29	45,777	18,275	1,043	39.9	46,002	11,347	933	24.7
	30 to 44	48,893	23,461	959	48.0	51,548	11,533	786	22.4
	45 to 59	43,614	16,461	844	37.7	43,910	6,593	579	15.0
	60+	35,962	7,398	581	20.6	42,326	1,403	272	3.3

6.3.3 Effort

Fishing effort can be expressed in a number of ways, for the current analysis fisher days have been used, however alternative metrics include fisher hours and fishing events. Overall, Tasmanian residents accounted for a total of 698,306 (SE 38,658) fisher days of effort during 2000-01, the estimated numbers of fishers and fisher days by fishing region are presented in Table 6.3.

Table 6.3: Annual recreational effort (number of fishers and fisher days) by fishing region during 2000-01, based on Tasmanian residents aged five years or older.
SE is standard error; values in italics indicate that fewer than 30 households provided information.

Region	Fishers	SE	Fisher days	SE
Western	13,656	1,543	51,392	9,940
Central	13,181	1,512	68,513	9,789
Eastern	17,723	1,794	67,277	10,003
West coast	3,981	780	26,870	10,833
West north coast	19,621	1,900	88,925	13,414
Tamar	12,789	1,697	41,645	7,315
East north coast	11,372	1,581	30,929	6,743
North east coast	12,095	1,513	51,644	10,685
Central east coast	18,512	1,998	65,012	9,994
South east coast	12,533	1,537	50,460	9,733
NFHB	15,917	1,666	38,030	5,484
Derwent	17,527	1,893	53,672	10,906
DEC	20,324	1,823	70,486	7,935

Effort levels disaggregated by fishing platform are provided in Figure 6.7 and revealed not only regional differences in overall effort levels but considerable variability in the relative importance of boat-based as opposed to shore-based fishing effort.

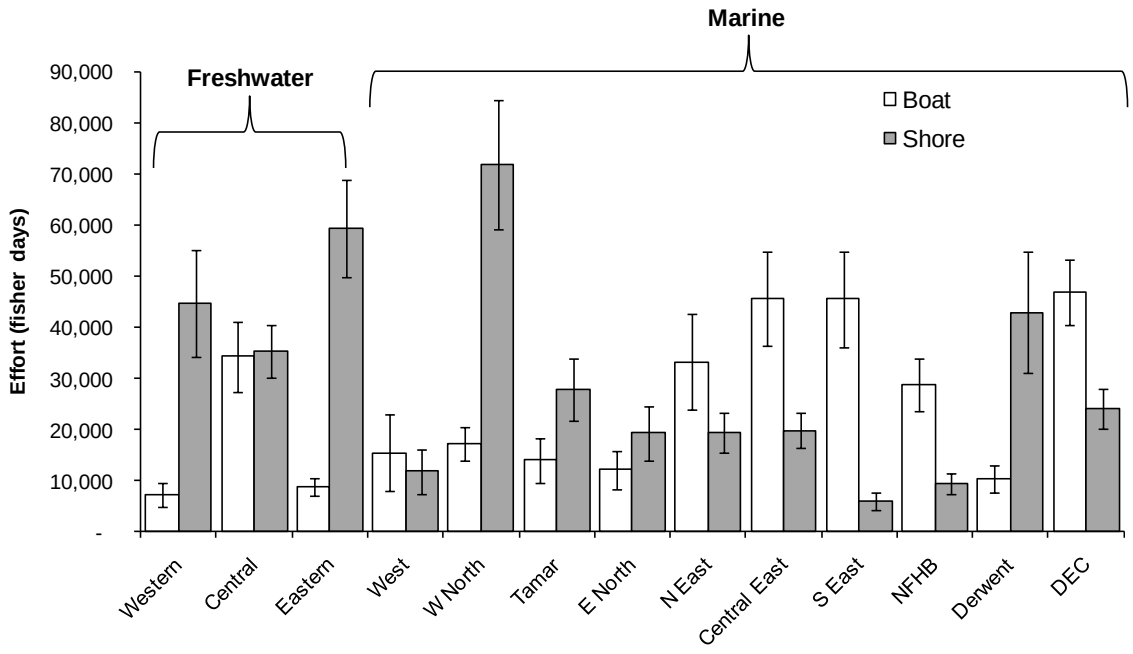

Fig. 6.7. Effort (fisher days) by fishing platform and fishing region during 2000-01 for Tasmanian residents aged five years or older. Error bars represent one standard error. NFHB Norfolk and Frederick Henry Bay; DEC D'Entrecasteaux Channel.

6.3.4 Catch

In recreational fisheries, catches can be split into retained (harvested) and released/discarded components. The harvested portion may be used for a range of purposes including consumption or as bait, whereas fish may be released because of regulation (e.g. size and/or bag limits), ethical reasons, undesirability of the species, and so on.

Re-analysed catch data for Tasmania are presented in Table 6.4 and estimated harvest weights compared with commercial production for the same period are presented in Table 6.5. The recreational harvest represented the major component (>50%) of the combined catch for several of scalefish species also taken commercially in Tasmania; namely flathead, mullet, cod, flounder, bastard trumpeter, barracouta, jackass morwong and silver trevally.

Catch (and effort) have been disaggregated by water body type (Table 6.6) and method (Table 6.7) to reveal fishery detail that is not apparent in the summary data. This analysis identified the importance of trout in the freshwater fisheries, and both flathead and Australian salmon in estuarine and inshore waters. By method it was evident that a wide variety of scalefish were taken by line fishing; trout (freshwater), flathead and Australian salmon (marine) being the dominant species. By contrast, trumpeter were the main species harvested by gillnet, dive catches were mainly comprised of rock lobster and abalone and rock lobster was the major catch taken by potting.

Table 6.4: Estimated annual catch (total, kept and released numbers) and percentage released/discarded for key species during 2000-01, based on Tasmanian residents aged five years or older. SE is standard error; + indicates value <1000; values in bold indicate relative standard error > 40%, values in italics indicate that fewer than 30 households recorded catches of the species/species group.

	Total		Kept		Released		%
	Number	**SE**	**Number**	**SE**	**Number**	**SE**	**released**
Trout	253,033	45,719	180,402	30,091	72,631	19,451	*28.7*
Atlantic salmon	13,981	4,133	13,227	4,096	+		*5.4*
Redfin	***16,219***	***8,166***	***8,896***	***7,292***	***7,323***	***3,543***	*45.2*
River blackfish	*9,618*	*2,967*	*7,282*	*2,561*	***2,336***	***1,091***	*24.3*
Australian salmon	374,043	78,528	268,262	71,782	105,781	20,383	*28.3*
Barracouta	29,114	6,042	23,177	5,478	5,937	2,048	*20.4*
Black bream	56,607	16,210	33,376	11,385	23,231	5,779	*41.0*
Blue warehou	*19,935*	*6,187*	*18,694*	*5,717*	***1,241***	***708***	*6.2*
Cod	98,735	21,176	59,892	17,433	38,843	8,408	*39.3*
Eel	9,083	2,163	6,903	1,902	2,180	716	*24.0*
Flathead	1,788,116	180,405	1,158,110	140,348	630,006	62,549	*35.2*
Flounder	54,103	13,423	50,241	12,987	**3,862**	**1,916**	*7.1*
Garfish	*17,753*	*5,110*	*14,869*	*4,628*	***2,885***	***1,715***	*16.2*
Gurnard	93,074	22,787	28,902	10,470	64,173	14,232	*68.9*
Jack mackerel	28,002	7,954	17,931	4,847	10,072	3,725	*36.0*
Jackass morwong	34,820	12,457	30,099	11,970	**4,721**	**2,042**	*13.6*
Leatherjacket	40,790	13,304	18,174	6,200	22,616	7,557	*55.4*
Mullet	108,666	27,073	76,162	23,968	32,503	8,691	*29.9*
Pike	11,313	3,357	9,622	3,185	**1,691**	**847**	*14.9*
Silver trevally	40,058	11,826	16,494	3,883	**23,564**	**10,166**	*58.8*
Trumpeter	51,669	13,887	46,186	12,067	**5,483**	**2,208**	*10.6*
Tuna	*7,112*	*2,461*	*6,839*	*2,338*	+		*3.9*
Whiting	12,653	3,796	7,490	2,414	**5,163**	**2,340**	*40.8*
Wrasse	70,486	13,636	20,431	5,699	50,055	11,944	*71.0*
Scalefish, other	61,402	17,523	23,712	9,458	37,691	12,894	*61.4*
Small baitfish	***6,465,619***	***5,950,703***	***6,210,329***	***5,943,857***	***255,291***	***242,899***	*3.9*
Sharks & rays	43,928	8,653	**8,128**	**3,424**	**35,800**	**7,371**	*81.5*
Rock lobster	143,561	23,006	76,385	13,791	67,176	12,148	*46.8*
Crustaceans, other	18,092	7,198	**13,361**	**6,943**	**4,731**	**1,358**	*26.1*
Southern calamari	26,977	7,826	25,344	7,309	**1,633**	**1,143**	*6.1*
Gould's squid	11,586	3,981	11,327	3,976	+		*2.2*
Cephalopod, other	***6,691***	***3,029***	+		***5,925***	***3,013***	*88.5*
Abalone	101,589	34,314	100,076	34,217	**1,513**	**669**	*1.5*
Scallop Bivalve, other	***114,749***	***53,531***	***113,581***	***53,326***	***1,168***	***804***	*1.0*
Other taxa	***39,120***	***16,281***	***31,940***	***12,987***	***7,180***	***6,402***	*18.4*

Table 6.5: Annual harvest (numbers), average weight and estimated harvest weight for key species taken by recreational fishers in Tasmania during 2000-01, based on Tasmanian residents aged five years or older, compared with commercial production in Tasmania. Commercial finfish catch data are based on General Fishing logbook returns for May 2000-April 2001, inclusive.

na not available; [A] based on limited data; [B] based on 1997-98 creel survey data; [C] other data sources utilised.

	Recreational					
Species	**Harvest (No.)**	**Av. weight (kg)**	**Estimated harvest (tonnes)**	**Commercial catch (tonnes)**	**Combined catch (tonnes)**	**% recreational**
Flathead	1,158,110	0.26	301.1	63.4	364.5	82.6
Australian salmon	268,262	0.35	93.9	485.0	578.9	16.2
Trout	180,402	na		-		
Mullet	76,162	0.27	20.6	13.7	34.3	60.0
Cod	59,892	0.47	28.2	4	32.2	87.6
Flounder	50,241	0.30[B]	15.1	10.5	25.6	58.9
Black bream	33,376	0.64	21.4	0	21.4	100.0
Bastard trumpeter	32,253	1.27	41.0	26.2	67.2	61.0
Jackass morwong	30,099	1.18	35.5	13.7	49.2	72.2
Gurnard	28,902	na		7.8	7.8	
Barracouta	23,177	1.93	44.7	15.1	59.8	74.8
Wrasse	20,431	0.59	12.0	88.4	100.5	12.0
Leatherjackets	18,174	0.44	8.0	16.7	24.7	32.4
Blue warehou	18,694	0.89	16.6	36.3	52.9	31.4
Jack mackerel	17,931	0.20	3.6	8.6	12.2	29.4
Silver trevally	16,494	0.28[A]	4.6	1.6	6.2	74.3
Garfish	14,869	0.12[A]	1.8	81.4	83.2	2.1
Striped trumpeter	13,933	2.20[B]	30.7	49.6	80.3	38.2
Atlantic salmon	13,227	na			-	
Pike	9,622	na		12.5	12.5	0.0
Redfin	8,896	na		-		
Sharks & rays	8,128	na		na		
Whiting	7,490	0.11	0.8	42.5	43.3	1.9
River blackfish	7,282	na		-		
Eels	6,903	na			-	
Tuna	6,839	3.56[A]	24.4	na		
Southern calamari	25,344	0.60	15.2	76.6	91.8	16.6
Gould's squid	11,327	0.50 [C]	5.7	39.7	45.4	12.5

Table 6.6: Annual recreational effort (fishers and fisher days) and harvest (numbers) of key species by water body type during 2000-01 based on Tasmanian residents aged five years or older.
SE is standard error; + indicates value <1000;values in bold indicate relative standard error >40%, values in italics indicate that fewer than 30 households recorded catches of the species/species group

	Lake		River		Estuary		Inshore		Offshore	
	Number	**SE**	**Number**	**SE**	**Number**	**SE**	**Number**	**SE**	**Number**	**SE**
Effort										
Fishers	23,814	2,003	19,507	1,910	47,663	2,972	73,504	3,457	3,534	754
Fisher days	123,991	14,975	66,720	9,729	161,428	16,103	348,596	24,862	5,728	1,332
Catch										
Trout	137,507	27,671	37,491	7,285	*4,939*	*1,772*	+			
Atlantic salmon	+		+		***2,559***	***1,042***	*10,282*	*3,625*		
Redfin	***8,673***	***7,291***	+							
River blackfish	***3,950***	***2,276***	*3,332*	*1,178*						
Australian salmon					93,974	27,087	173,677	50,236	+	
Barracouta					***3,454***	***2,200***	17,654	4,595	***2,069***	***1,997***
Black bream					28,137	10,821	*5,239*	*1,588*		
Blue warehou					***3,141***	***2,939***	*15,450*	*4,862*	+	
Cod					***33,775***	***13,464***	22,499	5,975	***3,618***	***1,813***
Eel	+		*5,248*	*1,702*	+		+			
Flathead					125,655	22,087	1,027,075	137,901	***5,379***	***2,388***
Flounder					***11,329***	***5,921***	38,912	10,427		
Garfish					***2,167***	***1,593***	*12,702*	*4,299*		
Gurnards					+		21,079	6,667	***7,537***	***4,720***
Jack mackerel					***5,017***	***2,509***	12,914	4,036		
Jackass morwong					***2,944***	***1,881***	16,699	5,840	***10,455***	***7,786***
Leatherjacket					***1,598***	***846***	16,454	6,143	+	
Mullet					21,097	6,495	**55,065**	**22,993**		
Pike					***2,031***	***1,334***	*7,591*	*2,889*		
Silver trevally					*6,636*	*2,272*	*8,391*	*2,243*	***1,467***	***1,437***
Trumpeter					+		39,694	10,397	***6,419***	***3,330***
Tuna							***1,692***	***713***	***5,146***	***2,233***
Whiting					+		*7,086*	*2,406*		
Wrasse					***1,658***	***877***	18,773	5,616		
Scalefish, other	+		+		***2,317***	***1,416***	**20,149**	**9,166**	+	
Small baitfish	***1,485***	***1,459***	***6,150,877***	***5,943,675***	***54,961***	***45,035***	***3,005***	***2,225***		
Sharks & rays					+		***6,850***	***3,118***	+	
Rock lobster					+		75,773	13,734		
Crustaceans, other	***1,055***	***1,026***			***11,825***	***6,863***	+			
Southern calamari					***3,214***	***2,805***	*22,130*	*6,767*		
Gould's squid					***3,611***	***2,934***	*7,716*	*2,280*		
Cephalopod, other							+			
Abalone					+		99,858	34,217		
Bivalve, other					***69,662***	***50,263***	*43,918*	*15,023*		
Other taxa					***14,784***	***10,381***	***17,156***	***7,134***		

Table 6.7: Annual recreational effort (fisher days) and harvest (numbers) of key species by fishing method during 2000-01, based on Tasmanian residents aged five years or older.
SE is standard error; + indicates value <1000, values in bold indicate relative standard error >40%, values in italics indicate that fewer than 30 households recorded catches of the species/species group.

	Line		Gillnet		Dive		Lobster pot		Other	
	Number	**SE**	**Number**	**SE**	**Number**	**SE**	**Number**	**SE**	**Number**	**SE**
Effort										
Fisher days	615,985	35,020	34,928	6,810	22,081	4,900	48,359	9,609	22,216	4,647
Catch										
Trout	180,203	30,091	+		+					
Atlantic salmon	***7,516***	***3,761***	*5,710*	*1,625*						
Redfin	***8,896***	***7,292***								
River blackfish	*7,282*	*2,561*								
Australian salmon	254,078	71,173	***7,834***	***3,259***					***6,351***	***3,879***
Barracouta	23,177	5,478								
Black bream	32,437	11,377	+						+	
Blue warehou	***9,368***	***4,202***	***9,325***	***3,824***						
Cod	55,238	17,198	***4,538***	***2,631***			+			
Eel	*6,549*	*1,885*							+	
Flathead	1,151,487	139,984	***3,341***	***1,392***					***3,282***	***1,392***
Flounder	+		***9,960***	***7,185***	***1,329***	***869***			***38,792***	***10,746***
Garfish	***6,166***	***3,058***							***8,702***	***3,458***
Gurnards	26,634	10,349	***2,268***	***906***						
Jack mackerel	16,899	4,616	+				+		+	
Jackass morwong	13,633	3,573	***16,332***	***10,763***					+	
Leatherjacket	*5,228*	*1,349*	***12,020***	***5,986***	+				+	
Mullet	37,192	9,987	***11,436***	***6,682***					***27,534***	***20,363***
Pike	*9,579*	*3,184*	+							
Silver trevally	*12,744*	*3,497*	***3,641***	***1,695***	+					
Trumpeter	*10,818*	*3,588*	*34,310*	*10,333*	+		+		+	
Tuna	*6,839*	*2,338*								
Whiting	*7,324*	*2,409*							+	
Wrasse	12,863	3,373	***7,357***	***4,604***	+				+	
Scalefish, other	*11,661*	*3,917*	***2,741***	***1,145***	+		+		***9,082***	***8,474***
Small baitfish									***6,210,329***	***5,943,857***
Sharks & rays	***7,276***	***3,397***	+							
Rock lobster					33,098	10,904	43,211	8,259	+	
Crustaceans, other	***1,988***	***1,315***	+		+		+		***11,069***	***6,824***
Southern calamari	*21,664*	*6,461*							***3,680***	***3,055***
Gould's squid	11,247	3,975	+						+	
Cephalopod, other	+				+		+		+	
Abalone					100,020	34,217			+	
Bivalve, other	+								***113,581***	***53,326***
Other taxa			+		***5,649***	***5,649***			***11,662***	***11,662***

6.3.5 Key species example

Flathead was the dominant species taken by recreational fishers during 2000-01 and has been used here an example of information relevant to describing the status of the fishery for this species. Similar analyses are possible for each of the main species groups.

Southern sand flathead (*Platycephalus bassensis*) and tiger flathead (*Neoplatycephalus richardsoni*) were the dominant species of flathead taken in Tasmanian waters. Of the total catch numbers, 79% (1,422,782; SE 145,166) were identified as southern sand flathead and just 6% (114,643; SE 39,621) as tiger flathead. The balance (250,691; SE 51,956) were reported as unspecified flathead.

The vast majority (>80%) of the catch was derived from the Central East and South East coasts, with the D'Entrecasteaux Channel and Norfolk-Frederick Henry Bay regions particularly significant (Figure 6.8A). By comparison, North coast catches, including the Tamar, were comparatively low while West coast catches were insignificant. About 35% of all flathead caught were released or discarded (Figure 6.8B). Boat based fishing accounted for the vast majority (92%) of the catch (Figure 6.8C), and virtually all of the catch was taken by line fishing (Figure 6.8D). Flathead catches were concentrated in inshore coastal waters with relatively small numbers also taken from estuarine and, to a lesser extent, offshore waters (Figure 6.8E).

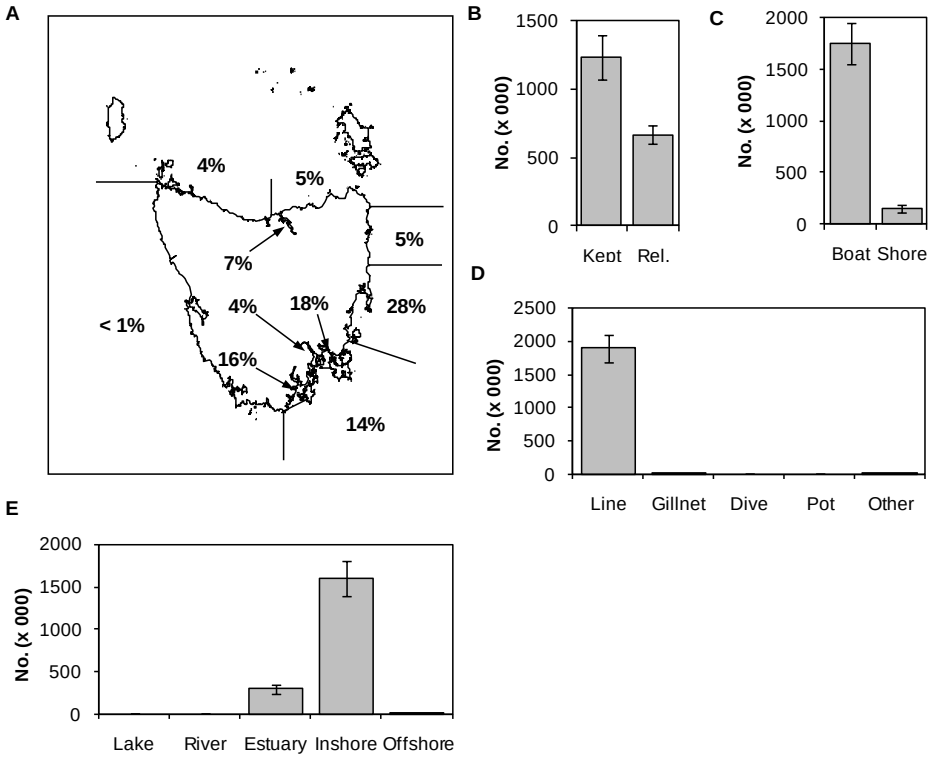

Fig. 6.8 Characteristics of the recreational fishery for flathead in Tasmania during 2000-01: A) proportion (%) of the total catch (numbers) by fishing region; B) total numbers kept and released; C) total catch (numbers) by boat and shore based fishing activities; D) total catch (numbers) by fishing method; and E) total catch (numbers) by water body fished

6.3.6 Conclusion

As demonstrated in this section, the *RecSurvey* package provides a platform from which analyses can be undertaken with flexibility, enabling disaggregation of data at a variety of scales. Importantly, users are able to undertake analyses that are relevant to management and assessment needs, producing estimates with their associated uncertainty.

For further examples of the type of outputs that have been produced using the *RecSurvey* package refer to Lyle *et al.* (2009) and Jones (2009).

BENEFITS

The primary benefit of this project has been the development of a statistically robust analytical framework, the *RecSurvey* package, which can be applied to re-analyse existing NRFS and future recreational fishing survey data. Importantly, the package enables users to specify analyses that are relevant to their needs, particularly in relation to data disaggregation, and provide estimates with associated statistical uncertainty. This latter point is especially important for interpreting the significance of outputs, noting that the survey methodology was designed to provide a big-picture perspective of recreational fishing. The *RecSurvey* package is also flexible enough for users to make decisions about how and what assumptions are used in the calibration and adjustment processes, enabling the sensitivity of estimates to differing assumptions to be assessed.

The package has been applied in the re-analysis of NRFS data for Tasmania and South Australia, and in the analysis of recently completed state-wide surveys in each of these states.

At a workshop held in October 2009, researchers from the CSIRO, and state fisheries agencies of New South Wales, Northern Territory, Victoria and Western Australia, along with government (DPIPWE) and recreational (RecFish Research and TARFish) stakeholders were provided with a practical demonstration of the package. This workshop generated considerable interest, with researchers expressing interest in a full re-analysis of NRFS data relating to their jurisdiction and in applying the package to the analysis of future recreational fishing surveys.

The key output arising from this project, namely the *RecSurvey* package (including functions and help files), an example database, worked data example and manual will be distributed electronically to each of the Australian fisheries research agencies and made available for download from the TAFI and FRDC websites.

FURTHER DEVELOPMENT

Opportunities for further development of the package include:

- alternative measures of error estimation, e.g. data re-sampling techniques, such as boot-strapping.
- simulation procedures to assess relationships between screening sample size and stratification and reliability of estimates associated with disaggregated data, such as for specific fisheries.
- incorporation of other sources of data, such as catch size composition information, that could be used in the estimation of harvest weights.
- formalised sensitivity analyses associated with different assumptions relating to calibration and adjustment.

The suggested database structure could also be refined to provide a more suitable input database model.

PLANNED OUTCOMES

The analytical module will ensure robust statistical approach is applied to the analysis of recreational survey data and, importantly, that the analyses are transparent and repeatable. The primary outcomes of reliable information for the recreational sector will include improved management of fishery resources and informed resource allocation through the accurate quantification of key user group's impacts.

Beneficiaries include resource managers along with recreational and commercial stakeholders through the provision of robust information about the recreational component of the fishery. Researchers tasked with conducting large-scale, off-site recreational surveys will benefit through access to a flexible and statistically robust analytical package that has been developed in a well supported and widely used statistical computing language.

CONCLUSION

Since the NRFS there have been several advances in the theory of calibration for multi-phase designs which have been applied to develop a statistical package to analyse survey data. The *RecSurvey* package has been implemented in the statistical computing language R and provides a flexible and transparent platform specifically designed for the phone-diary survey methodology. In addition to providing a step by step guide to analysis, with the capability for users to make decisions about what assumptions are applied in the calibration processes, the package provides recommendations on database structure and queries necessary to prepare data for analysis. A number of example analyses are provided to indicate the capability of the package to provide disaggregated data ouputs. For instance, catch and effort data can be readily disaggregated by fishing method, platform, region, target species, or combinations of these factors, with estimates provided with associated uncertainty.

Key NRFS data for Tasmania and South Australia were re-analysed using the *RecSurvey* package and compared with original estimates. Participation rates by region of residence and age group did not differ significantly between the original and re-analysed estimates. Furthermore, state-wide harvest estimates for the major species were not significantly different, indicating that the original analyses were generally robust. Detailed re-analysis of NRFS data for Tasmania was also undertaken and serves as an example of the type of outputs that can be achieved using the analytical package.

Several fisheries agencies have expressed interest in the application of the package, which has already been used to analyse recently completed state-wide surveys in Tasmania and South Australia and is expected to be used to analyse surveys in the Northern Territory and Queensland within the near future.

LITERATURE CITED

Bradford, E. (1998) National marine recreational fishing 1996: Scaling the diary survey results to give the total recreational harvest. NZ National Institute of Water and Atmospheric Research (NIWA) Technical Report 17.

Collett, D. (1991). *Modelling Binary Data*. Chapman Hall.

Crawley, M.J. (2007). *The R Book*. Wiley.

Dalgaard, P. (2000). *Introductory Statistics with R*. Springer, New York.

Estevao, V. and Särndal, C. E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2):127–147.

Higgs, J.B. (1999) Experimental catch estimates for Queensland residents. RFISH Technical Report No. 2. Results from the 1997 diary round. Queensland Fisheries Management Authority.

Higgs, J.B. (2001) Catch estimates for Queensland residents. RFISH Technical Report No. 3. Results from the 1999 diary round. Queensland Fisheries Service.

Henry, G.W. and Lyle, J.M. (2003) The national recreational and indigenous fishing survey. Final Report to the Fisheries Research and Development Corporation, Project 99/158. NSW Fisheries Final Report Series No. 40, 188p.

Jones, K. (2009) The 2007/08 South Australian recreational fishing survey. PIRSA Fisheries Management Series Paper No. 54.

Lapsley, M. and Ripley, B.D. (2008). *RODBC: ODBC Database Access*. R package version 1.2-4.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19.

Lumley, T (2010) *Survey: analysis of complex survey samples*. R package version 3.22-1. http://faculty.washington.edu/tlumley/survey/

Lyle, J.M. (2005) 2000/01 survey of recreational fishing in Tasmania. Tasmanian Aquaculture and Fisheries Institute, Technical Report Series No. 24, 97p.

Lyle, J.M., Coleman, A.P.M, West, L., Campbell, D., and Henry, G.W. (2002) An innovative methodology for the collection of detailed and reliable data in large-scale Australian recreational fishing surveys. In: *Recreational Fisheries: Ecological, Economic and Social Evaluation*, Pitcher, T.J., and Hollingworth, C.E. (eds). pp 207-226. Fish and Aquatic Resources Series No. 8, Blackwell Science, Oxford, UK.

Lyle, J.M, Tracey, S.R., Stark, K.E., and Wotherspoon, S. (2009). 2007-08 survey of recreational fishing in Tasmania. Tasmanian Aquaculture and Fisheries Institute report.

Pollock, K.H. (2003) Recreational angler surveys: the interaction of scale and optimal contact methods for effort and catch estimation. In Regional experience for global solutions. The proceedings for the 3rd World Recreational Fishing Conference, 21–24 May 2002, Northern Territory, Australia. Coleman, APM (ed.). Fisheries Report 67. Dept of Business, Industry and Resource Development.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Chapman Hall, second edition.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Särndal, C.E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley.

Spector, P. (2008). *Data Manipulation with R*. Springer.

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

INTELLECTUAL PROPERTY

This is not applicable to this project.

STAFF

Tasmanian Aquaculture and Fisheries Institute, University of Tasmania

Jeremy Lyle
Kate Stark

School of Mathematics and Physics, University of Tasmania

Simon Wotherspoon