Genstat®

Survey analysis

www.vsni.co.uk

VSNi

A Guide to Survey Analysis in Genstat®
(18th Edition)

by Steve Langton

Defra Environmental Observatory, 1-2 Peasholme Green, York YO1 7PX, UK.

Genstat is developed by VSN International Ltd, in collaboration with practising statisticians at Rothamsted and other organisations in Britain, Australia, New Zealand and The Netherlands.

Published by: VSN International, 2 Amberside, Wood Lane,
Hemel Hempstead, Hertfordshire HP2 4TP, UK
E-mail: info@genstat.co.uk
Website: http://www.genstat.co.uk/

First published 2009, for GenStat *for Windows* 12th Edition
This edition published 2015, for Genstat *for Windows* 18th Edition

Contents

Introduction

Surveys are widely used in many areas of modern life. Political opinion polls and the myriad of phone and postal surveys aimed at the general public spring instantly to mind. There are also vast numbers of more specialized surveys aimed at producing key facts for business, government, medical researchers and others. In addition, many scientific studies involve random sampling and may require the use of survey analysis methods.

The analysis of surveys is, in many cases, a fairly simple exercise compared to many other statistical analyses. Unfortunately that simplicity often tempts analysts to rely on unsuitable software, such as simple spreadsheet programs. Whilst these often give correct point estimates, they seldom produce valid standard errors and do not provide a means of identifying outlying or influential observations. The aim of this Guide is to show how the correct analysis can easily be achieved using Genstat's facilities for survey analysis.

Genstat can be used in two ways; the simplest, particularly for new users, is to use the menu system, and this Guide will show you how to perform all the analyses using menus. The second way is to use Genstat's own programming language, and this can be an efficient approach for many surveys since it allows the automation of repetitive tasks. The use of programming is not described in the main text, but a separate chapter introduces the principles and some key commands, whilst an Appendix gives the commands to generate all the analyses described in the main text. Those keen to learn to program in Genstat may prefer to read the programming chapter first and then refer to the Appendix whilst working through the earlier chapters.

The first stage in any survey is the design phase, but in this book we will concentrate on survey analysis, only briefly considering design issues. This should not be taken to imply that the design of a survey is not crucially important, but instead is a pragmatic decision based on the knowledge that many Genstat users will have to analyse surveys which they have not had the opportunity to design.

1 Basic principles

In this chapter we introduce some of the basic principles behind the analysis of surveys in Genstat. These principles will be illustrated using the small `Province` dataset; more realistic examples will be examined in later chapters. Analysis will use the **Single-stage Survey Analysis** menu (`SVSTRATIFIED` procedure), but the same basic principles apply to the more complex analyses available from the **General Survey Analysis** menu (`SVTABULATE` procedure).

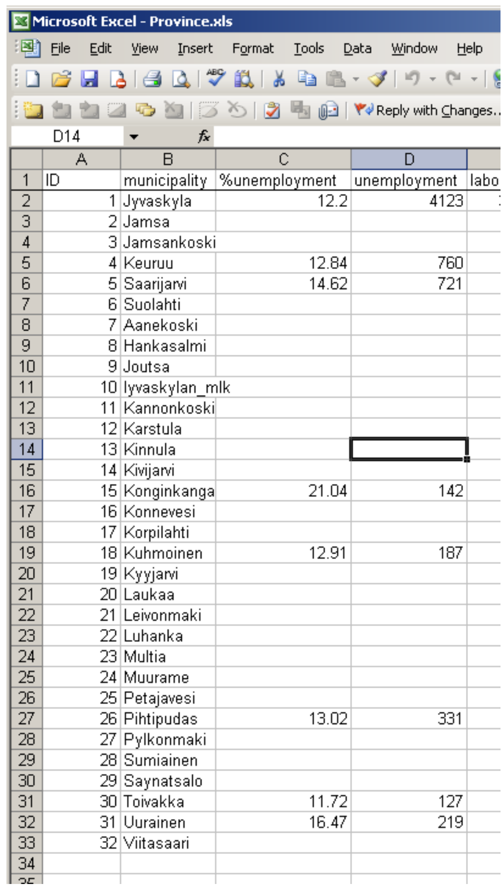
In this chapter you will learn about

- getting the data into Genstat
- how the data should be organized prior to analysis.
- identifying unusual observations, some of which may result from errors in data processing.
- defining strata and supplying strata sizes

1.1 Getting the data into Genstat

For the first example we shall use the `Province` population, taking a simple random sample of eight municipalities as shown in the Excel spreadsheet in Figure 1.1. The variables `%unemployment` and `unemployment` are shown only for the sampled municipalities, with blanks for the unsampled ones.

Excel is used for this dataset since it is one of the commonest formats used for small surveys, but Genstat can open files produced by a wide range of spreadsheet, database and statistical packages. More details can be found



	A	B	C	D	
1	ID	municipality	%unemployment	unemployment	labo
2	1	Jyvaskyla	12.2	4123	
3	2	Jamsa			
4	3	Jamsankoski			
5	4	Keuruu	12.84	760	
6	5	Saarijarvi	14.62	721	
7	6	Suolahti			
8	7	Aanekoski			
9	8	Hankasalmi			
10	9	Joutsa			
11	10	lyvaskylan_mlk			
12	11	Kannonkoski			
13	12	Karstula			
14	13	Kinnula			
15	14	Kivijarvi			
16	15	Konginkanga	21.04	142	
17	16	Konnevesi			
18	17	Korpilahti			
19	18	Kuhmoinen	12.91	187	
20	19	Kyyjarvi			
21	20	Laukaa			
22	21	Leivonmaki			
23	22	Luhanka			
24	23	Multia			
25	24	Muurame			
26	25	Petajavesi			
27	26	Pihtipudas	13.02	331	
28	27	Pylkonmaki			
29	28	Sumiainen			
30	29	Saynatsalo			
31	30	Toivakka	11.72	127	
32	31	Uurainen	16.47	219	
33	32	Viitasaari			
34					

Figure 1.1.

in the *Getting Started Guide* or by selecting **Importing data** from the on-line help.

To open the file, click on the **Excel Import Wizard** icon on the Genstat toolbar, or alternatively select **Open** from the **File** menu. The file is called `Province.xls` and can be found in the `data` subdirectory of the directory where Genstat is installed. (Alternatively, it can be found by selecting **Examples** from the **Help** menu, and then clicking on **Data Sets**, but this approach can, of course, be used only with the supplied example files). The data shown are in sheet `simple RS full pop`, and clicking on **Next** or **OK** as appropriate at the wizard's dialogue boxes to accept the default settings will successfully transfer the data into a Genstat spreadsheet window.

In general it is wise to start by calculating some simple descriptive statistics when investigating a new dataset. Looking at means, minima and maxima, and as well as graphical displays, such as scatter plots, can help identify the important features of the data. However, this example is so small that visual examination of the data is sufficient. From Figure 1.1 it is apparent that the first municipality has much higher numbers of unemployed people than the others, but that its unemployment rate is not particularly large; the number of unemployed stands out only because it has a much higher population than the other sampled regions. In terms of percentages, the distribution appears rather skewed, with the majority of municipalities having around 12% unemployment, but three of the eight having higher rates.

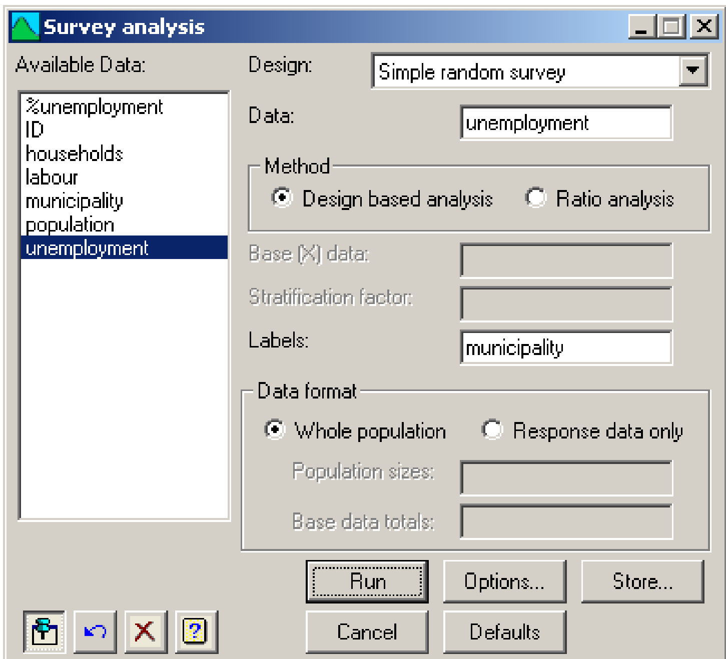


Figure 1.2.

To estimate the mean number unemployed per municipality and the total across all municipalities, we select **Survey Analysis** from the **Stats** menu, and then click on the **Single-stage Survey Analysis** sub-option. The menu shown in Figure 1.2 will open. Place the cursor in the **Data** box and double click on the variable `unemployment` to transfer it to the box. Then place the cursor in the **Labels** box and double click on the variable `municipality` to transfer it to the box. Clicking on **Run** produces the output below.

```
Survey analysis results
=======================


Data summary
------------

Y-variate (response data):  unemployment
Method:                     Design-based (expansion)
Variance method:            Conventional (Taylor series)
Deff:                       1.0000


               Total no. obs.  Imputed    Sample  Excluded  Sampling fraction
      Stratum
     All data              32       24         8         0              0.250


Estimated totals with 95% confidence limits
-------------------------------------------

                  Total      s.e.    %r.s.e.     Lower      Upper
      Stratum
     All data     26440     13282       50.2     -4968      57848


Estimated means with 95% confidence limits
------------------------------------------

                  Means      s.e.    %r.s.e.     Lower      Upper
      Stratum
     All data     826.2     415.1       50.2    -155.2       1808
```

The default output shown above starts with a summary of the type of analysis and the data used. `Deff` refers to the *design effect*; i.e. the ratio of the variance under the design used to the variance under simple random sampling. Since this analysis uses simple random sampling, the design effect is exactly one. Following this there is a table of the data that have been used, with a row for each stratum if the design is stratified. It is worth checking this carefully to ensure the number of sampled observations is as expected. The column headed `Imputed` shows the number of rows for which there are no data collected for the variable analysed (i.e. rows that have a blank in column D of Figure 1.1).

The following sections show the estimated means and totals. These are estimated using the usual methods for simple random sampling. The estimate of the mean is obtained by adding up the observed unemployment totals and dividing by the number of observations:

$$y = \sum y_i / n$$

The variance of the data is the sum of the squared differences between the observations and the mean.

These equations are identical to the usual ones used in non-sampling situations, but the equation for the standard error of the mean is different, since it includes a term known as the *finite population correction* (*fpc*), which is equal to one minus the number of sampled observations (n) divided by the number of units in the full population (N):

$$fpc = (1 - n / N).$$

The *fpc* is required because we are making inferences about a population of known size, N, whereas in ordinary estimation we are interested in a hypothetical infinite population. Note that, if we sample all the units in the population (so that n=N), the *fpc* equals zero, and the standard error of the mean is also zero. This is because we then know the size of the mean exactly and there is no sampling error associated with its estimation. Conversely, if n is very small compared to N, the *fpc* becomes very close to 1, and the equation for the standard error of the mean becomes similar to the standard version.

The figure labelled `%r.s.e.` is the *relative standard error* of the mean, and is simply the standard error of the mean (or any other statistic) expressed as a percentage of its estimate (in this case 415.1 / 826.2 * 100 = 50.2%). The relative standard error is often referred to as the *coefficient of variation* (%cv), but the latter term can be ambiguous since it is also used to describe the standard deviation of observations expressed as a percentage of the mean.

Finally, 95% confidence limits are shown for both the mean and the total. Limits calculated in this way can be expected to contain the true value 95% of the time. They are calculated using a t-statistic with 7 degrees of freedom, one less than the number of sampled units. If you wish to check the calculation, the appropriate value of the t-distribution can be found by selecting **Probability Calculations** from the **Data** menu. Notice that in this case, the lower limit is less than zero; simple random sampling with a sample size of eight is clearly not an effective sampling scheme for this dataset.

1.2 Saving results

In many cases the results in the output window will be sufficient, but often you will want to save the estimates in Genstat data structures. This might be to allow further analysis, or maybe to change the units in which they are measured. With large datasets containing many variables, you may want to save the estimates so that they can all be concisely displayed in the same spreadsheet. To save the estimates click on the **Store** button on the survey analysis menu. You will see the menu shown in Figure 1.3. In this case we are going to save the estimates of the totals and their standard errors. Click on the small boxes and the rectangles on the right become enabled, thus allowing us to type suitable names for saving them. These names can contain any of the 26 letters, plus % and _, and they are case sensitive. The numbers 0-9 can be used, but not at the start of the name. For more details see Section 1.4.3 of the *Syntax and Data Management Guide*, available from the help menu. In Figure 1.3 the **Display in Spreadsheet** box is also ticked; this is sensible when the results need to be saved, or cut and pasted to another package.

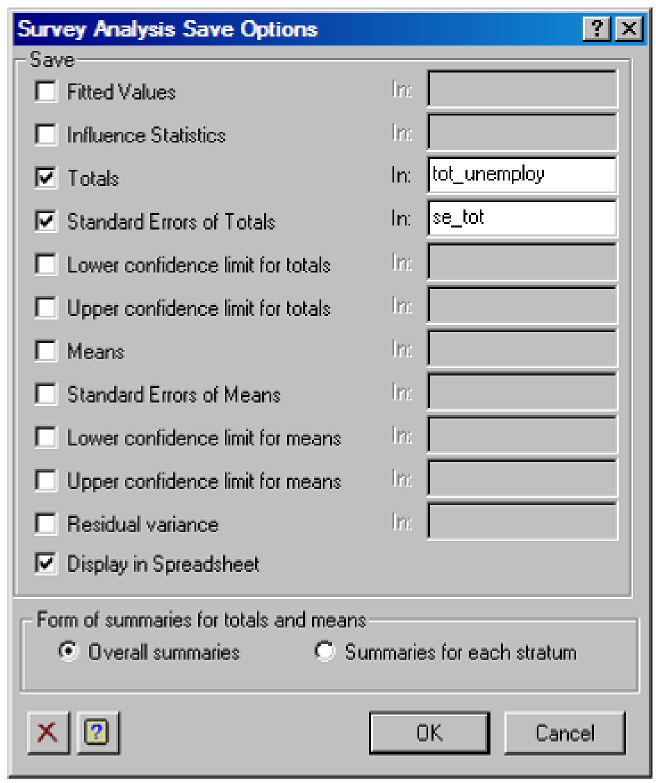


Figure 1.3.

1.3 Detecting outliers

The design-based analyses described above make no assumptions about the distribution of the data, in contrast to many other statistical techniques which assume a particular underlying distribution, often a Normal distribution. However, this does not mean that the results are unaffected by the presence of small numbers of unusually large or small values, often known as outliers. When extreme outliers do occur, it is important to be aware of them, because they may indicate that the analysis cannot be relied upon. In addition, they sometimes arise because of errors in data recording or processing, and so it is good practice to investigate any particularly large outliers to ensure that they are not the result of mistakes.

The methods provided for outlier detection can be seen in the **Design based Survey Analysis Options** menu (Figure 1.4), which can be opened by clicking on the **Options** button in Figure 1.2. If the **Scatterplot** box is ticked, a graphics window is produced containing a plot of the response variable against either the stratum number or, if the X parameter is set in order to carry out ratio analysis, a scatter plot of the response variable against X. These graphs are plotted on the log scale as survey data are frequently skew, which can make graphs on the natural scale uninformative.

With the current dataset, the scatter plot is not particularly informative, since there are so few data points and only one stratum (Figure 1.5). Notice how by clicking on the *data information tool* (highlighted on the toolbar) and then positioning the mouse over a point, information about the point is displayed. With large datasets this can be handy when trying to locate an observation in the data spreadsheet. More usefully with small datasets clicking on the **Influence** tick box (Figure 1.4) displays influence statistics. These

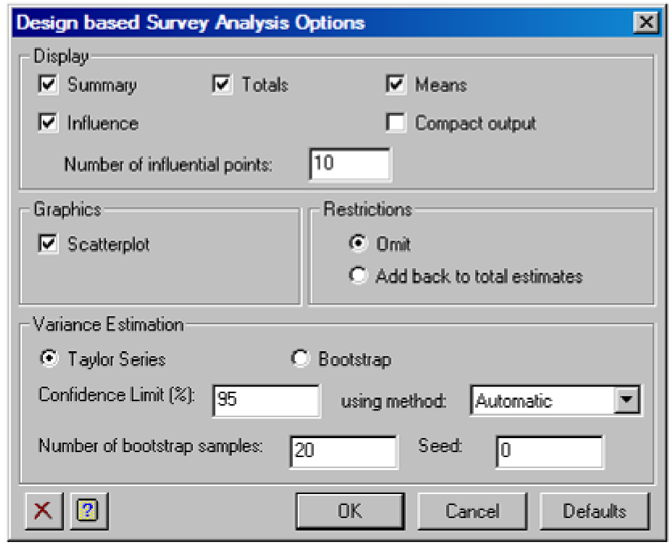


Figure 1.4.

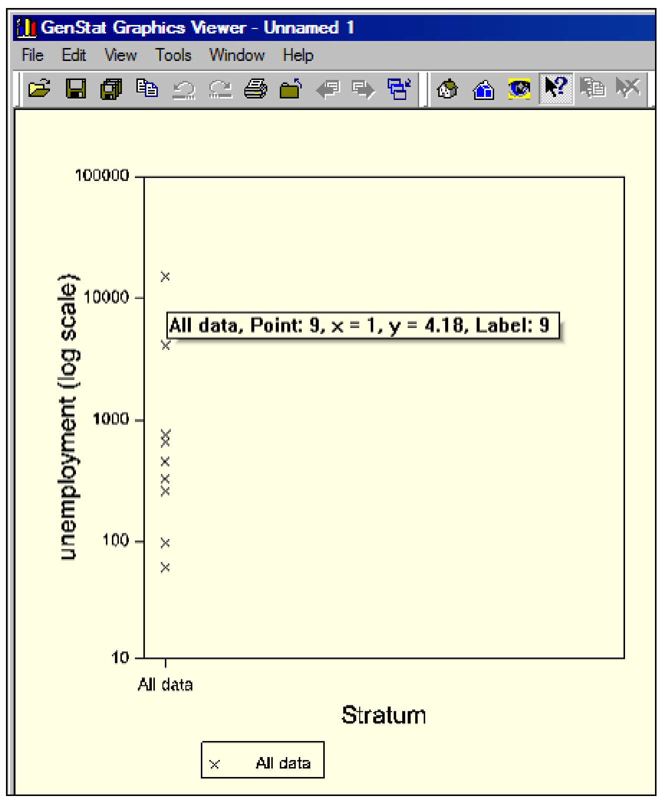


Figure 1.5.

are defined as the percentage change in the estimate of the grand total when the observation is replaced by a missing value (i.e. treated as if it was not sampled). By default the 10 highest observations are shown, but in this case only eight were sampled. For larger datasets this number can be increased using the options menu.

```
10 points with highest influence
--------------------------------

Unit          Stratum              Y           X  %influence
Jyvaskyla      All data        4123.0          *       57.00
Keuruu         All data         760.0          *        1.15
Saarijarvi     All data         721.0          *        1.82
Konginkangas   All data         142.0          *       11.83
Kuhmoinen      All data         187.0          *       11.05
Pihtipudas     All data         331.0          *        8.56
Toivakka       All data         127.0          *       12.09
Uurainen       All data         219.0          *       10.50

Percentage influence is calculated as the percentage change
in the grand total when each sampled observation is omitted.
```

Notice that in this case the figure from Jyvaskyla has an influence statistic of over 50%, confirming that these results should be treated with considerable caution.

1.4 Practical

This exercise involves verifying the influence statistic for Jyvaskyla by reanalysing the data without this observation. Start by saving the total for the full analysis as described above. Then go to the spreadsheet and form a copy of the unemployment column (select the **Column** option on the **Spread** menu, then click on **Duplicate**). Then delete the value in row one and repeat the analysis with this new variable. Finally calculate the influence using **Calculate** from the **Data** menu, as shown in Figure 1.6.

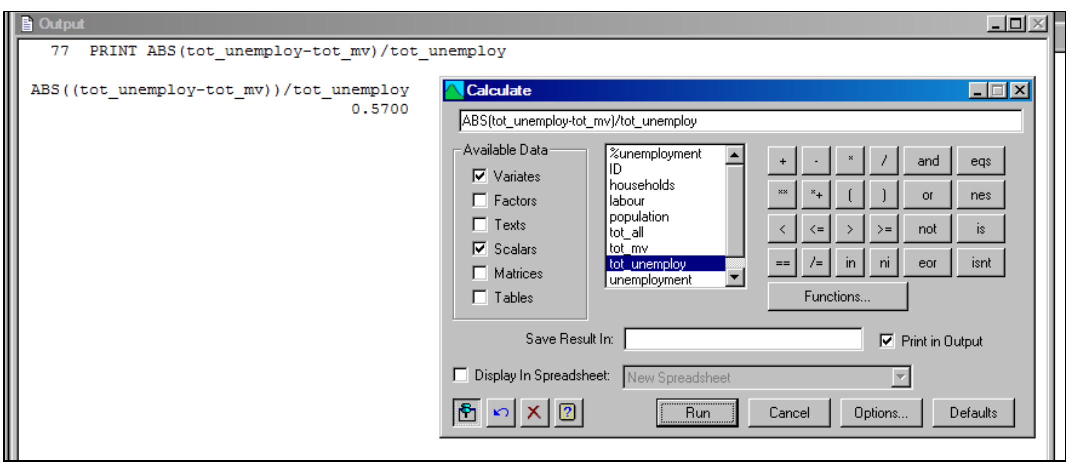
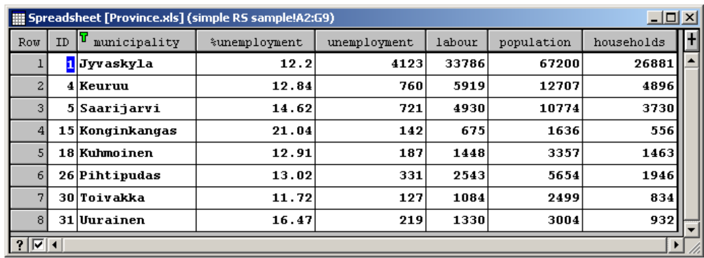


Figure 1.6.

1.5 Analysis with response data only

The analyses described so far in this chapter have been based on a dataset with one row for each unit in the population (in this case each municipality in the province), even if they were not sampled, or did not respond. This way of presenting the data avoids the problems associated with specifying the design, and is a particular advantage, as we shall see in the next chapter, for estimating totals by ratio analysis. However, it is not always a sensible or practical approach, particularly for very large datasets. In this section we will consider the alternative layout, where there is a row in the dataset only for those units that provide data for the final analysis, which generally means those units that have been sampled and have co-operated with the survey.

Figure 1.7 shows the Province data in this layout. The Genstat spreadsheet shown was created by loading sheet simple RS sample of Province.xls using the Excel wizard (see Section 1.1). To analyse the data in this format, we once again select **Survey Analysis** from the **Stats** menu, and then click on the **Single-stage Survey Analysis** sub-option. However, this time we click on the button for **Response data only** under **Data format** (Figure 1.8). The population sizes box then becomes enabled, allowing us to enter the total number of units in the population (i.e. the total number of rows in the full dataset



Spreadsheet [Province.xls] (simple RS sample!A2:G9)

Row	ID	municipality	%unemployment	unemployment	labour	population	households
1	1	Jyvaskyla	12.2	4123	33786	67200	26881
2	4	Keuruu	12.84	760	5919	12707	4896
3	5	Saarijarvi	14.62	721	4930	10774	3730
4	15	Konginkangas	21.04	142	675	1636	556
5	18	Kuhmoinen	12.91	187	1448	3357	1463
6	26	Pihtipudas	13.02	331	2543	5654	1946
7	30	Toivakka	11.72	127	1084	2499	834
8	31	Uurainen	16.47	219	1330	3004	932

Figure 1.7

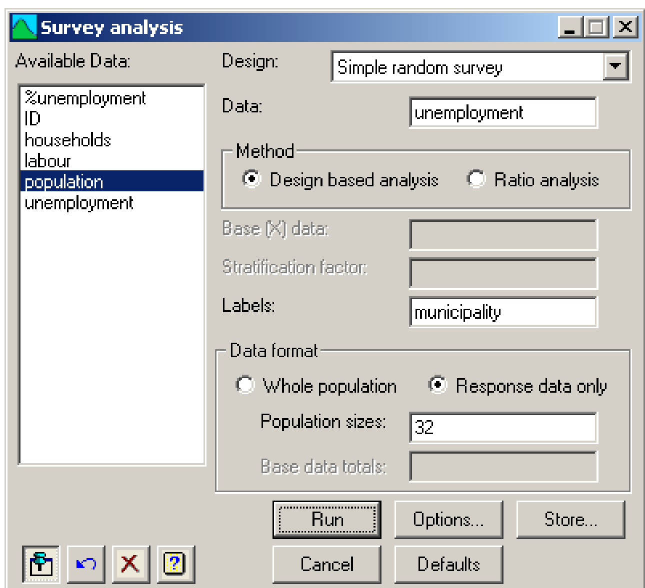


Figure 1.8

including unsampled municipalities, Figure 1.3). The analysis produced when the **Run** button is clicked is shown below; it is identical to the results obtained in Section 1.1 above.

```
Survey analysis results
=======================

Data summary
------------

Y-variate (response data):  unemployment
Method:                     Design-based (expansion)
Variance method:            Conventional (Taylor series)
Deff:                       1.0000


               Total no. obs.  Imputed    Sample  Excluded  Sampling fraction
      Stratum
     All data              32        24         8         0              0.250


Estimated totals with 95% confidence limits
-------------------------------------------

                  Total      s.e.  %r.s.e.    Lower     Upper
      Stratum
     All data     26440     13282      50.2     -4968     57848


Estimated means with 95% confidence limits
------------------------------------------

                  Means      s.e.   %r.s.e.     Lower     Upper
      Stratum
     All data     826.2     415.1      50.2    -155.2      1808
```

1.6 Stratified random samples – factors and tables

So far, all the analyses have been based on simple random sampling, that is selecting units (in this case municipalities) at random with equal probability. In many cases this is not an efficient approach and so stratified random sampling is used, with different sampling probabilities in different groups (*strata*). In order to analyse stratified random sampling designs in Genstat, it is necessary to construct a *factor* to indicate which stratum each unit belongs to, and so we will commence by learning more about factors.

For those familiar with the analysis of variance in Genstat, it is important to realize that the use of the word *stratum* is very different here. The strata in a survey are essentially similar to the blocks in a randomized block design; strata in a sample survey and blocks in a randomized block experiment are both generally selected to ensure that the units within a stratum or block are more homogeneous than those in different ones. The strata in analysis of variance are more akin to the stages or levels in a multistage survey.

Figure 1.9 shows the spreadsheet created by importing sheet `stratified sample` from `Province.xls`. Most of the columns are *variates*, that is numerical structures that can take any value, including negative values. Variates can be used in a wide variety of numerical calculations and statistical routines. The municipality column has a green ‘T’ in its title bar to indicate that it is a *text*. Texts can hold any textual strings, including numerical characters, and so cannot be used for standard numerical calculations. They are principally used for labelling observations, or recording comments.

The `stratum` column has a red exclamation mark by its name and this indicates that it is a factor. Factors are numerical structures that can only take certain predefined values; for example a factor for sex might take the values ‘male’ or ‘female’. Factors are essentially numerical structures, but they may be assigned textual *labels* to aid interpretation of the output (see Section 2.2). In this case there

Spreadsheet [Province.xls] (stratified sample!A2:H9)

Row	ID	municipality	stratum	%unemployment	unemployment	labour	population	households
1	1	Jyvaskyla	1	12.2	4123	33786	67200	26881
2	2	Jamsa	1	11.07	666	6016	12907	4663
3	4	Keuruu	1	12.84	760	5919	12707	4896
4	6	Suolahti	1	15.12	457	3022	6159	2389
5	21	Leivonmaki	2	10.65	61	573	1370	545
6	25	Petajavesi	2	15.08	262	1737	3800	1352
7	26	Pihtipudas	2	13.02	331	2543	5654	1946
8	27	Pylkonmaki	2	17.98	98	545	1266	473

Figure 1.9

are only two strata, and no textual labels have been defined, so only the values 1 and 2 (known as the *levels* of the factor) are allowed in the column. A factor can be created in a number of ways in the Genstat menu system.

- When using the Excel wizard, the final menu box, **Select Columns to Convert to Factors** suggests columns for conversion to factors. Highlighting the relevant column and clicking on the **Factor** button ensures that it becomes a factor.
- In the spreadsheet window, right clicking on the column gives a list of options, one of which is **Convert to Factor**

- From the **Spread** menu with the cursor in the column, select the **Factor** option and then **Convert to**

Figure 1.10 shows how this data layout can be analysed by selecting **Stratified random survey** in the **Design** drop-down list box. Note how, with the cursor in the **Stratification factor** box, the **Available Data** box only lists `stratum`, since this is the only factor in the spreadsheet. Since the spreadsheet only contains response data, the population size of each stratum must be specified. When there is more than a single stratum, these must be specified in a Genstat structure and, to minimize the risk of associating numbers with the incorrect stratum, it is best to use a *table*.

To create the table of population sizes, select the **New** option and **Create** sub-option from the **Spread** menu. Then click on the table item and tick the **Create from Existing Factors** box (left of Figure 1.11). At the next menu, click `stratum` across to the **Selected Factors** box (top right of Figure 1.11). Once the new table spreadsheet is created (bottom right), the table can be given a more appropriate name (make a right mouse click and select **Rename**) and the correct values can be entered (7 for stratum 1 and 25 for stratum 2).

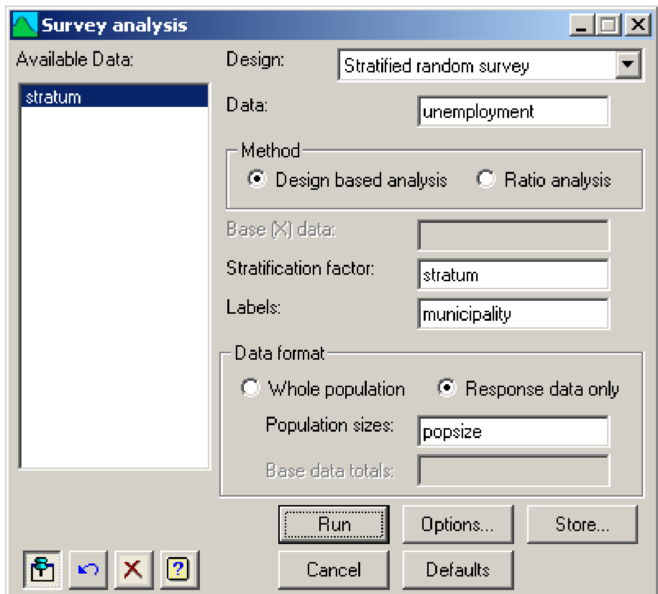


Figure 1.10

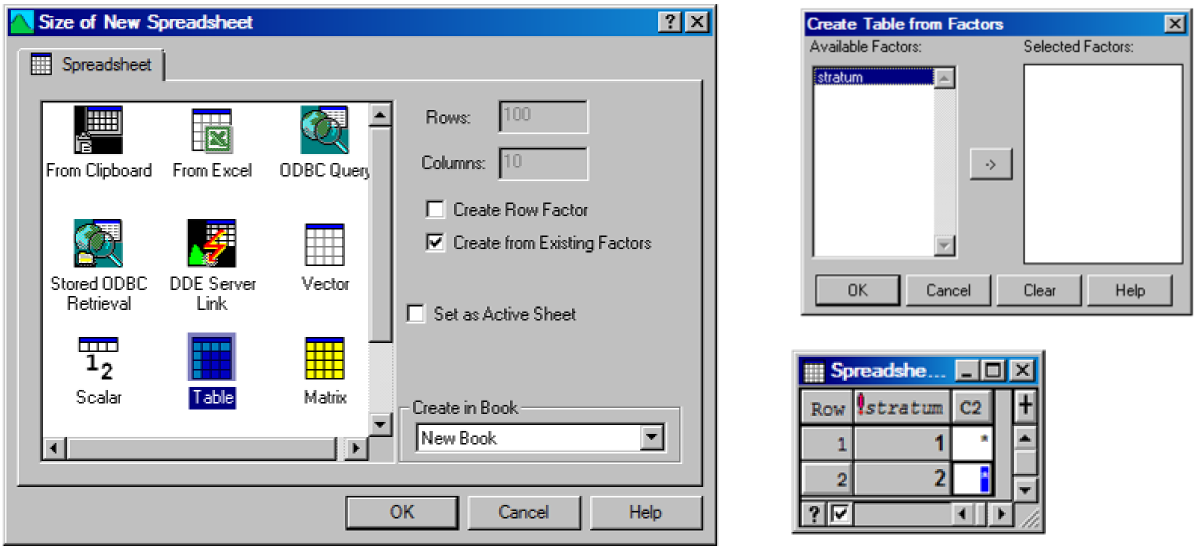


Figure 1.11

The results are shown below. Note that the design effect (`Deff`) is substantially less than 1.0 indicating that the stratification has produced a substantial gain in precision, relative to a simple random sample of the same size.

```
Survey analysis results
=======================

Data summary
------------

Y-variate (response data):  unemployment
Method:                     Design-based (expansion)
Variance method:            Conventional (Taylor series)
Deff:                       0.2065


               Total no. obs.  Imputed    Sample  Excluded  Sampling fraction
      stratum
            1               7        3         4         0              0.571
            2              25       21         4         0              0.160
        Total              32       24         8         0              0.250


Estimated totals with 95% confidence limits
-------------------------------------------
```

```
                  Total       s.e.  %r.s.e.    Lower     Upper
      stratum
            1     10510       4015      38.2     -2267     23288
            2      4700       1481      31.5       -14      9414
        Total     15210       4279      28.1      3081     27340


Estimated means with 95% confidence limits
------------------------------------------

                  Means       s.e.  %r.s.e.    Lower     Upper
      stratum
            1    1501.5      573.6      38.2    -323.8      3327
            2     188.0       59.3      31.5      -0.6       377
         Mean     475.3      133.7      28.1      96.3       854
```

1.7 Practical

Repeat the analysis above working from the full population dataset (sheet `stratified full pop` in `Province.xls`). The results should be identical, but are simpler to calculate because the population sizes for each stratum can be deduced by Genstat from the dataset, removing the need for the user to supply them in a separate data structure.

2 Estimating totals in stratified random surveys

In this chapter we shall examine the estimation of population totals and means from single-stage surveys, including the use of ratio estimation. This type of analysis is common in business surveys that seek to estimate total production, and we will illustrate it using data from the June Agricultural Survey in England. In particular, you will learn about

- ratio analysis
- the different types of output that Genstat will produce
- ways of handling outliers
- how to program the analyses in Genstat's programming language

Whilst some of the material in this chapter is of general applicability, other sections are specific to the **Single-stage Survey Analysis** menu, which runs the `SVSTRATIFIED` command. Those readers working on more complex surveys, or those more interested in cross-tabulations of the data, may prefer to go straight to Chapter 3 where we will consider the more general facilities available from the `SVTABULATE` procedure via the **General Survey Analysis** menu.

2.1 Design based estimators

The June Survey dataset is shown in Figure 2.1 below, and may be found in `June.gsh`. This is a relatively small subset of the full dataset, both in terms of units (nineteen thousand farms, compared to nearly two hundred thousand in the full survey population), and variables (eight, compared to around 150 in the full survery). It includes areas in hectares of various crops from the arable counties of the East of England, excluding very small holdings. Each row represents one agricultural holding (farm), and the spreadsheet contains all farms in the population, with missing values for those that were not sampled, or that did not respond.

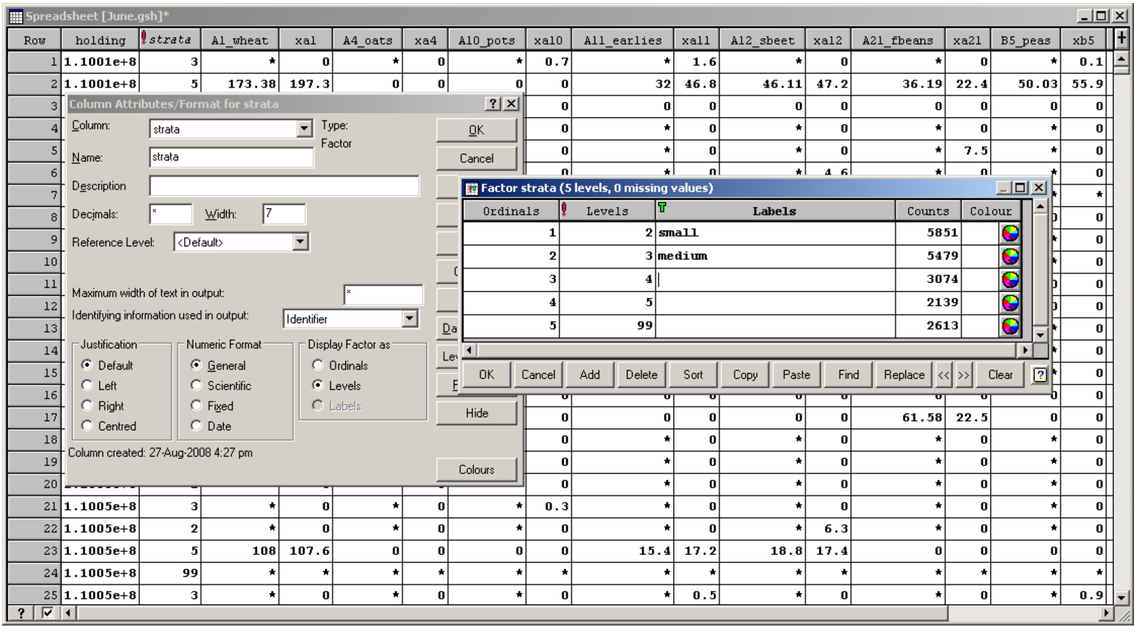


Figure 2.1

The first column shows a unique number for each agricultural holding (note that these have been altered and randomized to preserve confidentiality). The second is a factor (note the red exclamation mark by its name) indicating the stratification used to sample holdings for inclusion in the survey. The strata are indicated by the numbers 2-5 representing different economic sizes of farms, whilst 99s indicate new holdings of unknown size. This type of numeric coding is frequently used for factors, but it is good practice to replace them by more meaningful textual *labels*, as this removes a potential source of confusion in interpreting statistical output. This is achieved by right mouse clicking on the `strata` column, selecting **Column Attributes** from the context menu, and then clicking **Levels & Labels**. The labels can then be entered into the **Labels** column, as shown in Figure 2.1We will alter the labels in this way so that they read `small`, `medium`, `large`, `very large` for categories 2 to 5, and `new` for category 99. The categories can also be reordered by changing their numbers in the **Ordinals** column. In this case we will change the `new` category to have ordinal number 1, and renumber the others to become 2 to 5 (to match their levels), as this ensures they are in approximate order of contribution to the grand total.

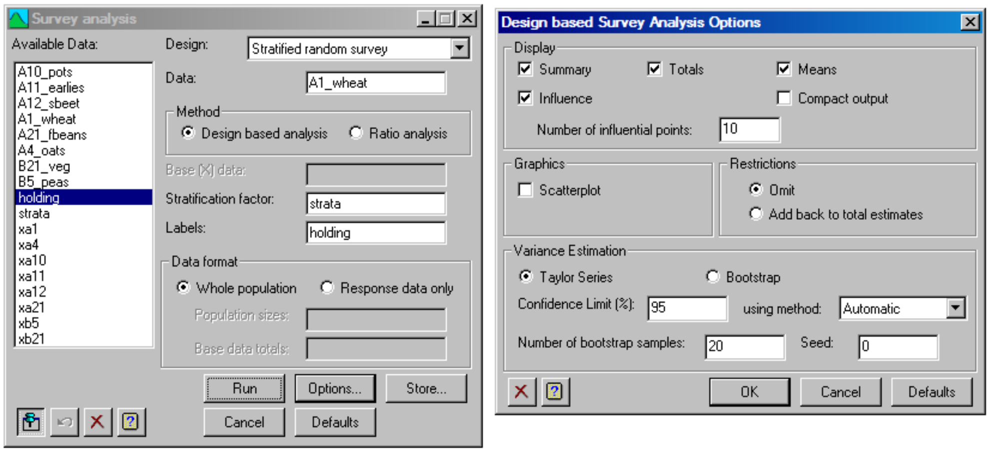


Figure 2.2

The `holding` and `strata` columns are shown in blue italics. This indicates that they have been *frozen* so that they always remain on the left of the window; this is done by selecting **Sheet** from the **Spread** menu with the cursor in the appropriate column, then **Freeze Columns**. The other change that will frequently be required when opening a spreadsheet for the first time is to set the numbers of decimal places shown. In particular, a field such as `holding`, containing long integer numbers will often appear in exponential format (e.g. 1.1001e+8). To set the number of decimal places, make a right mouse click with the cursor on the column, and then select **Column Attributes** before changing the **Numeric format** to **Fixed**.

Let us start by performing the conventional *design based* analysis (sometimes called *expansion raising*) on the area of wheat. This can be done in exactly the same way as the analysis of unemployment in Section 1.1; the menu settings are shown in Figure 2.2 and the resulting output is below.

```
Survey analysis results
=======================

Data summary
------------

Y-variate (response data):  A1_wheat
Method:                     Design-based (expansion)
Variance method:            Conventional (Taylor series)
Deff:                       0.3453
```

	Total no. obs.	Imputed	Sample	Excluded	Sampling fraction
strata					
new	2613	1387	1226	0	0.469
small	5851	4859	992	0	0.170
medium	5479	4357	1122	0	0.205
large	3074	2128	946	0	0.308
very large	2139	917	1222	0	0.571
Total	19156	13648	5508	0	0.288

Estimated totals with 95% confidence limits

	Total	s.e.	%r.s.e.	Lower	Upper
strata					
new	10539	1493	14.2	7610	13469
small	28466	1874	6.6	24787	32144
medium	110304	4568	4.1	101341	119266
large	180870	5787	3.2	169514	192226
very large	329479	6183	1.9	317348	341610
Total	659658	9916	1.5	640216	679100

Notice how, as would be expected from a sensible design, the sampling fraction is greater for the larger farms. It is also high for the new holdings stratum; since no background information is available for them, it is sensible to sample them intensively, in case they are large. In fact the sampling probabilities shown are not, in this example, the ones originally planned, because they are in fact probabilities of being sampled and responding; holdings sampled but not responding are treated in the same way as those not sampled. This is common practice in many surveys, but it is appropriate only if the non-responders can be regarded as being missing at random; by contrast if, for example, farms with more wheat are less likely to respond, the resulting estimates will be biased. Alternatives are to make more complicated adjustments based on a model of non-response, or to use some form of *imputation* (see Chapter 4).

The final estimate of approximately 660 thousand hectares has a relative standard error (coefficient of variation) of 1.5%; this is not bad, but, as we will see in the next section, it can be improved by use of ratio estimation.

2.2 Ratio estimation

Whilst the exact amount of wheat grown by a farmer will vary somewhat from year to year, it tends not to change dramatically. There is thus a high correlation between the responses to this question in the current survey and the responses received the last time farmers were asked it. This correlation between the response variable (in this case the current wheat area) and the *base data* or *auxiliary variable* (the previous area) can be used to produce improved estimates of the population total using *ratio estimation*. For this to work, the base data must also be known for the holdings not sampled in the current year (if only response data are in the spreadsheet the method can also be applied when only the stratum totals of the previous estimates are known).

Other situations where ratio estimation might help are as follows.

- In the `Province` example, the population size of each municipality could be used to improve the precision of the unemployment estimate.
- In a survey of car ownership in a particular area, the number of adults living in each household (perhaps taken from an electoral register) could be used as base data.
- In a field survey designed to estimate the population of an endangered species by sampling 1km squares, the area of suitable habitat in each 1km square might be used as base data.

To see why ratio estimation might improve precision, consider the graphs shown in Figure 2.3. The left hand graph illustrates the ordinary design based estimates; the variability of the observed values about the mean is used to estimate the standard errors (i.e. the quantities indicated by the red vertical lines). With ratio estimation, the variability of interest is about a line described by:

$$Y = rX$$

where r is the ratio calculated as

$$r = \bar{y} / \bar{x}.$$

The standard error is thus based on a variance calculated from the much smaller random errors shown on the right hand graph (again in red).

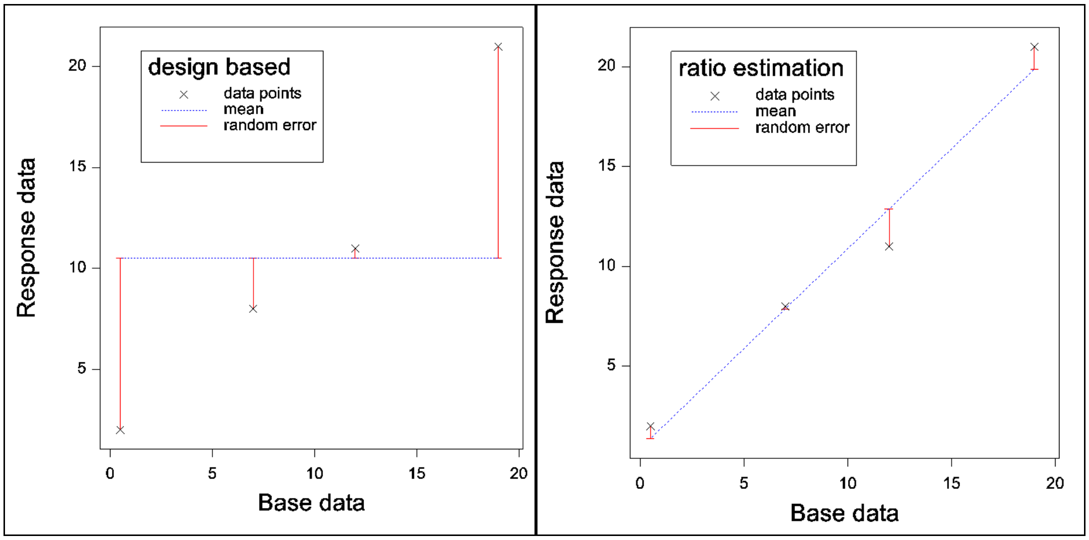


Figure 2.3

Before turning to the analysis, it is helpful to look back to Figure 2.1 to see the structure of the data. Looking down column `xa1` (the previous data for wheat), it can be seen that all holdings contain a value, with the exception of the new holdings in `strata` 99, which have not previously taken part in the survey. Genstat can analyse results like this provided the base data are either always present or always absent within a stratum. Ratio analysis is carried out using the usual **Single-stage Survey Analysis** menu, as is shown in Figure 2.4, and the output is shown below.

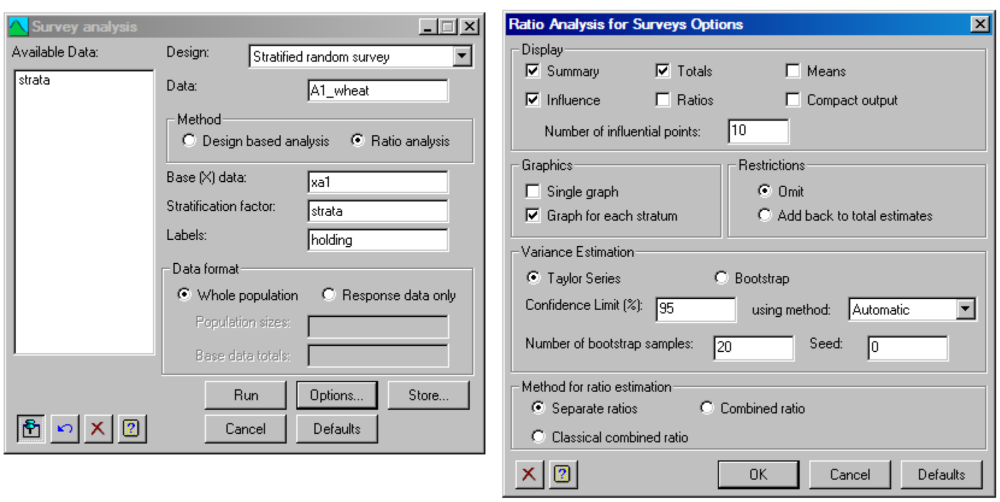


Figure 2.4

```
Survey analysis results
=======================

Data summary
------------

Y-variate (response data):  A1_wheat
X-variate (base data):      xa1
Correlation:                0.935
Ratio method:               separate
Variance method:            Conventional (Taylor series)
Deff:                       0.1159 (wrt design based srs)
Deff ratio analysis:        (Not calculated due to missing X)

               Total no. obs.  Imputed    Sample  Excluded  Sampling fraction
       strata
          new            2613      1387      1226         0              0.469
        small            5851      4859       992         0              0.170
       medium            5479      4357      1122         0              0.205
        large            3074      2128       946         0              0.308
   very large            2139       917      1222         0              0.571
        Total           19156     13648      5508         0              0.288


Estimated totals with 95% confidence limits
-------------------------------------------

                  Ratio     Total      s.e.  %r.s.e.    Lower     Upper
       strata
          new         *     10539      1493     14.2     7610     13469
        small     0.821     55549      1596      2.9    52417     58682
       medium     0.859    164976      2777      1.7   159527    170425
        large     0.905    207290      3978      1.9   199483    215098
   very large     0.912    317537      2200      0.7   313221    321854
        Total     0.896    755892      5758      0.8   744602    767182

Estimates in strata with ratio=* are based on simple raising
The ratio shown in the total row is the combined ratio estimator

* MESSAGE: Default seed for random number generator used with value 622571


10 points with highest influence
--------------------------------

Unit        Stratum             Y           X  %influence
233540082    small           80.0       13.80      0.1048
233860038    small           71.9        0.00      0.1096
281070004    medium         195.2       48.80      0.1484
343460118    large         1116.6      112.90      0.5008
344230042    large            0.0      263.00      0.1178
381130006    new            425.0           *      0.1189
387050023    new            451.1           *      0.1262
388090049    large          439.4       69.00      0.1860
481490005    small           74.2        0.00      0.1131
614160015    very large     722.0      224.00      0.1157
```

```
Percentage influence is calculated as the percentage change
in the grand total when each sampled observation is omitted.
```

A few extra items are now shown in the output. Firstly the correlation between the response data and the base data is shown; this will give a good indication of whether the ratio analysis will be more effective than a design based analysis. In this case the correlation is 0.935, suggesting that it should be highly effective. In the case of ratio analysis two *design efficiency* figures (*deff*) are usually quoted: one comparing the stratified sampling with a simple random sample of equivalent size, and one comparing the ratio analysis with a design based one. In this example the latter cannot be calculated due to the missing base data in the new holdings stratum.

In the table of total estimates, the ratio of response data to base data for the responding holdings is shown for each stratum. The estimated total is obtained by multiplying the sum of all base data in the stratum by the ratio. Since the base data are all missing from the new holdings stratum, no ratios can be calculated and the estimate of the total wheat area for the stratum is calculated using the design-based analysis (hence the estimate of 10539ha for new holdings, with s.e. of 1493ha is identical to that produced in Section 2.1). In all other strata, where estimates use ratio estimation, the standard errors are considerably lower than those of Section 2.1. The result is that the standard error of the estimate of the total area of wheat in the region is now less than 6,000ha, compared to almost 10,000ha without the use of the base data.

The `SVSTRATIFIED` command can produce a variety of different output, and to see exactly how the calculations are performed it is helpful to use a *compact* style of output by clicking the **Compact output** box on the **Options** menu. This is designed to produce a comprehensive summary of the analysis that can nevertheless fit onto a single sheet of paper, provided the number of strata is not too large. It can be used only with *plain text* output, which can be obtained by selecting **Output** on the **View** menu

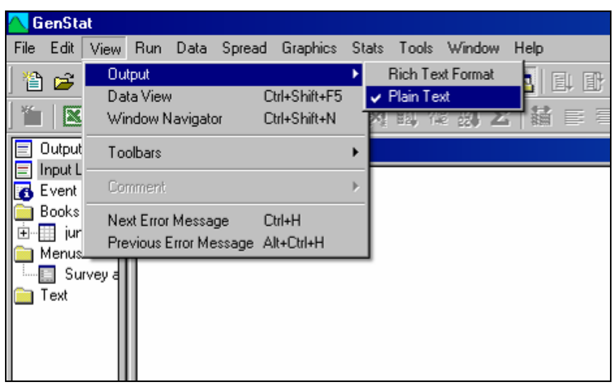


Figure 2.5

and then clicking on **Plain Text** (Figure 2.5). To get the full information shown below, the output width should be set to 110 characters or more by selecting **Options** from the **Tools** menu, and then altering the setting on the **Text Editor** tab (Figure 2.6).

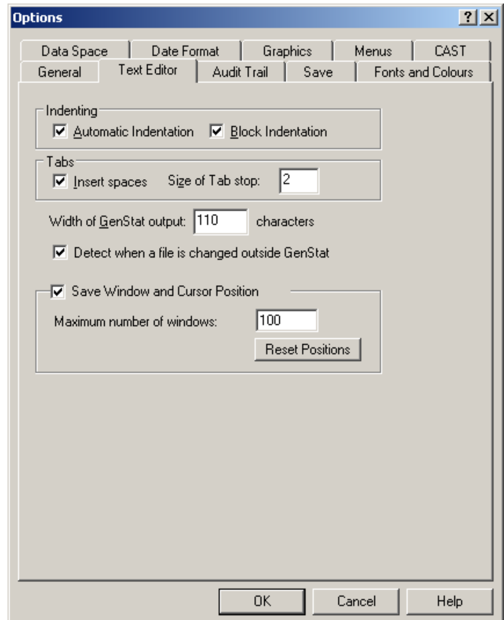


Figure 2.6

Figure 2.7 shows the output produced with this option set. The first difference in the compact output is that the table of observations now has two extra columns giving the number of observations greater than zero for the matched pairs of response (*y*) and base (*x*) data from those holdings responding to the survey (for example, looking at the spreadsheet in Figure 2.1, rows 1 and 4 are excluded from these figures because they have missing values for `a1_wheat`). These numbers of non-zero observations are important in interpreting datasets, such as this one, where there are many zeros, as otherwise the sample size can give a misleading impression of the robustness of estimates.

Totals for the responding units are shown in the table of `estimated totals`, again calculated using only the matched pairs of *y* and *x* figures in holdings where ratios are estimated. These are the figures used to calculate the ratio. For example, in stratum `small` the ratio is:

$$r = \Sigma y_i / \Sigma x_i = 4826 / 5879 = 0.8209$$

```
Survey analysis results
=======================


Data summary
------------

Y-variate (response data):  A1_wheat
X-variate (base data):      xa1
Correlation:                0.935
Ratio method:               separate
Variance method:            Conventional (Taylor series)
Deff:                       0.1159 (wrt design based srs)
Deff ratio analysis:        (Not calculated due to missing X)

                    Numbers of observations      Sampling      Matched data
                total  imputed   sample excluded fraction      y>0      x>0
       strata
          new    2613     1387     1226        0    0.469       82        0
        small    5851     4859      992        0    0.170      260      332
       medium    5479     4357     1122        0    0.205      499      563
        large    3074     2128      946        0    0.308      619      656
   very large    2139      917     1222        0    0.571      994     1028
        Total   19156    13648     5508        0    0.288     2454     2579

Estimated totals
----------------

                Matched sample          All data    Raising factor  Estimated totals
                sum y    sum x    ratio    sum x    ratio expans'n  imputed      all     s.e. %r.s.e.
       strata
          new    4945        *        *        *    2.131    2.131     5594    10539     1493    14.2
        small    4826     5879   0.8209    67667   11.510    5.898    50723    55549     1596     2.9
       medium   22588    26287   0.8593   191992    7.304    4.883   142388   164976     2777     1.7
        large   55661    61524   0.9047   229123    3.724    3.249   151629   207290     3978     1.9
   very large  188230   206309   0.9124   348037    1.687    1.750   129308   317537     2200     0.7
        Total  276250   299999   0.8965   836819    2.736    3.478   479642   755892     5758     0.8

95% confidence limits for total are 744602 to 767182

Estimates in strata with ratio=* are based on simple raising
The ratio shown in the total row is the combined ratio estimator
```

Figure 2.7

The column to the right of the ratios shows the totals of the base (x) data for all units in the population. The estimates of the stratum totals (headed `all`) are obtained by multiplying these by the ratio. Again using the `small` stratum as an example:

Total = $r\, \Sigma x_i$ = 0.8209*67667 = 55549

(where summation is over the whole population).

The `imputed` column contains the estimated total for the unsampled/non-responding holdings. This is the difference between the total estimated wheat areas shown in column `all` and the total of the response data shown in the first numeric column. In the `small` stratum:

Imputed total = 55549 – 4826 = 50723ha

Comparison between the `imputed` and `all` columns thus provides an easy way of seeing how much of the estimated total in each stratum comes from real data, and how much is imputed from unsampled or non-responding holdings. Similarly looking up and down the imputed column shows where estimation is most critical. In this example, whilst the greatest estimated wheat area is in the `very large` stratum (318 thousand hectares), only 129 thousand hectares of this is imputed, compared to 188 thousand hectares obtained directly from farmers' responses. The imputed totals are actually higher for the `medium` and `small` strata due to their lower sampling fractions, suggesting that these strata are key to the accuracy of the overall estimate for this variable. This is confirmed by the size of the standard errors for these strata.

Raising factors are also shown in the table; these are more commonly known as *survey or sampling weights*. The design-based estimates shown in Section 2.1 are obtained by multiplying the response totals (column `sum y`) by the *expansion raising factor*, whilst the ratio estimates shown in Figure 2.7 are obtained by multiplying the response totals by the *ratio raising factor*[1]. Thus for the `small` stratum:

Design-based estimate = 4826*5.898 = 28466ha

ratio estimate = 4826*11.510 = 55549ha

Thus these two columns are useful for highlighting strata where, as in this case, the estimates using the two methods differ substantially.

---

[1] In the terminology of Lehtonen & Pahkinen (1994, *Practical methods for the design and analysis of complex surveys*) the raising factor is the adjusted weight, formed by multiplying the sampling weight by the g-weight.

2.3 Dealing with outliers

If the **Influence** box is ticked, as in Figure 2.4 the following list of influence statistics is produced.

```
10 points with highest influence
--------------------------------

Unit         Stratum               Y            X  %influence
233540082     small             80.0        13.80      0.1048
233860038     small             71.9         0.00      0.1096
281070004     medium           195.2        48.80      0.1484
343460118     large           1116.6       112.90      0.5008
344230042     large              0.0       263.00      0.1178
381130006     new              425.0            *      0.1189
387050023     new              451.1            *      0.1262
388090049     large            439.4        69.00      0.1860
481490005     small             74.2         0.00      0.1131
614160015     very large       722.0       224.00      0.1157

Percentage influence is calculated as the percentage change
in the grand total when each sampled observation is omitted.
```

Note that the number of influential points shown can be increased if needed by using the **Options** menu (Figure 2.4), and that the variable listed in the **Labels** box (in this case holding number) is used to label the units; by default the row number is displayed.

Just because an observation is influential, it does not follow that it is incorrect, or that any adjustment is necessary. However, if resources are available to carry out checks on some of the data-points, it is sensible to concentrate on these observations in order to maximize the reliability of the final estimate. The magnitude of the influence statistics is one guide to the effort which it is sensible to expend. In the example shown, most of the units have values of just over 0.1%; since the total estimate of the wheat area is 750 thousand hectares, this implies that they change the estimate by around 750ha, which is small compared to the standard error of nearly 6,000ha (this comparison can also be made by comparing the influence statistics with the relative standard error). Hence, investigating these influential points will not have much impact on the overall estimate, unless there is some systematic error causing a large number of units to all influence the total in the same direction; this might happen, for example, if a number of holdings had recorded their wheat areas in acres not hectares.

There is however one influence value that is much larger than the rest; the holding in unit 343460118 changes the overall estimate by around 0.5%. What is more, the information shown in the table is suggestive of a typing error. The recorded wheat area is 1116.6ha compared to 113ha the previous year; such a large increase would be highly unusual, whereas a change from 113ha to 116.6ha would be much more plausible. Checking the original survey form did indeed reveal that the farmer had written 116.6ha but that this had been miss-keyed as 1116.6ha.

Once an outlier has been identified, it is necessary to decide what to do with it. In the case above this is straightforward; the miss-key should be corrected in the Genstat spreadsheet (and in the database from which it was formed, if appropriate) and the analysis repeated. The appropriate row in the spreadsheet can easily be found to do this by clicking on the binoculars icon on the toolbar and searching for the holding number. In other cases, one of the following actions might be needed.

- The observation can be replaced by a missing value. This is the correct course of action if it is clear that the data are unreliable, but the correct value cannot be found, possibly because the farmer could not be contacted. This is perhaps more likely to occur in an anonymous survey, when it is impossible to re-contact the respondent to find the correct values. To insert the missing value, simply find the appropriate row in the spreadsheet, highlight the value and press the **Delete** key.
- The unit can be removed from the population. This is quite unusual but may be necessary if, for example, investigation shows that the farm was actually outside the geographic area covered by the survey. This can be achieved by restricting the unit out of the data, as described in the next question.
- The unit can be given a weight of exactly 1.0 in the analysis (*added back*). This means that it contributes to the total estimate, but is ignored for the purposes of extrapolating the results to the unsampled units. This is done when the unit is not representative of the survey population as a whole. It can also be achieved using a restriction, but it is important to understand the reasons for this approach and so it will be considered in more detail in the next section.

It is also important to stress that none of these actions are appropriate when the outliers are not due to any errors and when the units are genuinely representative of the survey population. In this case the original analysis must stand, although, when one or more units strongly influence the results, it may be appropriate to publish the estimates with a warning that this is the case, or to publish results with and without the outlier(s).

2.4 Using restrictions

In this section we will look at how restrictions can be used to exclude an observation from the population, or to deal with an outlier that needs to be given a weight of 1.0 because it is not representative of the wider survey population.

A restriction is generally used in Genstat to confine an analysis to a specified subset of the data, but on a temporary basis, so that the full dataset is still stored within Genstat, thus allowing rapid removal of the restriction. When analysing single-stage surveys with SVSTRATIFIED, restrictions may be used to exclude a unit totally from the survey population.

Restrictions can be created most easily by using the **Spread** menu. For example, let us suppose that investigations have shown that holding number 343460118, the outlier identified above, was actually not in the survey region at all, but was a farm in Scotland. It should therefore be completely removed from the dataset, and one option would be simply to delete this row. However, this can sometimes cause

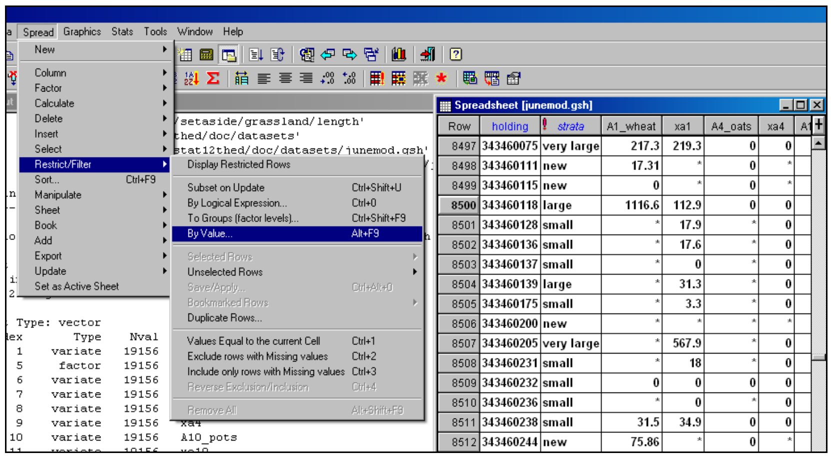


Figure 2.8

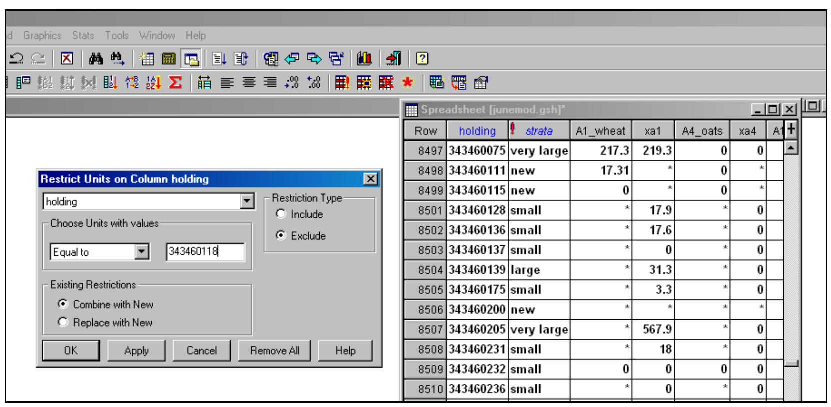


Figure 2.9

problems, perhaps because we have other spreadsheets that would also require modification, so instead we will use a restriction to exclude it. First we select **By value** on the **Restrict/filter** option on the **Spread** menu (Figure 2.8) and then enter the holding number (Figure 2.9). We want to exclude, not include this unit, so we click the appropriate radio button. Once we click either the **Apply** or **OK** button row 8500 vanishes from the spreadsheet window, as shown in Figure 2.9. In some situations it may be still useful to see the excluded data, and this may be achieved by clicking on the cross at the top of the scrollbar; the row excluded by the restriction then appears in red (Figure 2.10).

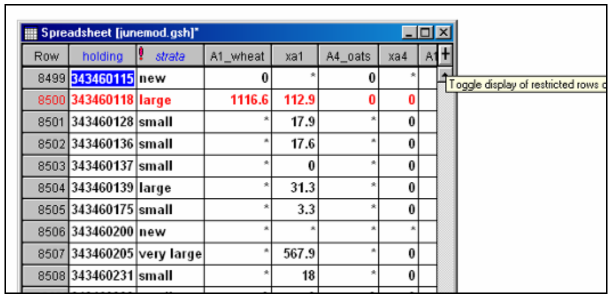


Figure 2.10

Once the restriction has been applied, the analysis can be re-run to produce the output shown below.

```
Survey analysis results
=======================

Data summary
------------

Y-variate (response data):  A1_wheat
X-variate (base data):      xa1
Correlation:                0.944
Ratio method:               separate
Variance method:            Conventional (Taylor series)
Deff:                       0.0924 (wrt design based srs)
Deff ratio analysis:        (Not calculated due to missing X)

               Total no. obs.  Imputed    Sample  Excluded  Sampling fraction
       strata
          new            2613     1387      1226         0              0.469
        small            5851     4859       992         0              0.170
       medium            5479     4357      1122         0              0.205
        large            3074     2128       945         1              0.308
   very large            2139      917      1222         0              0.571
        Total           19156    13648      5507         1              0.287
```

Estimated totals with 95% confidence limits

	Ratio	Total	s.e.	%r.s.e.	Lower	Upper
strata						
new	*	10539	1493	14.2	7610	13469
small	0.821	55549	1596	2.9	52417	58682
medium	0.859	164976	2777	1.7	159527	170425
large	0.888	203405	2868	1.4	197777	209033
very large	0.912	317537	2200	0.7	313221	321854
Total	0.892	752007	5055	0.7	742096	761918

Estimates in strata with ratio=* are based on simple raising
The ratio shown in the total row is the combined ratio estimator
Totals and means exclude restricted (excluded) data

The excluded column now shows that one unit from the large stratum is excluded from the calculations.

In the analysis of a single-stage survey using the SVSTRATIFIED command, restrictions can also be used when one or more units are not considered as representative of the wider population. They are then excluded from the main calculations but are 'added back in' to the final estimates. This is equivalent to giving them a weight of 1.0 in the analysis. It is achieved by clicking on the add back to total estimate radio button on the options menu (Figure 2.11).

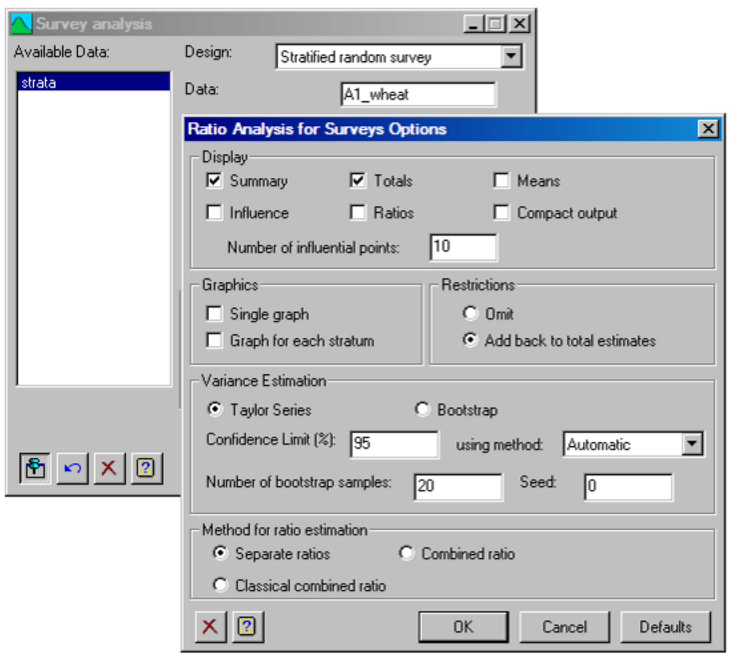


Figure 2.11

Let us now suppose that the response from `holding` 343460118 was indeed correct, but that is not considered representative of the wider population. This might be because of some exceptional factor that did not apply to other holdings. When using a ratio analysis, it is also permissible to use this approach if the base data are thought to be incorrect. For example, suppose that investigations on `holding` 343460118 showed that the area of 1116ha was correct, but that the previous value of 113ha was incorrect and the true base value could not be ascertained. Thus the apparent ten-fold increase in the wheat area is misleading and should not be extrapolated to other holdings. The modified estimates of the totals is then as shown below.

```
Estimated totals with 95% confidence limits
-------------------------------------------

                  Ratio     Total      s.e.   %r.s.e.     Lower     Upper
       strata
          new         *     10539      1493      14.2      7610     13469
        small     0.821     55549      1596       2.9     52417     58682
       medium     0.859    164976      2777       1.7    159527    170425
        large     0.888    204522      2868       1.4    198893    210150
   very large     0.912    317537      2200       0.7    313221    321854
        Total     0.892    753123      5055       0.7    743212    763035

Estimates in strata with ratio=* are based on simple raising
The ratio shown in the total row is the combined ratio estimator
Totals and means include restricted (excluded) data
```

Notice that the new total estimate is now equal to the previous total estimate when the holding was completely excluded, plus the observed value for the holding which has been ‘added back’ to the total:

New estimate = 752007 + 1116 = 753123

It is important not to over-use this approach. It can be tempting to assume that just because an observation is influential, it is atypical and should be added back to the total in the way described above. This is incorrect and can lead to an undesirable degree of subjectivity in results, with outliers being removed until an expected value is achieved. Instead the approach should be used only in exceptional circumstances, where the unit is clearly qualitatively different to the rest of the population, or where there is a problem with the base data.

2.5 Practical

The approach of adding an outlier back to the total is equivalent to putting the observation in its own stratum, which is therefore sampled at a rate of 100%. To show that this is the case, duplicate the stratum factor and create an extra factor level. Remove all restrictions and edit the duplicated stratum factor to take this new value for holding 343460118 before running the analysis again.

2.6 The combined ratio estimator

As we have seen, the analysis of the wheat area produced robust results. There was a single large outlier, and the ratios look logical, with an increasing trend with increasing farm size. This is not always the case, particularly when numbers of sampled observations in each stratum are smaller and the distribution of the data is more skew. Consider the example of variable `a11_earlies`, which gives the area of early potatoes grown on each holding.

```
Estimated totals with 95% confidence limits
-------------------------------------------

              Ratio     Total      s.e.   %r.s.e.     Lower     Upper
      strata
         new      *       989     263.8      26.7       471      1506
       small  1.293      2270     307.8      13.6      1666      2874
      medium  0.763      5099     417.6       8.2      4279      5918
       large  0.978      9131     507.8       5.6      8135     10128
  very large  0.912     32980     954.1       2.9     31108     34852
       Total  0.916     50469    1227.6       2.4     48062     52876
```

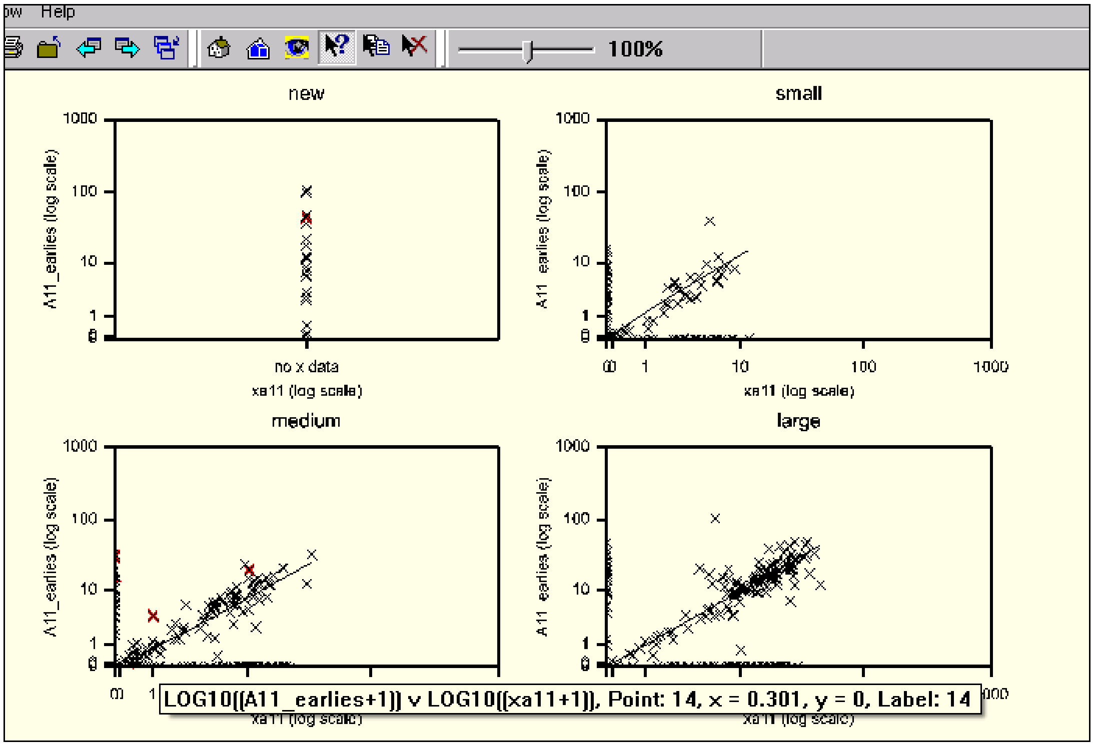


Figure 2.12

In this case, whilst the number of observations is the same as for the wheat example, there are far fewer non-zero values, resulting in larger relative standard errors. This can be seen in the plots in Figure 2.12, which have been produced using the **Graph for each stratum** button on the **Options** menu – notice the use of the data information tool (on toolbar, with arrow and question mark) to reveal details of a point on the graph. The ratios show a less logical trend, which could be simply a product of random variation; it is difficult to see why the ratio for medium sized farms should be much lower than that for either small or large ones. It may therefore be preferable to use a robust estimator of the ratio, pooling information from all strata. This can be achieved by clicking either **Combined ratio** or **Classical combined ratio** from the **Options** menu. An extract of the output is shown below in compact style.

```
Estimated totals
----------------

               Matched sample            All data    Raising factor  Estimated totals
               sum y     sum x     ratio     sum x    ratio expans'n  imputed       all
       strata
          new     464        *         *         *    2.131     2.131      525       989
        small     284      220    0.9161      1755    5.944     5.898     1406      1691
       medium     816     1070    0.9161      6684    7.303     4.883     5144      5960
        large    2509     2566    0.9161      9338    3.473     3.249     6204      8713
   very large   23510    25790    0.9161     36179    1.405     1.750     9517     33027
        Total   27584    29646    0.9161     53956    1.826     3.478    22795     50379
```

The results shown are for the setting **Combined ratio**; the overall ratio is applied to the sum of the base (x) data for holdings not sampled, and then this is added to the observed response (y) data. For example, for the small farms stratum:

Estimate of total = (1755-220)*0.9161 + 284 = 1691ha

The classical combined ratio is the form presented in most textbooks, in which the base data total is simply multiplied by the overall ratio:

Estimate of total = 1755*0.9161 = 1608ha

In general the two variants give similar results, but when sampling ratios are high the classical combined ratio can occasionally produce illogical estimates, where the total for the whole stratum is estimated to be less than that for the sampled units.

2.7 Saving and exporting results

Most of the results displayed in the output produced by the `SVSTRATIFIED` command can be saved using the **Save** menu (Section 1.2). Saving results can be useful for two reasons. Firstly, the saved structures can then be used within Genstat for further calculations or in the production of graphs. Secondly, it is often necessary to export the results to other packages (for example, Excel) for presentation or other purposes.

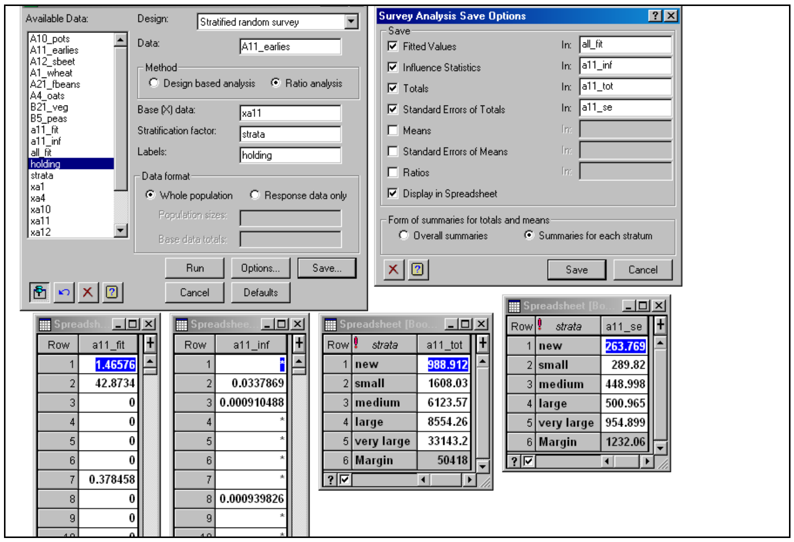


Figure 2.13

Figure 2.13 shows the options set to save the fitted values, the influence statistics, the totals and the standard errors of the totals. The **Display in Spreadsheet** box is ticked and the resulting spreadsheets are shown below. Note that in this case the totals and their standard errors have been saved as table structures, which means they are labelled with the stratum names. Alternative, if the **Overall summaries** button is selected, scalar quantities are created, saving just the overall total figures.

The fitted values are often useful, for example in constructing estimates for sub-populations. When ratio analysis is used they are equal to the base value times the appropriate ratio, or when the design based estimation is used (as in the new holdings stratum here), they are simply set to equal the mean.

When tables of results are to be exported into other packages, there are four possible approaches.

1. Cutting and pasting from the output window. Simply highlighting the output and selecting **Copy** from the **Edit** menu (or the **Copy** button or **ctrl-c**) can be adequate, particularly if pasting into a word processing

package from output in rich text format. If using plain text output it will be necessary to use a font such as *courier* in the word processing package.

2. Copy special. When copying from a plain text output window, pasting into a word processing package does not give a true table, and this can result in poor alignment of columns. Selecting **Copy Special** from the **Edit** menu gives a variety of options that allow results to be copied in a proper tabular form.

Both of the above approaches will copy results only with the precision shown in the output window, and so they are not advisable when further numerical processing is intended. The following options avoid this problem.

3. Copying from spreadsheets. The data to be exported are put in one or more spreadsheets by selecting **New** and then **Data in Genstat** from the **Spread** menu. The required cells are then highlighted and copied. This method results in the data being pasted with full precision, as long as the **Paste with full precision** box is ticked on the **File** tab of **Spreadsheet Options** (which can be opened from the on the **Tools** menu on the menu bar).
4. Saving from spreadsheets. Once the data have been put in a spreadsheet, Genstat allows them to be saved in a wide variety of formats for import into spreadsheets or other statistical packages. This is generally the best approach when large amounts of data are to be exported.

3 General survey analysis

So far all the analyses considered have used simple random sampling or stratified random sampling, and their aim has been to estimate a population mean or total. In this chapter we will learn how the **General Survey Analysis** menu and the SVTABULATE procedure can be used for the following more complex situations

- designs with unequal sampling weights
- cross-tabulations of means, totals and ratios
- Wald tests of differences between means
- means, totals and ratios for sub-populations
- two stage samples

The table below compares the features of SVTABULATE with the SVSTRATIFIED command used for the analyses in Chapters 1 and 2.

	SVSTRATIFIED	SVTABULATE
Menu	Single-stage Survey Analysis	General Survey Analysis
Main purpose	Estimation of population means and totals	Cross tabulations
Stages	One-stage only	One- or two-stage
Survey weights	Calculated internally	Usually supplied explicitly, but can be calculated
Quantile estimation	No	Yes
Ratio estimation	Yes	Yes, but cannot directly produce population totals
Wald tests	No	Yes
Restrictions	Used to exclude unit from population, or add back in	Define subpopulations

In this chapter we will deal with datasets where the weights are supplied. Information on how survey weights are calculated and modified can be found in Chapter 4.

All analyses described in this chapter are carried out using the menu system. If you are interested in using Genstat's command language, you may find it helpful to read it in conjunction with Chapter 5 on programming and Appendix 1 which lists the commands to achieve the same analyses.

3.1 Farm Business Survey datset – merging data

We shall illustrate the next few sections with data from the Farm Business Survey in England. This is a single-stage stratified random sample, but the survey weights have been adjusted by *calibration* (Section 4.5) so that they are not equal within a stratum. For the purposes of this chapter we will treat the calibration weights as if they were sampling weights; this is not strictly correct, but it is generally a conservative assumption (i.e. standard errors will be larger than the true values). In Section 4.5 we will show how the correct standard errors can be calculated.

The data available here consist of the farm's net margin, income from farming, income from other activities, and subsidy payments. There is also information on the farmer's sex, age and level of education. This dataset is in the Excel file `FBSdata.xls`. A separate Genstat file, `FBS_England.gsh`, contains the information needed for analysis, namely the survey weights and strata, plus some other information on the farms.

In order to merge these files for analysis, we first need to import the Excel file into Genstat using the *Excel import wizard* (the Excel icon on the second row of the toolbar), as described in Section 1.1. An additional complication is that the spreadsheet has an extra line of text in row 2, giving the variable names in the survey database (Figure 3.1). These can be read in as column descriptions by clicking the appropriate box at the **Select Options for Importing Excel Data** window which is displayed by the wizard (Figure 3.2). Column descriptions can be

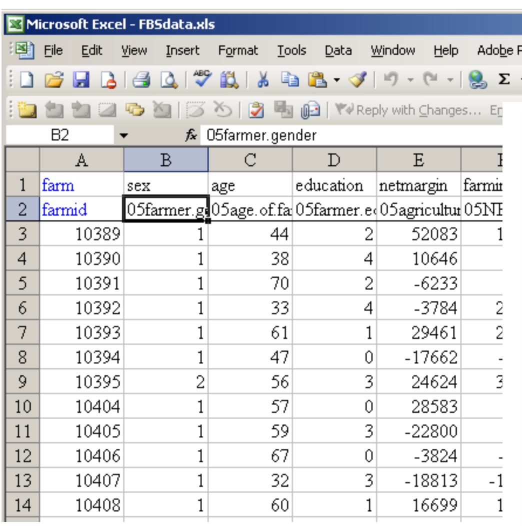


Figure 3.1

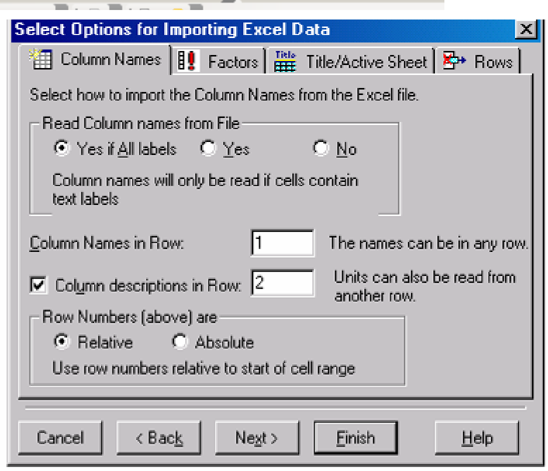


Figure 3.2

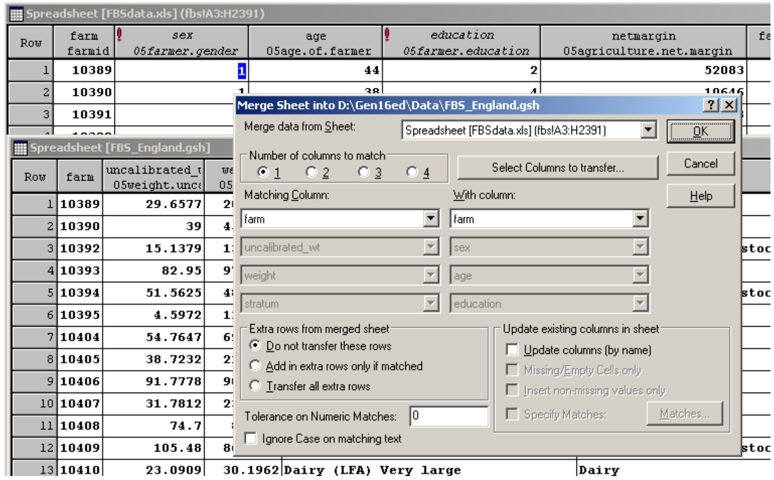


Figure 3.3

particularly useful to provide a fuller definition of each column when it is desired to keep the column names themselves brief. To allow `Sex` and `education` to be used in cross-tabulation, they should be set to be factors, either in the wizard or by right mouse-clicking in the columns and selecting **Convert to Factor** from the context menu (see Section 1.6).

Next the file `FBS_England.gsh` needs to be opened in Genstat; then with it as the active window select **Manipulate** from **Spread** menu and then **Merge** from the sub-options. The `FBSdata.xls` dataset has some extra farm businesses in it that are not required for the analyses here, so the **Do not transfer these rows** button should be clicked (Figure 3.3). The completed file should then be saved so that these operations do not need to be repeated. A version of the merged file is provided as `FBS_England_Merged.gsh`, if you do not wish to do this yourself. This version also has labels added to the education factor to aid its interpretation. The four columns on the left have been frozen (**Freeze columns** from **Sheet** on the **Spread** menu).

3.2 Cross-tabulation

To illustrate cross-tabulation we will produce a table of mean incomes by farmers' sex using the **General Survey Analysis** menu with **Stratified random survey** selected in the **Design** box. Figure 3.4 shows the appropriate settings. Factor `stratum` has been clicked across into the **Stratification factor(s)** box, and factor `sex` across into the **Classification factor(s)** box. More than one factor can be specified, if required; they should be separated with commas. The variate `weights` has been entered into the **Weights** box, and the variate `farm` into the **Labels** box.

Note that, unlike with the `SVSTRATIFIED` command, it is not always necessary to supply the stratum population sizes. This is because `SVTABULATE` can deduce them from the sum of the weights in each stratum. However, if preferred, the weights can be left unset and population sizes supplied instead. In Figure 3.4 the labels have been set to the farm numbers; this makes the influence statistics easier to interpret than if they were labelled by the row numbers, which would be the case if this box is left blank.

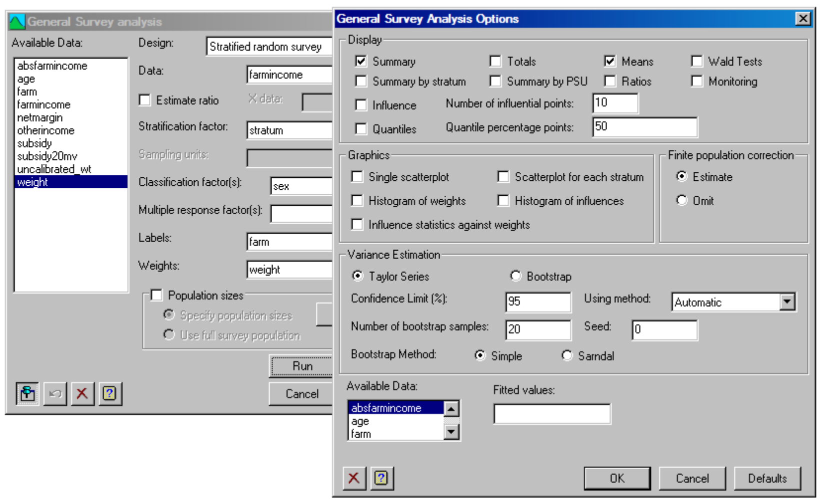


Figure 3.4

The output produced when the **Run** button is clicked is shown below[2]. At the top is a summary of the analysis. This includes information on the range of weights. More detail on the range of the weights and the response data, as well as the number of observations per stratum, can be obtained by ticking the **Summary by**

[2] The methodology used for calculating survey estimates in Genstat is similar to that used in the US Census Bureau's Cenvar package - see http://www.census.gov/ipc/www/imps/download.htm.

stratum box on the **Options** menu. In this case it might be wise to investigate the large range in weights, since there is a more than one hundred-fold difference between the minimum and maximum weights.

```
Survey analysis results
=======================

Summary of analysis
-------------------

Y-variate (response data):               farmincome
Method:                                  Design-based (expansion)
Stratification factor:                   stratum
Number of strata:                        75
Components for variance calculation:     Between sampling units
Confidence interval method:              tdistribution (95% limits)
Total number of responses:               1776
Survey weights:                          weight
Weights range:                           Min = 1.483  Mean = 34.71  Max = 185.8
Sum of weights:                          61653


Means with 95% confidence limits
--------------------------------

                      n   Sum wts      Mean     s.e.   %RSE/CV     Lower    Upper
 05farmer.gender
          male     1723     59740     21403     1884      8.80     17708    25098
        female       53      1913     12823     5757     44.89      1532    24114
          Mean     1776     61653     21137     1830      8.66     17547    24726

Standard errors based on Taylor series approximations. Confidence limits use t-
distribution with 1701 d.f.
```

Looking at the results themselves, the mean income for female farmers is much smaller than that for males, but the sample size is small for the latter, with a relative standard error approaching 50% of the mean. The sum of the weights for each category is also shown, and this shows that the low sample size for women farmers reflects the low estimated number in the population, rather than being due to a particularly low sampling level. Given the large standard errors, it is difficult to tell if this represents a real difference between the mean farming incomes of men and women in the population. Ticking the **Wald Tests** box on the **Options** menu (Figure 3.4) produces the following output.

```
Adjusted Wald test
------------------


                   Means
 05farmer.gender
         male      21403
       female      12823


Test of null hypothesis that the means above are equal
Test statistic F = 1.98 with 1 and 1701 d.f.
Probability = 0.160
```

The Wald test indicates that there is a probability of 0.160 (i.e. 16%) of observing an F-statistic at least as large as this, even if there is no real difference between the means. Hence we cannot reject the null hypothesis that the means are different.

One point to note is that the calculation of Wald tests requires knowledge of the covariances between the dummy variables representing the different cells in the table. This involves the use of a different algorithm that is much slower with large datasets. Hence, except for small datasets, it is best not to calculate Wald statistics (or to save variance-covariance matrices of estimates) unless they are genuinely needed.

Further information on the reliability of these means can be obtained by displaying the influence statistics. These are defined in a similar way to those produced by the `SVSTRATIFIED` command; they indicate the percentage change in the estimate of the grand total (or equivalently, the mean) when the observation is replaced by a missing value and its weight is redistributed across the other observations in each stratum. Farm 14501 has by far the biggest influence statistic with respect to the grand total, with an income of over £3 million. This is surprisingly large for a farm in a stratum classified as `small`, and it should therefore be checked. Influence statistics are also shown for the individual cells in the table (i.e. for the male and female cells, as opposed to the grand total). Farm 14501 is again large, but there are some even larger statistics for those with female farmers; not surprisingly, given the small sample size.

```
10 points with highest percentage influence on grand total
----------------------------------------------------------

farm         stratum                                  Weight  farmincome  %influence
10891        Dairy (Lowland) Medium                    47.75      283184       0.957
12452        Specialist Poultry Small                  54.12       24971       0.900
12506        Specialist Poultry Small                  42.96       23521       0.703
12518        Specialist HNS Very large                 35.63      323580       0.690
14501        Specialist Poultry Small                  30.58     3273062       7.612
14583        Cereals Very large                        24.58      583178       1.021
14595        General Cropping Very large               23.96     -448626       0.999
14601        General Cropping Very large               14.75      678310       0.688
14848        Mixed Very large                          49.57      250916       0.786
43140        Other Horticulture Very large             27.24      625315       1.143

* Note: The influence value is the percent change in the estimate when the observation is
omitted

10 points with highest percentage influence on individual cells
---------------------------------------------------------------

farm          05farmer.gender      Weight  farmincome  %influence
10477        female                 16.20       81457        5.38
14459        female                 28.74      -91126       10.54
14501        male                   30.58     3273062        7.76
14598        female                 20.55      408924       34.18
15856        female                 45.35       29849        5.44
16005        female                 74.26      -17954        5.44
43214        female                 28.10      151862       16.87
43295        female                  4.88      443265        8.82
43471        female                 30.16      117888       13.81
48360        female                 40.19      -29156        4.78
```

- Note: The influence value is the percent change in the estimate when the observation is omitted

3.3 Sub-populations

Tables of means or totals can be classified by two or more factors, but in practice this can make the output more difficult to interpret, particularly if the factors have many levels. If only some of the factor levels are of interest, more concise tables may be produced by confining the analysis to this *subpopulation*. For example, suppose we were interested in the effect of educational qualifications on the farming income of male farmers. Rather than having to interpret two-way tables classified by sex and education, we can restrict the analysis to male farmers only, so that only the cells of interest are shown.

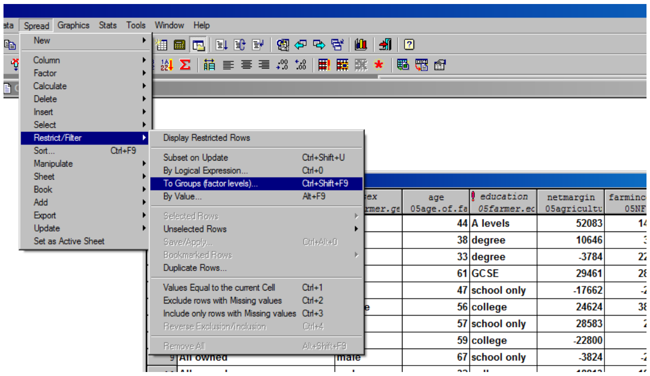


Figure 3.5

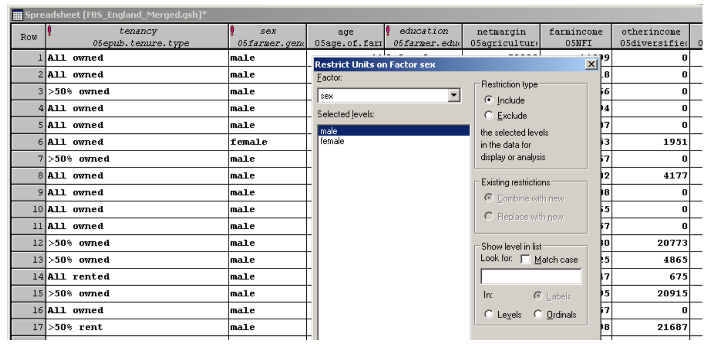


Figure 3.6

The first stage in this analysis is to apply the restriction by selecting **To Groups (factor levels)** from the **Restrict/Filter** submenu of the **Spread** menu (Figure 3.5) with the spreadsheet window active. `Sex` can then be selected from the drop down list of factors and `male` highlighted as shown in Figure 3.6, and when the **Apply** button is clicked, row 6 relating to a female farmer disappears from view. In order to check that the restriction is operating as intended, particularly with complex restrictions, it may be helpful to click on the black cross in the top right hand corner of the spreadsheet window, level with the variable names; rows excluded from the dataset by the restriction are then shown in red (Figure 3.7).

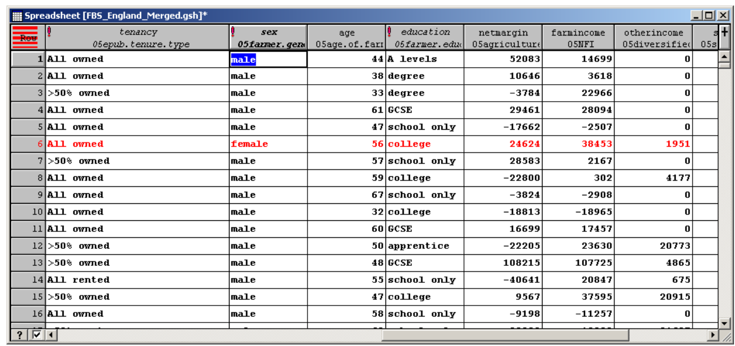


Figure 3.7

Once the restriction has been used to define the sub-population of interest, the analysis can be specified, as in Figure 3.4 but with `education` as the classification factor. The output is shown below.

```
Survey analysis results
=======================

Summary of analysis
-------------------

Y-variate (response data):            farmincome
Method:                               Design-based (expansion)
Stratification factor:                stratum
```

```
Number of strata:                       75
Components for variance calculation:    Between sampling units
Confidence interval method:             tdistribution (95% limits)
Total number of responses:              1776
Survey weights:                         weight
Weights range:                          Min = 1.483  Mean = 34.71  Max = 185.8
Sum of weights:                         61653
Note: statistics above relate to the whole sample, not just the subsetdefined by the
restriction


Means for subpopulation defined by restriction in farmincome with 95% confidence limits
---------------------------------------------------------------------------------------

                    n   Sum wts      Mean      s.e.   %RSE/CV     Lower     Upper
    education
  school only     526     19874     13807      1510     10.93     10846     16768
         GCSE     230      8536     30082     11729     38.99      7078     53087
     A levels     121      4123     20041      3081     15.37     13997     26084
      college     511     16356     20886      1680      8.04     17590     24181
       degree     222      6789     38041      5063     13.31     28110     47972
     postgrad      41      1645      9757      4682     47.98       574     18940
   apprentice      36      1323     15941      3389     21.26      9294     22587
        other      36      1094     25402      8467     33.33      8796     42008
         Mean    1723     59740     21403      1884      8.80     17708     25098
```

The `summary of analysis` section is identical to that in the previous section, since this relates to the population as a whole. However, in the section headed `'means for subpopulation...'` the sample size (`n`) and sum of weights for the overall mean are less than those in the full population; reference to the previous section will show that this row is identical to that for male farmers, confirming that the analysis is now confined to male farmers only.

3.4 Practical

Construct tables of `farmincome` tabulated by `sex` for farmers in the education category `school only` and, separately, for those with college education. Save the means and their standard errors in suitably named tables by clicking on the **Store** button and display them in spreadsheets next to each other in order to make it easy to make comparisons.

3.5 Counts and proportions

So far, all the analyses in this section have aimed to estimate means or totals, but sometimes we may instead want to estimate the proportion of the population that has a particular characteristic. For example, as a result of the analysis of farm income by sex in Section 3.2, we might be interested in the proportion of farmers who are women. To answer this question we rerun the analysis, but with the **Data** box left blank (Figure 3.8). Genstat then produces the following results.

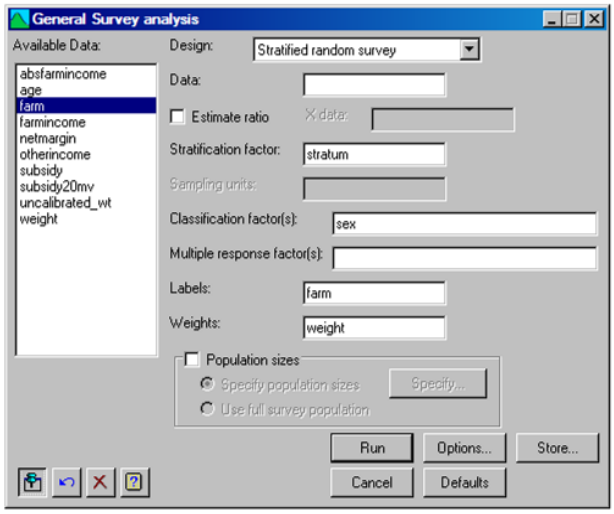


Figure 3.8

```
Survey analysis results
=======================

Summary of analysis
-------------------

Y-variate (response data):                Count
Method:                                   Design-based (expansion)
Stratification factor:                    stratum
Number of strata:                         75
Components for variance calculation:      Between sampling units
Confidence interval method:               tdistribution (95% limits)
Total number of responses:                1776
Survey weights:                           weight
Weights range:                            Min = 1.483  Mean = 34.71  Max = 185.8
Sum of weights:                           61653


Counts with 95% confidence limits
---------------------------------
```

```
                     n   Sum wts     Total      s.e.   %RSE/CV     Lower     Upper
 05farmer.gender
        male      1723     59740     59740     573.7      0.96     58615     60866
      female        53      1913      1913     302.4     15.81      1320      2506
       Total      1776     61653     61653         *         *         *         *

Standard errors based on Taylor series approximations. Confidence limits use t-
distribution with 1701 d.f.


Proportions with 95% confidence limits
--------------------------------------

                     n   Sum wts      Mean      s.e.   %RSE/CV     Lower     Upper
 05farmer.gender
        male      1723     59740    0.9690  0.004905      0.51    0.9594    0.9786
      female        53      1913    0.0310  0.004905     15.81    0.0214    0.0406
        Mean      1776     61653    1.0000  0.000000      0.00    1.0000    1.0000

Standard errors based on Taylor series approximations. Confidence limits use t-
distribution with 1701 d.f.
```

Notice that **Totals** now produces tables of counts, whilst **Means** produces proportions. The **Counts** are equal to the sum of the weights shown in the analyses in Section 3.2 (Genstat has effectively analysed a variate with a value of 1.00 for each unit), but they are now accompanied by standard errors and confidence limits. The **Counts** for the grand total have no standard error, as the number of units (farms in this case) in the population is always taken to be a known constant; in practice it also subject to error, but these errors are not a consequence of the sample design of the current survey and so cannot be estimated from it.

When two or more classification factors are specified, the proportions are expressed relative to the grand total. For example, an analysis by `sex` and `education` shows that 0.009 (i.e. just under 1%) of farmers in the population are female with a degree. If instead we wish to know what proportion of farmers with a degree are female, it is necessary to first restrict the analysis to farmers with a degree, and then to re-run the analysis as specified in Figure 3.8.

3.6 Ratios

The **General Survey Analysis** menu (`SVTABULATE` command) can also estimate ratios, although, unlike the **Single-stage Survey Analysis** menu examined in Chapter 2, it cannot use these ratios directly to estimate a population total. To demonstrate this, we will estimate the ratio of `subsidy` to `farmincome` for farms in England. A complication is that many farms had negative farm incomes for the year of the survey. So we will restrict the analysis to those with a farm income greater than zero, using the **By Value** sub-option from the **Restrict/Filter** option of the **Spread** menu (Figure 3.9).

To specify the ratio analysis, the **Estimate ratio** box should be ticked and `farmincome` clicked across to the **X data** box (i.e. the denominator of the ratio), with `subsidy` in the data box (numerator, see Figure 3.10). `Farmsize` has been specified as the **Classification factor** and the output is shown below.

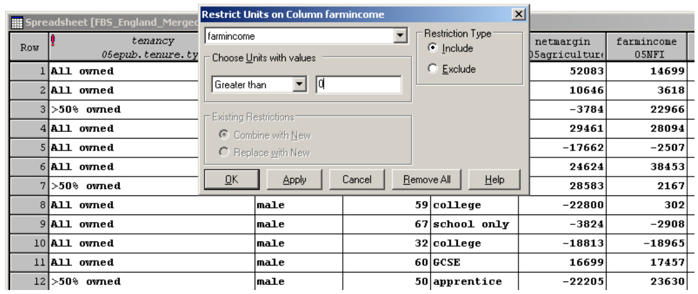


Figure 3.9

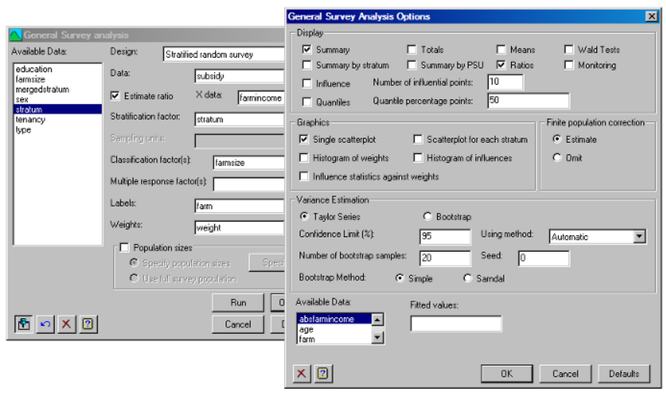


Figure 3.10

```
Survey analysis results
=======================

Summary of analysis
-------------------

Y-variate (response data):             subsidy
X-variate:                             farmincome
Correlation:                           0.109
Method:                                Design-based (expansion)
Stratification factor:                 stratum
Number of strata:                      75
Components for variance calculation:   Between sampling units
```

```
Confidence interval method:             tdistribution (95% limits)
Total number of responses:              1776
Survey weights:                         weight
Weights range:                          Min = 1.483  Mean = 34.71  Max = 185.8
Sum of weights:                         61653
Note: statistics above relate to the whole sample, not just the subsetdefined by
the restriction
```

Ratios for subpopulation defined by restriction in subsidy with 95% confidence limits

farmsize	n	Sum wts	Ratio	s.e.	%RSE/CV	Lower	Upper
Part-time	157	11403	0.8863	0.06970	7.86	0.7496	1.0230
Small	375	16878	0.6468	0.15203	23.50	0.3486	0.9450
Medium	309	8276	0.8592	0.06757	7.86	0.7266	0.9917
Large	276	5400	0.8115	0.05119	6.31	0.7111	0.9119
Very large	296	5197	0.5545	0.04173	7.52	0.4727	0.6364
Margin	1413	47154	0.6970	0.04944	7.09	0.6000	0.7940

Standard errors based on Taylor series approximations. Confidence limits use t-distribution with 1701 d.f.

When interpreting ratios such as these, it is always wise to plot a scatter plot of the two variables, since the mean ratios shown in the table may reflect more complex relationships between the variables. Influence statistics are also available for the estimation of ratios and are once again useful in detecting outliers; when X data are provided these are calculated as the percentage change in the estimate of the ratio when the observation is replaced by a missing value.

Figure 3.11 shows the scatter plot produced by ticking the **Single scatterplot** option on the **General Survey Analysis Option** menu (Figure 3.10). The scatter plots are plotted on the log-scale (except where negative values are present)

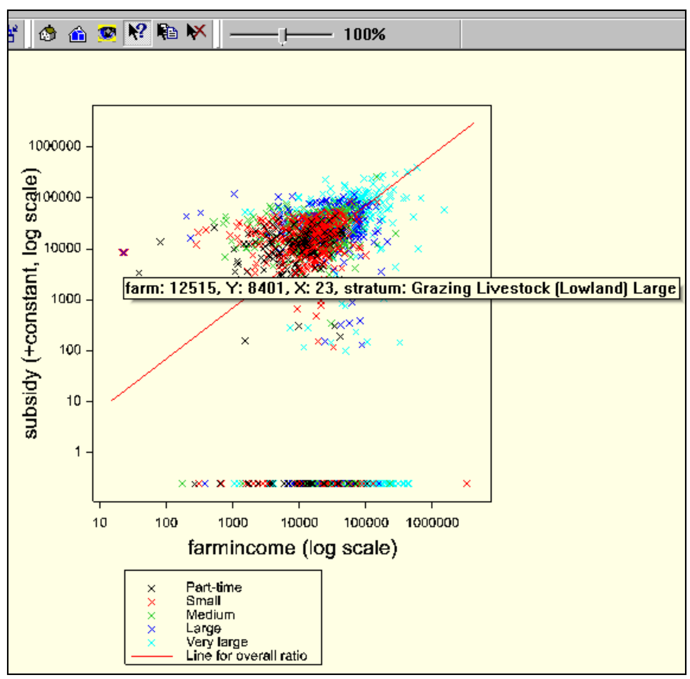


Figure 3.11

since survey variables are frequently strongly skewed, as is the case here. The line representing the relationship described by the overall ratio (i.e. $y = 0.697x$ in this case) is shown on the graph; alternatively plots of the ratios for each level of the classification factor(s) can be obtained by ticking the box for **Scatterplot for each stratum**. A number of features are apparent in Figure 3.11. The overall correlation is not very high; the summary of analysis shown above indicates that the correlation is 0.109, but this is for the full dataset, not just the sub-population with positive farm incomes. A few points have extremely high ratios of subsidy for income, including the one that has been highlighted using the **Data info** button (arrow and question mark) on the toolbar. The information includes the variable from the **Labels** box (or the row number if this is blank), allowing the data to be checked in the spreadsheet if necessary. At the other extreme there is a row of points along the bottom of the graph, representing farms with no subsidy claim; to allow these points to be shown on the log scale, a small constant has been added to them.

3.7 Quantiles and bootstrapping

It is apparent from the previous sections that the distribution of the farm income data is markedly non-Normal. The distribution is skewed to the left, with a few very large values. In this respect it is rather like a log-Normal distribution, but there is also a significant number of negative values. In situations like this, comparisons between means may give an over-simplified picture of the true differences between groups. A more complete assessment can be made by looking at the quantiles of the distribution and Figure 3.12 shows how this may be done. The output is shown below.

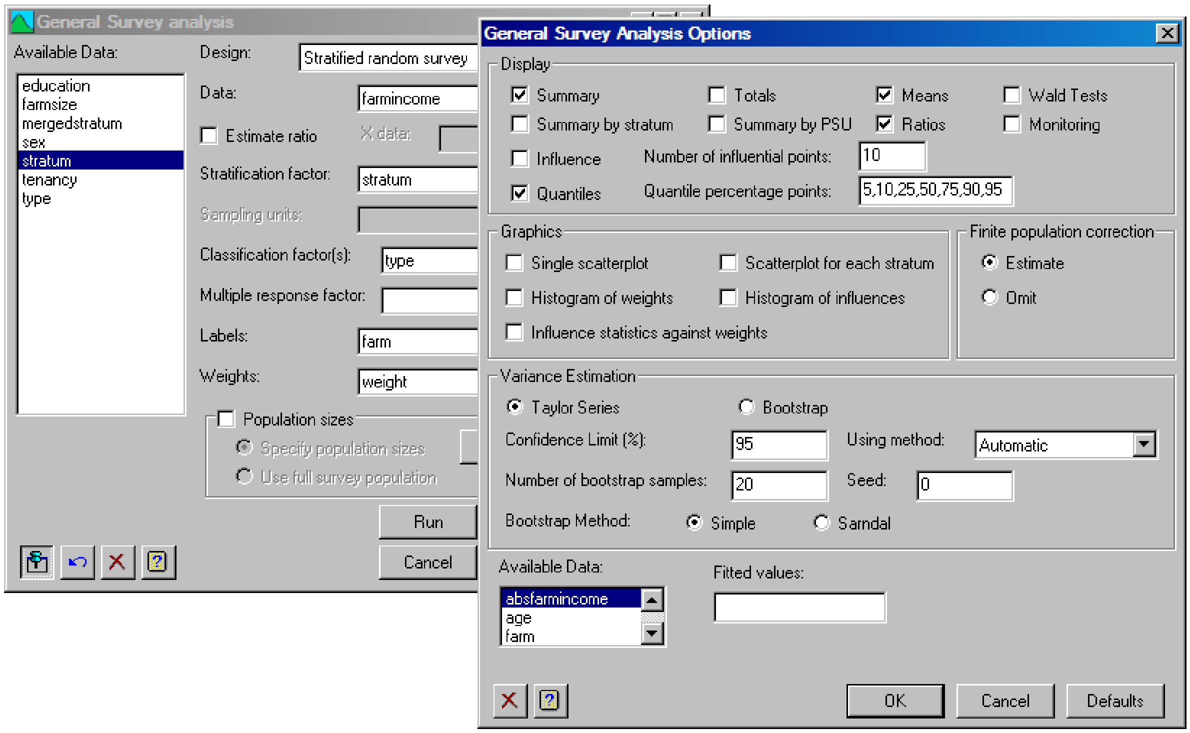


Figure 3.12

Means with 95% confidence limits

05farm.type	n	Sum wts	Mean	s.e.	%RSE/CV	Lower	Upper
Dairy	290	12289	27064	1751	6.47	23629	30499
Upland Grazing Livestock	234	5974	11775	1244	10.57	9335	14216
Lowland Grazing Livestock	221	8835	5265	984	18.68	3336	7194
Cereals	339	13125	14084	1955	13.88	10250	17918
General cropping	188	6589	26678	3847	14.42	19133	34224
Pigs	60	1156	29032	5849	20.15	17561	40503
Poultry	64	1643	97532	60195	61.72	-20532	215596
Mixed	177	6176	17385	3162	18.19	11184	23586
Horticulture	203	5866	32710	4441	13.58	23999	41421
Mean	1776	61653	21137	1830	8.66	17547	24726

Standard errors based on Taylor series approximations. Confidence limits use t-distribution with 1701 d.f.

Quantiles

05farm.type	q5%	q10%	q25%	q50%	q75%	q90%	q95%
Dairy	-8028	1100	9003	18216	32808	65184	92758
Upland Grazing Livestock	-13230	-5698	1178	9211	17974	30477	40681
Lowland Grazing Livestock	-13434	-7844	-2801	3871	11265	21796	28089
Cereals	-33265	-21439	-3627	8768	27385	50521	66395
General cropping	-24768	-11164	3246	16593	35080	74370	97450
Pigs	-28237	-17698	-2933	17032	48635	100309	155137
Poultry	-71528	-10013	5849	24971	67515	128713	186460
Mixed	-31761	-24042	-2335	11403	23968	72628	92128
Horticulture	-23058	-13406	1742	12950	41754	72524	136013
Margin	-23400	-11501	1034	11683	27495	55663	84133

The definition of a quantile is that the specified percentage of the population is less than or equal to the value shown. Thus the table indicates that 25% of dairy farms have an income of £9003 or lower. The 50% quantile (q50%) is also known as the median, whilst the 25% and 75% values are the lower and upper quartiles. In the current example, the importance of looking at the quantiles can be seen by comparing the means and medians between dairy and horticultural farms. The mean is slightly higher (although not significantly so) for horticultural farms, but the median is markedly higher for dairy farms; the horticultural farm mean is being strongly influenced by a minority of very large enterprises; 5% have incomes above £136,000.

The table of quantiles above does not show standard errors. This is because the Taylor-series approximation used to estimate variances for the other statistics is not applicable to quantiles. When standard error estimates are required, Genstat can calculate them by *bootstrapping*. Bootstrapping involves sampling with replacement from the original sample in each stratum to form a large number of bootstrap populations. The relevant statistics are then calculated for each bootstrap sample and estimates of the standard errors are derived from the variance of the distribution of these bootstrapped estimates. Alternatively, if sufficient bootstrap samples are used (ideally several thousand), confidence limits can be determined directly from the distribution of the bootstrapped estimates.

Two basic methods of bootstrapping are provided within Genstat. The *simple* method is the approach used in non-survey settings in which observations are selected at random, with replacement, from the original sample ignoring the survey weights. When weights vary within a stratum, each observation remains associated with its weight, so that the sum of the weights in each bootstrap sample will not be exactly equal to the sum of the weights in the original populations. This approach ignores the finite nature of the population, but this is seldom a problem in practice, except when the sampling proportion is very high.

The second method is known as *sarndal*[3] and involves first constructing a pseudo-population, with each unit being replicated *w* times, where *w* is the appropriate weight, rounded to the nearest integer. For stratified designs, the process is carried out separately in each stratum. Sampling is then carried out, without replacement, using the inverse of the weights as inclusion probabilities. For reasons of computational simplicity, the bootstrap sample sizes are not fixed, and will therefore differ slightly from the one in the original sample. This method takes account of the finite nature of the population, but it is computationally slower.

Figure 3.13 shows the settings for bootstrapping the tables of farm incomes for each farm type. Two hundred bootstrap samples have been specified, and the **Using method** list box has been left at the default of **Automatic**; this forms confidence limits from the t-distribution, using standard error from the bootstrapped samples, when less than four hundred bootstrapped samples are used, but otherwise uses percentile limits. The **Seed** option has been left at its default of zero; this option should be set to a number with four or more digits if you want to be able to repeat the analysis and obtain identical results. If it is left at zero, a fresh set of random numbers is used to construct

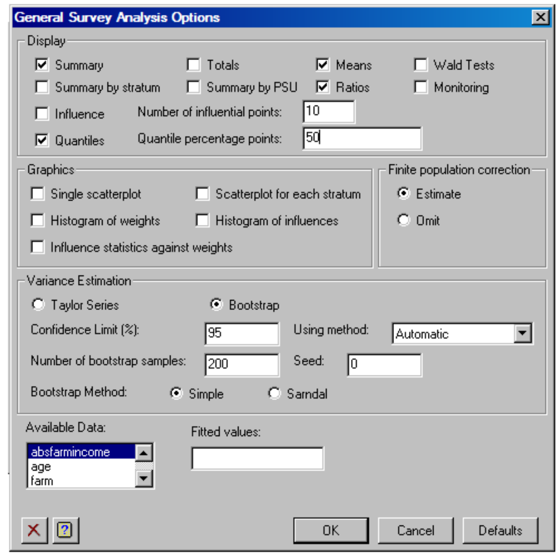


Figure 3.13

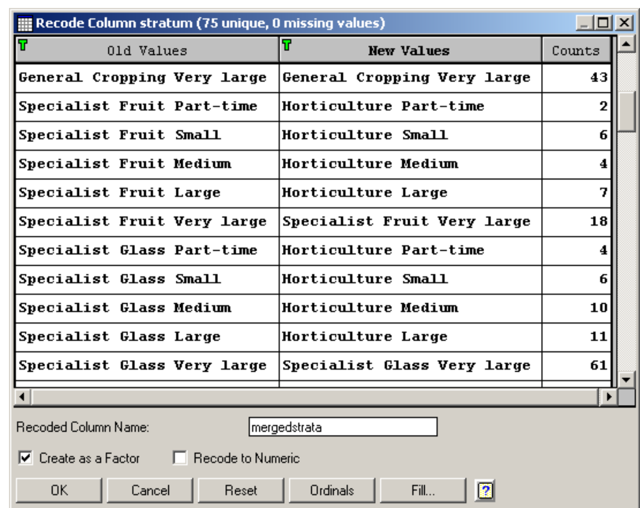


Figure 3.14

[3] Sarndal, C., Swensson, B. & Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag, New York. See page 442.

the bootstrapped samples, so that slightly different results will be produced each time the command is run.

One complication with the analysis is that bootstrapping requires a reasonable sample size in each stratum to produce reliable results. The FBS dataset contains some very small strata, and so it is best to form a new stratification variable, combining the smaller strata where necessary, before using bootstrapping. To achieve this, **Recode** should be selected from the **Factor** sub-menu of the **Spread** menu, with the cursor in the existing `stratum` factor. The strata can then be combined as required. Figure 3.14 shows this process; the specialist fruit and glass categories have been edited to combine them into size categories for all horticulture, with the exception of the very large-size categories, where sample sizes are more reasonable.

Results of the analysis are shown below.

```
Survey analysis results
=======================

Summary of analysis
-------------------

Y-variate (response data):              farmincome
Method:                                 Design-based (expansion)
Stratification factor:                  mergedstratum
Number of strata:                       48
Components for variance calculation:    Resampling sampling units
Bootstrap method:                       simple
Number of bootstrap samples             200
Confidence interval method:             tdistribution (95% limits)
Total number of responses:              1776
Survey weights:                         weight
Weights range:                          Min = 1.483  Mean = 34.71  Max = 185.8
Sum of weights:                         61653


Means with 95% confidence limits
--------------------------------

                                n   Sum wts     Mean     s.e.    %RSE/CV     Lower    Upper
                    type
                   Dairy      290     12289    27064     1937       7.16     23266    30863
 Upland Grazing Livestock     234      5974    11775     1340      11.38      9147    14404
Lowland Grazing Livestock     221      8835     5265      960      18.24      3382     7148
                 Cereals      339     13125    14084     1848      13.12     10459    17708
        General cropping      188      6589    26678     3787      14.19     19251    34105
                    Pigs       60      1156    29032     6574      22.64     16138    41925
                 Poultry       64      1643    97532    59159      60.66    -18499   213563
                   Mixed      177      6176    17385     3423      19.69     10671    24099
            Horticulture      203      5866    32710     4497      13.75     23890    41530
                    Mean     1776     61653    21137     1786       8.45     17633    24640
```

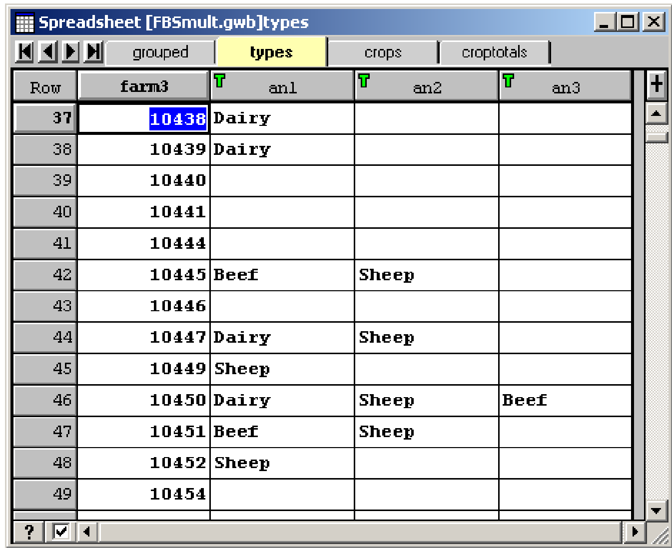
```
Standard errors based on 200 bootstrap samples. Confidence limits use t-distribution with 1728
d.f.


Quantiles with 95% confidence limits
------------------------------------

                                  q50%         s.e.        Lower        Upper
                     type
                    Dairy        18216         1598        15083        21349
 Upland Grazing Livestock         9211         1819         5643        12779
Lowland Grazing Livestock         3871          737         2426         5316
                  Cereals         8768         1730         5375        12161
         General cropping        16593         1413        13822        19364
                     Pigs        17032         7782         1768        32296
                  Poultry        24971         5566        14055        35887
                    Mixed        11403         2190         7107        15699
             Horticulture        12950         3295         6487        19413
                   Margin        11683          597        10511        12855

Standard errors based on 200 bootstrap samples. Confidence limits use t-distribution with 1728
d.f.
```

3.8 Multiple-response tables

All the classification factors used in the analyses up to this point have had a single value for each unit. Thus, for example, farms have been classified to the most appropriate type on the basis of their activities. A farm with both dairy cattle and cereal crops, will be classified to one group or the other, depending on which enterprise is more economically important; it cannot be in both the dairy and cereals categories simultaneously.

Sometimes it is more helpful to form tables classified by a *multiple-response factor*, where each unit can contribute to two or more cells in the same table. For example, suppose that in a questionnaire respondents are asked to state which languages they can speak and have a number of boxes in which to respond. Using multiple-response factors a table can be formed with a row for each language, so that, for example, some people contribute to both the `French` and `German` rows. More details on how Genstat handles multiple-response factors can be found in the *Syntax and Data Management Guide* (available from the **Genstat Guides** option on the **Help** menu.

In this section we will concentrate on how the **General Survey Analysis** menu can be used to form tables from multiple-response data from surveys, using the FBS dataset as an example. The data describe the types of livestock found on each farm, and may be found in the file `FBSmult.gwb`; note that this is a Genstat *workbook* with several different worksheets within it, whereas the files that we have used previously are .GSH files containing a single worksheet. For illustration purposes, the worksheets `grouped` and `types` present the same data in two alternative formats. We shall start by examining the data in sheet `types` (Figure 3.15). This is the format that would arise if farmers were asked which livestock they had on the farm, and given three different text boxes to record their results. The available responses are dairy (cattle), beef (cattle), sheep or pigs. The data in the spreadsheet are in text columns (note the green T by the variable names); they could equally well be in factors, but the next step requires the data as texts, so they should be converted to texts before proceeding.



Row	farm3	an1	an2	an3
37	10438	Dairy		
38	10439	Dairy		
39	10440			
40	10441			
41	10444			
42	10445	Beef	Sheep	
43	10446			
44	10447	Dairy	Sheep	
45	10449	Sheep		
46	10450	Dairy	Sheep	Beef
47	10451	Beef	Sheep	
48	10452	Sheep		
49	10454			

Figure 3.15

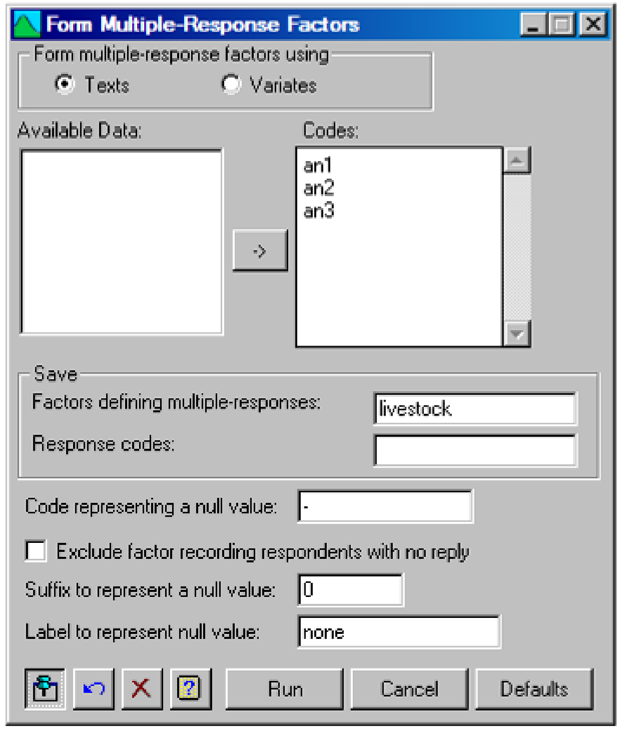
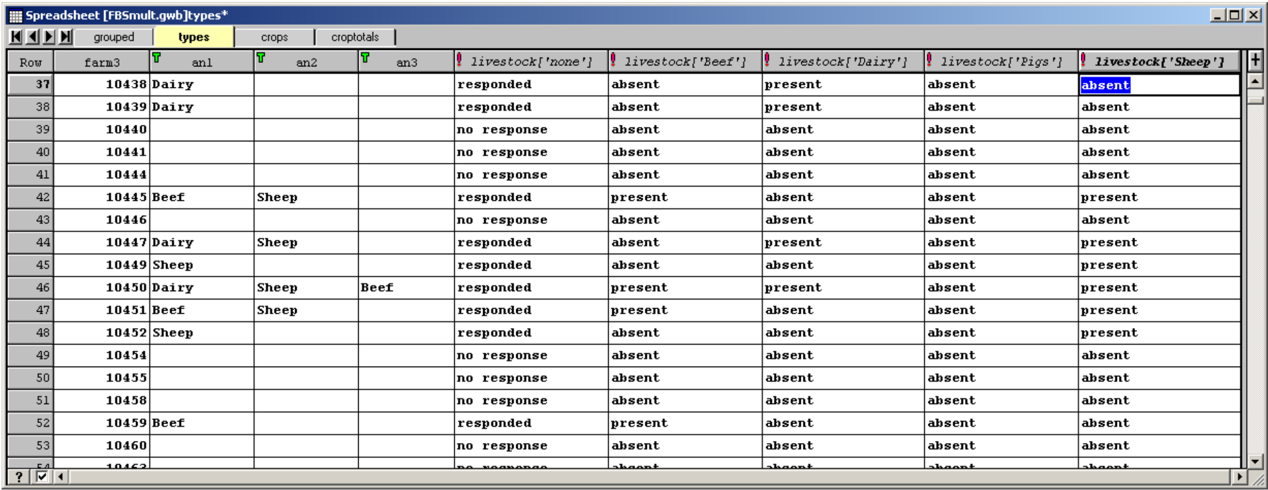


Figure 3.16

To form the multiple-response factors, select **Form Multiple-Response Factors** from the **Data** menu to open the **Form Multiple-Response Factors** menu (Figure 3.16). The three text structures are clicked across, suitable names are given for the new factors to be created, and labels are defined to represent a null value.

Whilst not strictly necessary for the analysis, it is useful to add the new multiple-response factors to the spreadsheet (**Data in Genstat** from the **Add** option of the **Spread** menu), in order to understand how Genstat stores the information. Genstat creates a series of five new factors, four for the different types of livestock and one for null responses. All the factors have the levels 0 and 1, with 1's being represented by the factor label `present` for the livestock types and `no response` for the null factor (Figure 3.17).



Spreadsheet [FBSmult.gwb]types*

grouped | types | crops | croptotals

Row	farm3	an1	an2	an3	livestock['none']	livestock['Beef']	livestock['Dairy']	livestock['Pigs']	livestock['Sheep']
37	10438	Dairy			responded	absent	present	absent	absent
38	10439	Dairy			responded	absent	present	absent	absent
39	10440				no response	absent	absent	absent	absent
40	10441				no response	absent	absent	absent	absent
41	10444				no response	absent	absent	absent	absent
42	10445	Beef	Sheep		responded	present	absent	absent	present
43	10446				no response	absent	absent	absent	absent
44	10447	Dairy	Sheep		responded	absent	present	absent	present
45	10449	Sheep			responded	absent	absent	absent	present
46	10450	Dairy	Sheep	Beef	responded	present	present	absent	present
47	10451	Beef	Sheep		responded	present	absent	absent	present
48	10452	Sheep			responded	absent	absent	absent	present
49	10454				no response	absent	absent	absent	absent
50	10455				no response	absent	absent	absent	absent
51	10458				no response	absent	absent	absent	absent
52	10459	Beef			responded	present	absent	absent	absent
53	10460				no response	absent	absent	absent	absent

Figure 3.17

When data are supplied in a separate spreadsheet to the main data, it is essential to check that they are correctly matched, since mismatched data (e.g. if one sheet has been sorted by farm type and the other by farm number) are a frequent cause of errors. One option is to merge the two spreadsheets as in Section 3.1. In other cases it may be preferable to keep them separate, particularly if the dataset is so large that a merged file would be excessively big. When this is

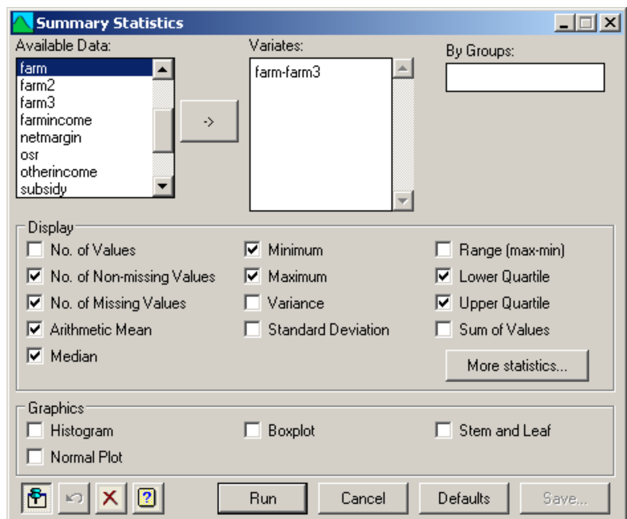


Figure 3.18

a case, a check should always be carried out before analysis. There are various ways of doing this, but one option is to use **Summary Statistics** from the **Summary Statistics** sub-option on the **Stats** menu. This is shown in Figure 3.18. To avoid calculating a new variable, the expression `farm-farm3` has been typed in the **Variates** box, `farm` being the farm identifier in the main dataset, and `farm3` the identifier in the multiple dataset. Results are shown below: as expected, the calculation always produces a result of zero indicating that the datasets are correctly matched.

```
Summary statistics for Y[1]
===========================

      Number of observations = 1776
    Number of missing values = 0
                        Mean = 0
                      Median = 0
                     Minimum = 0
                     Maximum = 0
              Lower quartile = 0
              Upper quartile = 0
```

The analysis can now be specified using the **General Survey Analysis** menu (Figure 3.19). The means produced are shown below.

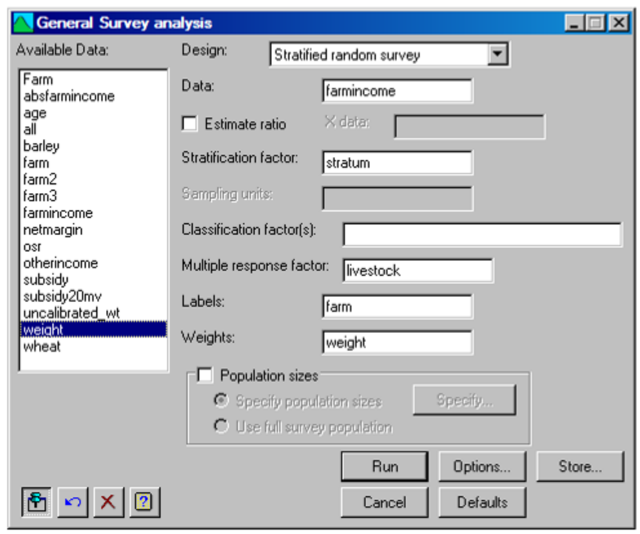


Figure 3.19

```
Means with 95% confidence limits
--------------------------------

                     n   Sum wts      Mean      s.e.   %RSE/CV     Lower     Upper
     mrfac[1]
        none       893     31783     24882      3457     13.89     18101     31663
         Beef      446     14308     10787      1272     11.79      8292     13283
        Dairy      257      9950     29394      2194      7.46     25092     33697
         Pigs       61      1564     27272      6819     25.00     13897     40646
        Sheep      510     14778     11819      1079      9.13      9702     13935
         Mean     1776     61653     21137      1830      8.66     17547     24726

Standard errors based on Taylor series approximations. Confidence limits use t-
distribution with 1701 d.f.
```

Notice that the sums of the numbers of observations (`n`) and the weights (`Sum wts`) are now higher than in the margin of the table (row labelled `mean`). This is because all farms are represented at least once in the individual rows, but those with more than one livestock type are included in two or more rows.

3.9 Two-stage samples

Whilst many surveys employ a single level of sampling, in others two or more levels are used. Sometimes this is necessary because a complete sampling frame is unavailable. For example, in a survey of educational performance, we may lack the complete list of all pupils in all schools (*sampling frame*) that would be needed to sample by random, or stratified random, sampling. However, if a complete list of schools exists, we can sample from these at random and then obtain pupil lists from the selected schools in order to implement a second stage of sampling to select pupils within each of these schools.

With increasing computerization of administrative data, particularly in industrialized countries, a complete sampling frame is more often available, thus allowing the use of single-stage sampling. For a given sample size, a single-stage survey will nearly always be more precise than a two-stage one. However, a two-stage approach may still be the most cost-effective solution when there are substantial overheads that are proportional to the number of higher level units. To return to the educational survey example, if we used a simple random sample of one hundred pupils, these might come from many different schools, making the survey expensive if visits were needed to each school. For the same cost it might be possible to sample, for example, twenty pupils from each of ten schools, using a two-stage design. In this situation the increased number of pupils in the two-stage design might well outweigh the inherent inefficiency of the design.

File `Malawi7.gsh` contains data from a multi-stage survey of households in Malawi[4]. A minimum of three Extension Planning Areas (EPAs) were selected at random from the seven Agricultural Development Divisions (ADDs), and then two villages were selected at random from each EPA. It is thus a two-stage stratified design, with the ADDs being the strata, the EPA the first stage (primary) sampling units, and villages as the secondary sampling units. Weights are supplied in this file; we shall demonstrate how they are calculated in the next chapter.

Figure 3.20 shows the **General Survey Analysis** menu for analysis of the number of households enumerated in each village (column `GTIS-hh`). Notice that `ADD` is listed in the **Classification factor(s)** box as well as the **Stratification factor** one; if this was not done the same estimate of the grand total would be produced, but the output table would not be classified by `ADD`. When the **Run** button is clicked the following warning appears.

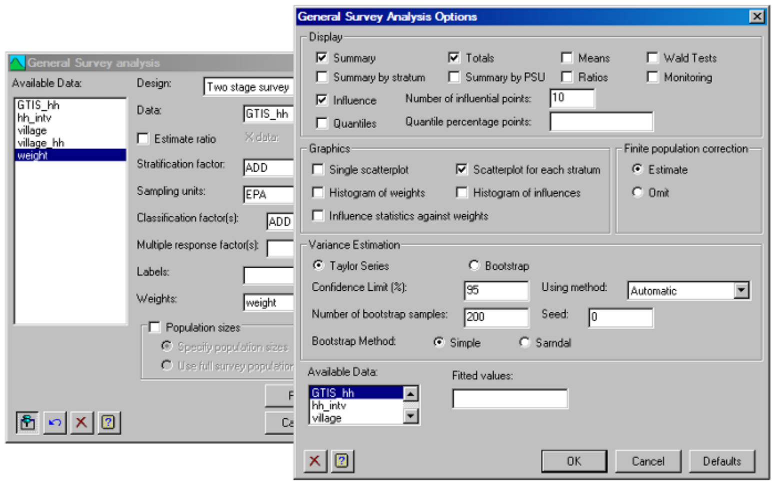


Figure 3.20

```
******** Warning 12, code UF 2, statement 292 in procedure SVTABULATE

Insufficient information to calculate FPC.
```

[4] Data from the Malawi Ground Truth Investigation Study are supplied by permission of Dr Roger Stern, Statistical Services Centre, University of Reading, U.K. We have used data from seven of the eight strata (ADDs) where adequate numbers of secondary units were sampled.

Because only survey weights have been supplied, rather than full information on the number of primary units in each stratum and secondary units in each primary unit, Genstat cannot calculate the finite population correction (FPC) and it prints a warning to this effect. The warning can be suppressed, if desired, by clicking on the **Omit** button under **Finite population correction** on the options menu. In this situation Genstat uses the *ultimate clusters* form of analysis, basing the variance estimates only on the variance between primary units, ignoring the variance between secondary units, except insofar as it is reflected in the differences between primary units. This is a reasonable approach for large surveys if, as is frequently the case, the variance between secondary units is comparatively small.

Output is shown below and is basically similar to that produced from a single-stage survey, apart from the extra summary information relating to the primary sampling units (PSUs).

```
Survey analysis results
=======================

Summary of analysis
-------------------

Y-variate (response data):            GTIS_hh
Method:                               Design-based (expansion)
Stratification factor:                ADD
Number of strata:                     7
Primary sampling units:               EPA
Number of PSUs sampled:               26
Components for variance calculation:  Between PSUs (ultimate clusters)
Confidence interval method:           tdistribution (95% limits)
Total number of responses:            52
Survey weights:                       weight
Weights range:                        Min = 60.00  Mean = 429.5  Max = 1597
Sum of weights:                       22335

Totals with 95% confidence limits
---------------------------------

                    n   Sum wts     Total      s.e.   %RSE/CV     Lower     Upper
          ADD
     Blantyre       8      1775    350446    125578     35.83     87608    613285
      Karonga       6       696     77172     14089     18.26     47683    106661
      Kasungu       8      3958    177856     20648     11.61    134640    221072
     Lilongwe      10      8653    390058     65016     16.67    253977    526138
     Machinga       8      2524    239382    111709     46.67      5573    473191
        Mzuzu       6      3113    295730     68280     23.09    152818    438641
       Salima       6      1615    330997     52661     15.91    220777    441218
        Total      52     22335   1861641    201336     10.81   1440241   2283042

Standard errors based on Taylor series approximations. Confidence limits use t-
distribution with 19 d.f.
```

Whilst the ultimate clusters approach is often a reasonable approximation, it is generally preferable to include the contribution from variance between secondary units (EPAs) in the analysis. This can be done by supplying the number of EPAs in each `ADD`. (We could also supply the number of villages per EPA, but since the supplied weights are assumed to represent the inverse of the combined probability of selection at both stages, this information can be calculated from the number of EPAs per `ADD`.) This information is best supplied in a table classified by `ADD`. It is also possible to supply the figures in a variate with one row for each stratum. However, if this is done, great care must be taken to ensure that the strata are listed in the correct order.

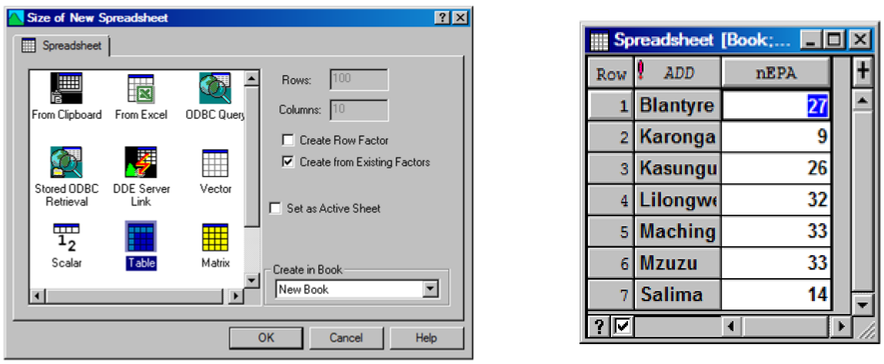


Figure 3.21

Figure 3.21 shows the process of creating a table. We first select **Create** from the **New** sub-option of the **Spread** menu, and then select **Table** and tick **Create from Existing Factors**. At the next menu we chose `ADD` as the classifying factor to produce a new spreadsheet. The relevant values can then be added into the table, as is shown in the right-hand image of Figure 3.21.

Once the table has been created it can be used to supply the population sizes by ticking the **Population sizes** box and clicking on the **Specify** button as is shown in Figure 3.22. The results below show that specifying the full design in this way causes a substantial change in the variance estimates in this example.

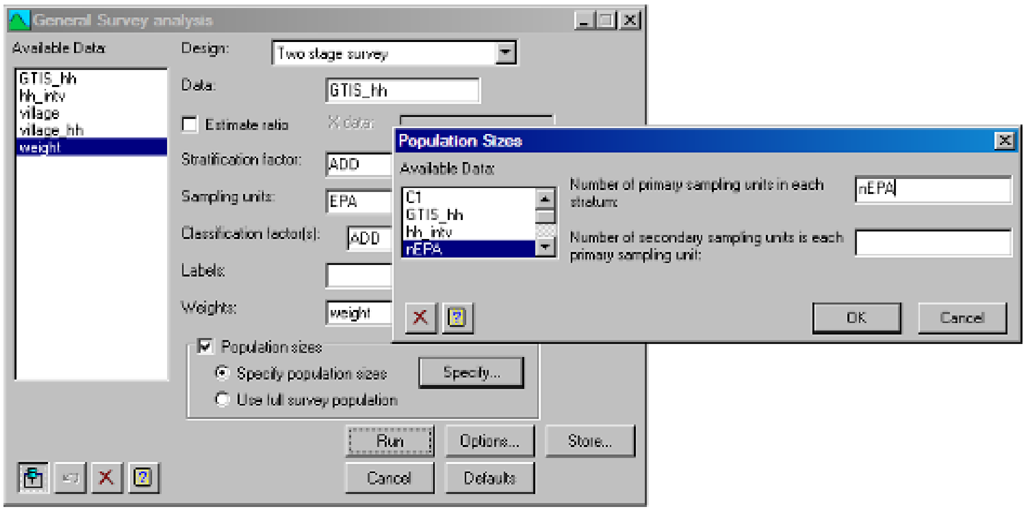


Figure 3.22

```
Survey analysis results
=======================

Summary of analysis
-------------------

Y-variate (response data):                GTIS_hh
Method:                                   Design-based (expansion)
Stratification factor:                    ADD
Number of strata:                         7
Primary sampling units:                   EPA
Number of PSUs sampled:                   26
Components for variance calculation:      Between PSUs & within PSUs
Confidence interval method:               tdistribution (95% limits)
Total number of responses:                52
Survey weights:                           weight
Weights range:                            Min = 60.00  Mean = 429.5  Max = 1597
Sum of weights:                           22335
```

```
Totals with 95% confidence limits
---------------------------------

                   n   Sum wts     Total      s.e.   %RSE/CV     Lower     Upper
         ADD
    Blantyre       8      1775    350446    125704     35.87     87344    613549
     Karonga       6       696     77172     18441     23.90     38575    115769
     Kasungu       8      3958    177856     34478     19.39    105693    250019
    Lilongwe      10      8653    390058     86528     22.18    208953    571162
    Machinga       8      2524    239382    130909     54.69    -34613    513377
       Mzuzu       6      3113    295730     74636     25.24    139514    451945
      Salima       6      1615    330997     77638     23.46    168499    493496
       Total      52     22335   1861641    231415     12.43   1377285   2345998

Standard errors based on Taylor series approximations. Confidence limits use t-
distribution with 19 d.f.
```

4 Weights and imputation

In the previous chapter all the datasets included a column of survey weights, so there was no need to calculate them prior to analysis. This is frequently how complex datasets are supplied to researchers for further analysis. However, if you are analysing a survey from the outset, you may need to calculate a set of survey weights before using the methods in Chapter 3. It is possible to avoid calculating weights explicitly by using the **Population sizes** box on the **General Survey Analysis** menu. However, this is generally sensible only for small surveys, or for single-stage surveys where the methods described in Chapter 2 are adequate. For larger surveys with many variables it is usually easier to calculate the weights, not least because there will often be a need to modify them in some way, for example to deal with unusual observations.

In this chapter you will learn how to create survey weights, how to modify them to allow for outliers or missing data, and how to use calibration weighting to ensure that they reflect known population totals. You will also learn how imputation can be used to allow for missing values in a dataset.

4.1 Creating survey weights

We shall illustrate how to create survey weights using the June Agricultural Survey data introduced in Chapter 2. File `Juneresponse.gwb` contains two sheets; sheet `response` contains figures from those farms that were selected for and responded to the survey, whilst sheet `nfarm` holds a table showing the total number of farms in each stratum of the survey population. See Section 1.6 for details of how to create such tables.

Open the file so that the data are sent to the Genstat server, and then open the **Create Survey Weights** menu from the **Survey Analysis** option on the **Stats** menu.

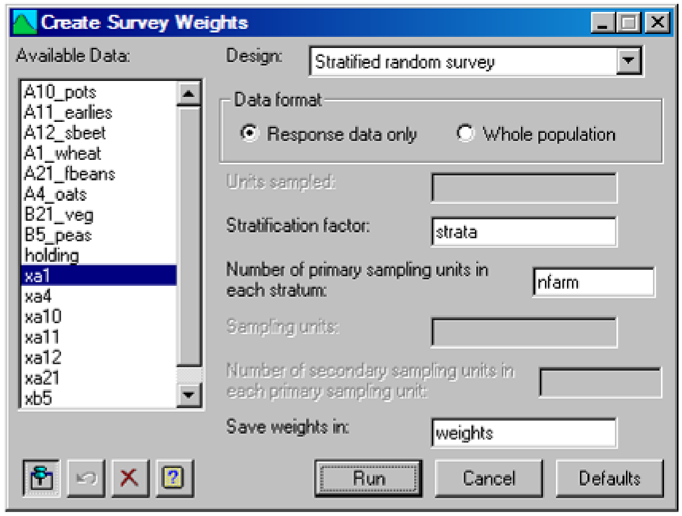


Figure 4.1

Figure 4.1 shows this menu with the appropriate settings. Since a stratified random survey is specified, the boxes relating to sampling units are greyed out, but data can be entered in these in the same way for two-stage designs. When the **Run** button is pressed a brief summary of the weights is created in the output window.

```
Create Survey Weights
=====================

Summary of weights
------------------

Survey weights:     weights
Weights range:      Min = 1.750 Mean =3.478 Max = 5.898
Sum of weights:     19156


Weights summary by stratum
--------------------------

                  mean wt
       strata
          new       2.131
        small       5.898
       medium       4.883
        large       3.249
   very large       1.750
Weights are constant within each stratum
```

To understand where these weights come from, it is useful to display some of the output of the same data from the **Single-stage Survey Analysis** menu, originally displayed in Section 2.1:

	Total no. obs.	Imputed	Sample	Excluded	Sampling fraction
strata					
new	2613	1387	1226	0	0.469
small	5851	4859	992	0	0.170
medium	5479	4357	1122	0	0.205
large	3074	2128	946	0	0.308
very large	2139	917	1222	0	0.571
Total	19156	13648	5508	0	0.288

In this output the sampling fraction is the number of observations in the sample divided by the number of units (farms in this case) in the whole population; for example, for the `new` stratum 1226/2613=0.469. The weights calculated above are the inverses of the sampling fractions (i.e. 2613/1226 = 2.131 = 1/0.469); these are known as *probability weights*. It should be noticed that in this case, the 'sampling fraction' actually represents a combination of the processes of sampling and response (or non-response). Treating non-response in this way (as if it were really part of the sampling process) is common practice, and is valid if it is believed that non-response occurs approximately at random with respect to the variables to be analysed. It is an approach that should be used with caution when response rates are low, and it will produce biased results if the probability of response is related to the data analysed; for example, if holdings with large wheat areas are more likely to respond. More sophisticated forms of non-response adjustment are needed in these situations.

It is often useful to store the new weights in the main datafile. With the spreadsheet `Juneresponse.gwb` open at the `responses` sheet, select **Data in Genstat** from the **Add** option of the **Spread** menu. At the next menu, move `weights` across to the box on the right and click **Add**. Weights will be added to the far right-hand size of the spreadsheet, but it can be moved to the left, if desired, by hovering the mouse over the variable name so that the cursor changes to a hand, holding the left mouse button down and dragging it across.

4.2 Practical

Using the weights created above, analyse the wheat data (A1_wheat) with the **General Survey Analysis** menu. Verify that it gives the same results as those shown in Section 2.1. You may notice that the confidence limits are very slightly different. This is because different approximations are used to calculate the degrees of freedom for the t-statistic; the approximation used by the **General Survey Analysis** menu (SVTABULATE command) is cruder, but is generally applicable[5].

4.3 Modifying weights for missing data

Sometimes survey respondents fail to supply data for all of the questions (item non-response). For example, the Juneresponse.gwb dataset contains a column berror which identifies units where anomalies were detected in the responses to section B of the survey during the validation process. It may therefore be sensible to exclude these units from the analysis of questions B5_peas and B21_veg.

One option, when we are interested in estimating a mean or a ratio, is simply to exclude these items from the analysis using a restriction (see Section 3.3). When we want to estimate the population total, this approach is not sensible, since estimates would relate to the subpopulation without such errors and hence would be biased downwards relative to the full population. Instead it is necessary to form a new set of weights, treating the units with anomalies as if they were unsampled, provided, of course, that it is reasonable to regard these units as being missing at random. This could be done by forming a new dataset of valid responses to section B of the survey, excluding the suspect data, and then repeating the process described in Section 4.1. However, it is generally preferable to use modified weights within the existing dataset, so that the suspect observations remain in the dataset, but are ignored in the analysis.

---

[5] SVSTRATIFIED uses the effective degrees of freedom described by, for example, Sampford (1962, *An introduction to sampling theory*) which weights the degrees of freedom according to each stratum's contribution to the variance. SVTABULATE takes d.f. as the total number of primary sampling units less the number of strata.

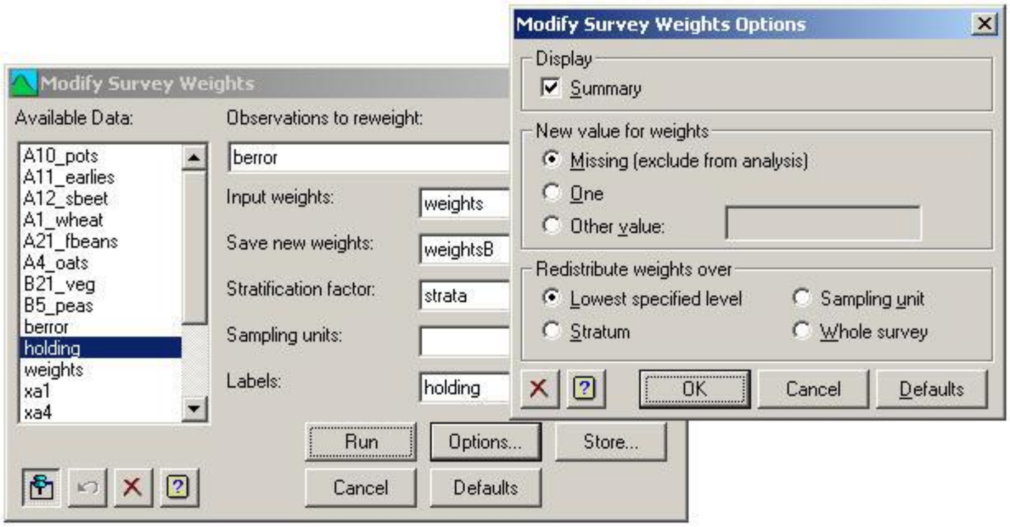


Figure 4.2

Figure 4.2 shows the **Modify Survey Weights** menu with the appropriate settings. The **Observations to reweight** box can be used to supply a list of the appropriate observations (see next section), but it is often easier to indicate these using a variate of 0's and 1's, where the 1's indicate the observations that need reweighting. This is precisely what column `berror` contains, and so it is clicked across into the box. The **Options** menu can be left with the standard default settings, as shown in Figure 4.2. Since the **Missing (exclude from analysis)** button is selected, the missing observations will have their weights set to missing values.

In order to ensure that the weights still produce estimates totals for the full population, the weights previously assigned to the observations now treated as missing must be redistributed to other observations. This reallocation may be done over the whole survey, within each stratum, or, in the case of a two-stage survey within each primary sampling unit. By default the **Lowest specified level** is used; in the case of a stratified random survey like this, that means that redistribution is within each stratum.

4.4 Modifying weights for outliers

In Section 2.3 we considered the various approaches for dealing with outliers in the context of the **Single-stage Survey** menu. Whilst the same principles apply to all surveys, the way of achieving the modified analyses is rather different using the **General Survey Analysis** menu.

It is worth making the point once again that, just because an observation is influential, it is not necessarily appropriate to adjust the analysis to reduce this influence. On the contrary, unless there is evidence to suggest that the record is erroneous, or in some way different to the rest of the population, the original analysis should stand. However, particularly with a statistically literate audience, one option may be to report results with and without the outlier, so that readers can judge the impact for themselves. The analysis without the outlier is obtained by treating the observation as missing, as in the previous section.

Sometimes it is required to retain an observation as a valid response but to reduce its weight. There are various methods that routinely use such modified weights in order to produce robust, but biased, estimates of population totals. We will not consider these methods here, but instead deal with the simpler situation where an observation although correct, is not considered representative of the wider population. We shall illustrate this using the June Survey dataset and considering the problem of how to estimate the ratio of between the area of wheat grown in the survey year and the area grown in the previous year. This is the same example that we used to illustrate outliers in Section 2.3, with the **Single-stage Survey Analysis** menu.

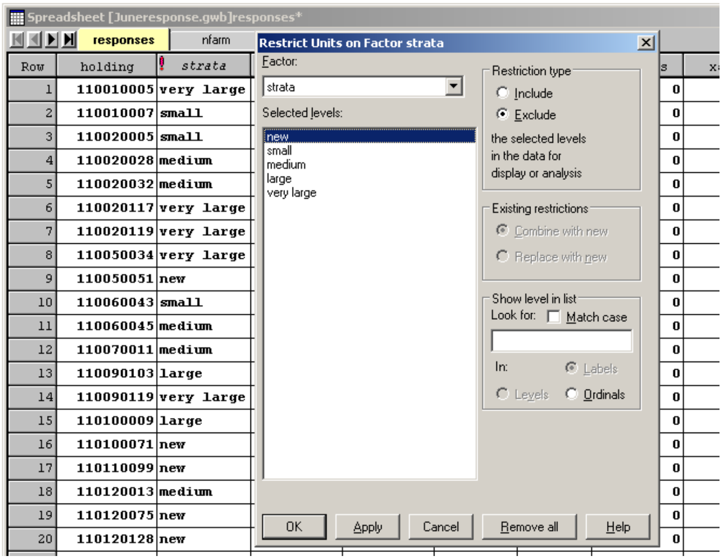


Figure 4.3

The analysis with all observations included is shown below. Because no previous crop areas are available for farms in the new stratum, the analysis must be restricted to the subpopulation excluding this stratum. This is achieved by selecting **To Groups (factor levels)** from the **Restrict/Filter** option on the **Spread** menu (Figure 4.3). The analysis is then produced using the settings shown in Figure 4.4.

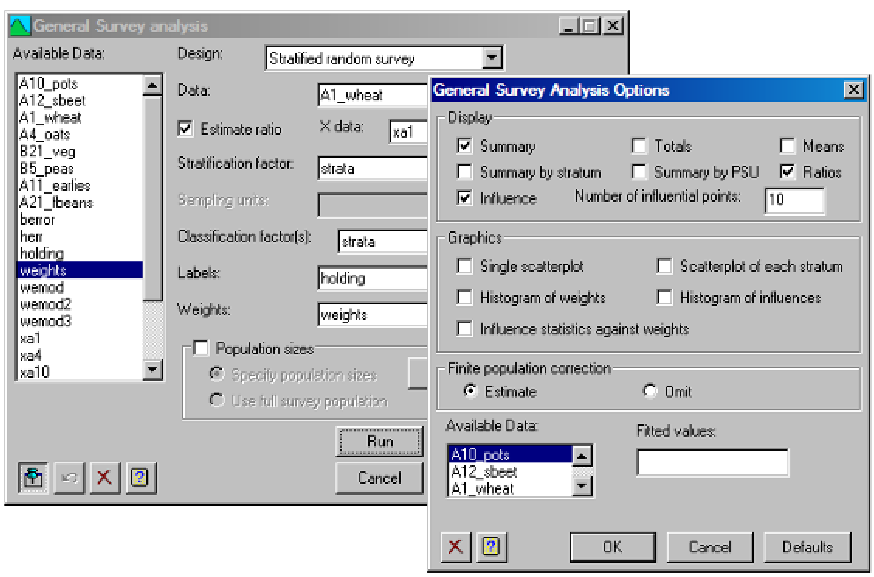


Figure 4.4

```
Survey analysis results
=======================

Summary of analysis
-------------------

Y-variate (response data):              A1_wheat
X-variate:                              xa1
Correlation:                            0.935
Method:                                 Design-based (expansion)
Stratification factor:                  strata
Number of strata:                       5
Components for variance calculation:    Between sampling units
Confidence interval method:             tdistribution (95% limits)
Total number of responses:              5508
Survey weights:                         weights
Weights range:                          Min = 1.750  Mean = 3.478  Max = 5.898
Sum of weights:                         19156
Note: statistics above relate to the whole sample, not just the subset defined by
the restriction


Ratios for subpopulation defined by restriction in A1_wheat with 95% confidence
limits
----------------------------------------------------------------------------------

                     n   Sum wts     Ratio      s.e.   %RSE/CV     Lower    Upper
       strata
          new        0         0         *         *         *         *        *
        small      992      5851    0.8209   0.04604      5.61    0.7307   0.9112
       medium     1122      5479    0.8593   0.02163      2.52    0.8169   0.9017
        large      946      3074    0.9047   0.01990      2.20    0.8657   0.9437
   very large     1222      2139    0.9124   0.00609      0.67    0.9004   0.9243
       Margin     4282     16543    0.8965   0.00772      0.86    0.8813   0.9116

Standard errors based on Taylor series approximations. Confidence limits use t-
distribution with 5503 d.f.


10 points with highest percentage influence on overall ratio
------------------------------------------------------------

holding      strata          Weight    A1_wheat        xa1  %influence
232480050    large            3.249        21.2      212.6      0.0852
232980220    very large       1.750         0.0      345.8      0.0844
281070004    medium           4.883       195.2       48.8      0.1147
343460118    large            3.249      1116.6      112.9      0.5087
344230042    large            3.249         0.0      263.0      0.1185
347310134    large            3.249         0.0      187.1      0.0844
383090082    large            3.249       330.0      136.0      0.1040
388090049    large            3.249       439.4       69.0      0.1889
614160015    very large       1.750       722.0      224.0      0.1400
615950014    large            3.249         0.0      216.7      0.0977

* Note: The influence value is the percent change in the estimate when the
observation is omitted
```

The observation with the highest influence is holding number 343460118, which increased its wheat area from just over a hundred hectares to well over a thousand. In fact, as described in Section 2.3, this is in fact a transcription error and the true value was only 116.6ha. However, for the purposes of illustration, let us suppose that the wheat area of 1116.6ha was correct, but that this increase was dictated by an unusual requirement of an environmental scheme that applied to no other farm in the country. Hence it would be incorrect to extrapolate this result to other farms in the `large` stratum, so the holding should instead be given the weight of 1.0 and treated as if it was in its own stratum.

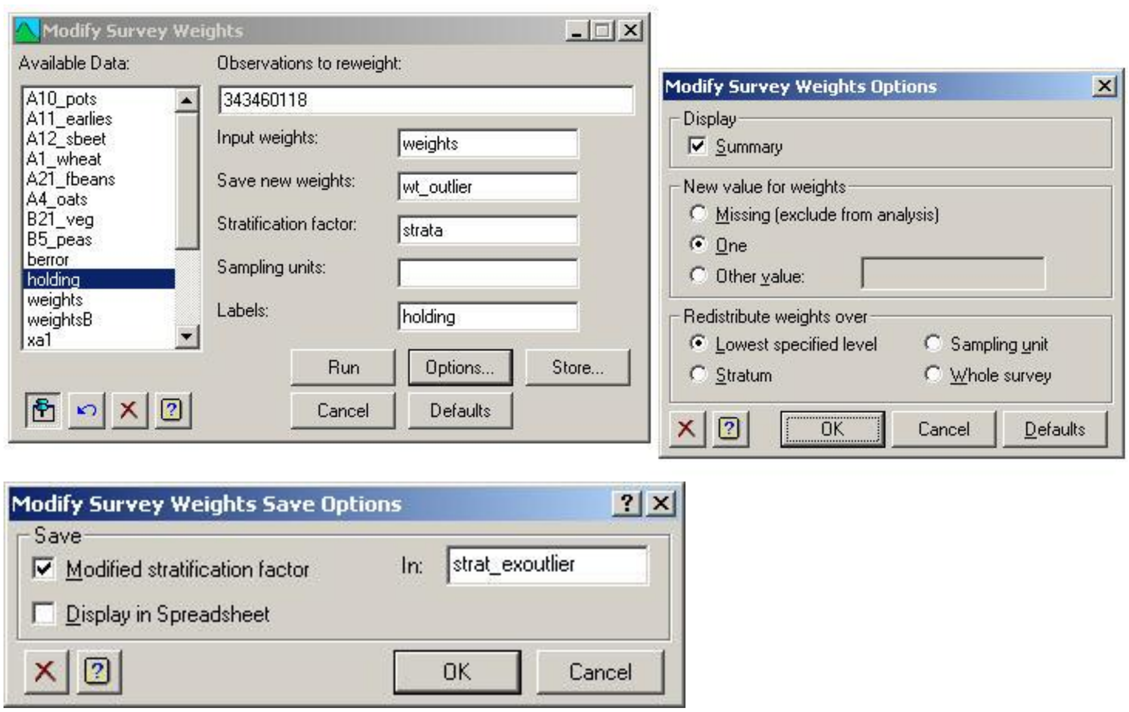


Figure 4.5

Figure 4.5 shows how this may be achieved using the **Modify Survey Weights** menu. With small numbers of outliers, it is generally simplest just to list the observation(s) in the **Observations to reweight** box, making sure that the **Labels** box is set to the appropriate variable (by default, if this is unset, row numbers are used). However, if preferred, the outliers can be identified using a variate of 0's and 1's, as in the previous section. As well as changing the default **New value for weights** from **Missing** to **One**, it is helpful (although not essential) to define a new stratification factor by clicking on the **Store** button, as shown in the lower image of Figure 4.5. The analysis can then be rerun, exactly as in Figure 4.4 except that the **Stratification factor** is set to `strat_exoutlier` and **Weights** are `wt_exoutlier`. The modified analysis is shown below; as expected, the outlier is now in its own stratum with a total weight of 1.0 and a ratio of 9.89 (i.e. 1116.6/112.9).

```
Survey analysis results
=======================

Summary of analysis
-------------------

Y-variate (response data):              A1_wheat
X-variate:                              xa1
Correlation:                            0.935
Method:                                 Design-based (expansion)
Stratification factor:                  strat_exoutlier
Number of strata:                       6
Components for variance calculation:    Between sampling units
Confidence interval method:             tdistribution (95% limits)
Total number of responses:              5508
Survey weights:                         wt_exoutlier
Weights range:                          Min = 1.000  Mean = 3.478  Max = 5.898
Sum of weights:                         19156
Note: statistics above relate to the whole sample, not just the subset defined by the
restriction


Ratios for subpopulation defined by restriction in A1_wheat with 95% confidence limits
--------------------------------------------------------------------------------------

                          n   Sum wts     Ratio      s.e.   %RSE/CV     Lower     Upper
strat_exoutlier
            new           0         0         *         *         *         *         *
          small         992      5851     0.821   0.04604      5.61     0.731     0.911
         medium        1122      5479     0.859   0.02163      2.52     0.817     0.902
          large         945      3073     0.888   0.01436      1.62     0.860     0.916
     very large        1222      2139     0.912   0.00609      0.67     0.900     0.924
       Outliers           1         1     9.890   0.00000      0.00     9.890     9.890
         Margin        4282     16543     0.893   0.00672      0.75     0.880     0.906

Standard errors based on Taylor series approximations. Confidence limits use
t-distribution with 5502 d.f.
```

4.5 Calibration weighting

Calibration weighting is an approach that can be used to modify an initial set of weights, either to remove bias or to ensure that the weights reproduce known population totals. We shall illustrate the approach using the FBS dataset, using data from sheet `crops` of `FBSmult.gwb`; this lists areas of wheat, barley and oilseed rape for each of the FBS farms, whilst sheet `croptotals` gives the estimates of the English national areas of these crops from the much larger June Survey. Using the original weights representing the inverse of the probability of selection, which are in the variate `uncalibrated_wt`, we can estimate total areas and compare these with the June survey areas. There are some substantial differences, particularly for oilseed rape, and so we will use calibration to ensure that the FBS totals match the June ones.

The initial FBS estimate of the oilseed rape area is 584 thousand hectares, compared with a June Survey result of 464 thousand hectares.

```
Survey analysis results
=======================

Summary of analysis
-------------------

Y-variate (response data):              osr
Method:                                 Design-based (expansion)
Stratification factor:                  stratum
Number of strata:                       75
Components for variance calculation:    Between sampling units
Confidence interval method:             tdistribution (95% limits)
Total number of responses:              1776
Survey weights:                         uncalibrated_wt
Weights range:                          Min = 4.597  Mean = 34.72  Max = 146.0
Sum of weights:                         61655


Totals with 95% confidence limits
---------------------------------

                      n    Sum wts     Total      s.e.   %RSE/CV     Lower    Upper
      Alldata
     All data      1776      61655    584285     26494      4.53    532320   636250

Standard errors based on Taylor series approximations. Confidence limits use t-
distribution with 1701 d.f.
```

When such large discrepancies occur, careful checking is needed to ensure that the discrepancy is genuine, and is not the result of an artefact, such as a difference in definition between the two data sources. For the purposes of illustration, let us assume that this difference is genuine, and results from the chance selection of an FBS sample containing too many farms with large areas of rape. It is therefore sensible to use calibration to reduce the weight associated with such farms, so that they are correctly represented in estimates of population totals despite being accidentally over-sampled.

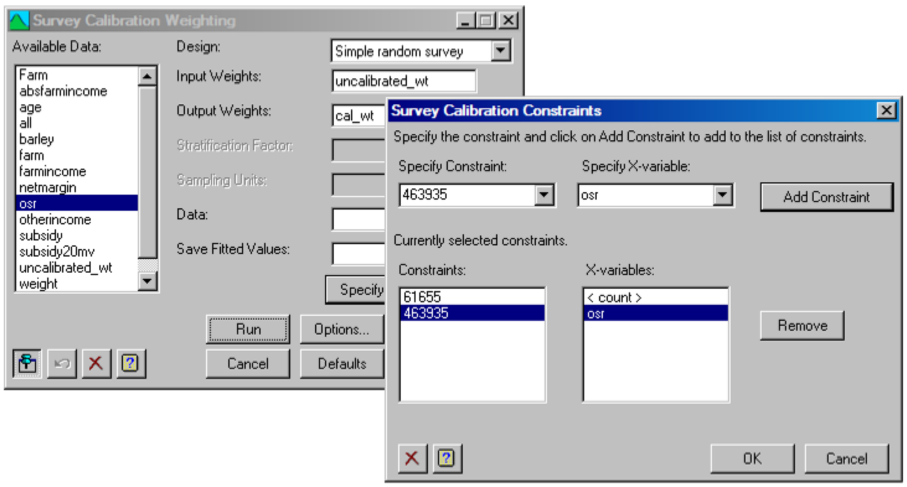


Figure 4.6

Figure 4.6 shows how this is carried out using the **Calibration Weighting** option of the **Survey Analysis** menu. Calibration can be done separately in each stratum of a stratified design, but this depends on having good estimates of the population totals relating to the separate strata. Since sheet `croptotals` just contains a single national figure for all strata, in this instance we will specify a simple random survey as the design, so that a single calibration is used across all strata.

Note that the **Data** box can be left empty; this is used only when it is required to produce estimates of population totals with standard errors allowing for the calibration process. The approach relies on the relationship between calibration and regression analysis of surveys, calculating standard errors using the variance about the regression line, in the same way that ratio analysis calculates standard errors about the ratio line (see Section 2.2). The calibration menu only allows the calculation of population totals, but the **Save Fitted Values** box allows fitted values to be saved and passed to the `SVTABULATE` command (**General Survey Analysis** menu) in order to calculate other statistics (see the practical in Section 4.7). Once a calibration analysis has been run, the fitted values for other variables may be calculated without the need to repeat the calibration by selecting the **Fitted Values** button in the **Method** section of the **Survey Calibration Weighting Options** menu.

Calibration involves specifying one or more constraints, such as the weighted estimate of the rape area equalling 464 thousand hectares; the initial weights are then modified to achieve these constraints whilst minimising the difference between the initial and calibrated weights. The constraints are supplied by clicking on the **Specify Constraints** button, and then supplying them using the top two boxes in the **Survey Calibration Constraints** menu. Thus in Figure 4.6 the national estimate of the rape area, 463935 hectares, has been entered in the first box and `osr` has been specified as the corresponding variable which is multiplied by the new weights to achieve the constraint value. Alternatively, the constraint value may be supplied in a Genstat structure of type scalar or table; suitable structures are listed in the drop down list. When the constraint is correctly specified, clicking the **Add Constraint** button moves it into the list of **Currently selected constraints**.

During calibration it is generally necessary to ensure that the sum of the weights remains constant, since this represents the size of the population. This is achieved by specifying a constraint equal to the sum of the original weights, 61655 in this case. The corresponding x-variable is left unset. When this constraint is added Genstat displays the x-variable as `<count>` (see Figure 4.6) and analyses it as if a vector of 1's had been provided.

When the **Run** button is clicked a summary of the changes to the weights is produced, as shown below. Note that with large datasets the process may take some time, particularly with the iterative methods (truncated linear or logistic), and it may be helpful to tick the **Monitoring** box in the **Options** menu in order to check how the calculations are progressing.

```
Survey calibration
==================

Method:                              linear
Stratification factor:               No_strata
Number of strata:                    1
Total number of data values used:    1776
Input weights:                       Min = 4.597 Median = 28.87 Max = 146.0
Adjusted weights:                    Min = 2.914 Median = 28.52 Max = 149.3
Correlation input & adjusted wts:    0.996

                  Target    Initial    % error      Final    % error
   Constraint
        Count      61655      61655       0.00      61655       0.00
          osr     463935     584285      25.94     463935       0.00
```

The output lists the constraints and the percentage error from the target value for both the initial and calibrated weights; the latter should of course be zero, if the algorithm has reached a satisfactory convergence. Whilst the output gives some basic statistics comparing the old and new weights, including their correlation, it is sensible to examine a graph of the new calibrated weights against the initial ones. This can be obtained by ticking the **Weights plot** box on the **Options** menu and is shown in Figure 4.7. Whilst the adjustments to the weights are generally small, a number of farms with initial weights around about 25 have much smaller calibrated weights. The data information tool (see Section 1.3) can be used to find out more about these points; the initial weight is 23.4 and the bottom point has a calibrated weight of 4.12. These points represent farms with high rape areas, and so reducing their weights pulls the estimate of the total rape area down towards the

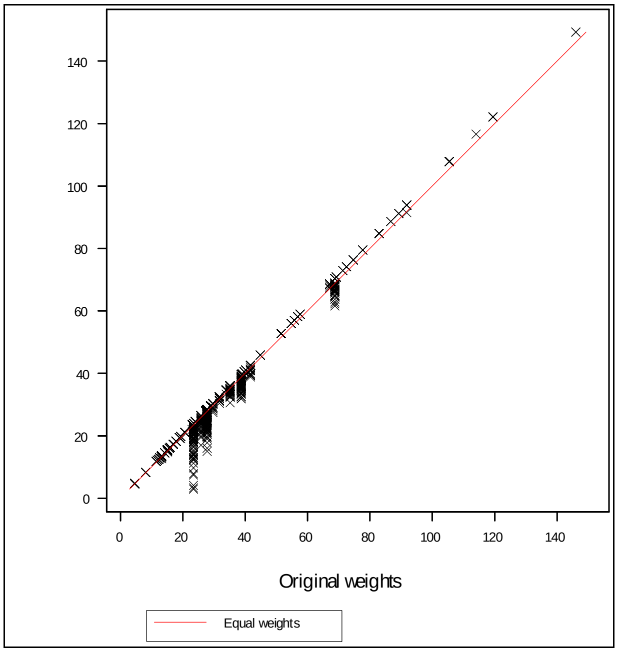


Figure 4.7

constraint value. If these adjustments are considered excessive, it may be preferable to use either the truncated linear or logistic methods, both of which impose lower and upper bounds on the adjustments to the initial weights; the former still uses a linear scale to relate the two sets of weights, whilst the latter uses a logit-like transformation. The bounds are specified as limits on the *g-weights* (that is the multipliers applied to the original weights); by default they are set to 0.1 and 10, so that all calibrated weights must be at least one tenth of the initial weight and not more than ten times as big.

Particularly when working with multiple constraints, it is generally helpful to run a number of calibrations using different methods, different limits and even different combinations of the possible constraints. The various plots of the weights can then be compared in order to decide upon one that achieves the desired aims without excessive adjustment to the weights of particular units. Failure to check the graphs can result in the use of unsatisfactory calibration weights, and hence problems with highly influential observations in the subsequent analyses.

4.6 Calibration by groups

In the above example, a single national estimate for the area of oilseed rape was available. If instead an estimate was available for each `farmsize` category, this information could be supplied as a table, and that is what is done in the example below, using the table in `FBSosrbysize.gsh`. The analysis is run in exactly the same way as is shown in Figure 4.6, except that the constraint is set to the table `osrbysize`, rather than the total 463935.

```
Survey calibration
==================

Method:                                 linear
Stratification factor:                  No_strata
Number of strata:                       1
Total number of data values used:       1776
Input weights:                          Min = 4.597 Median = 28.87 Max = 146.0
Adjusted weights:                       Min = 4.597 Median = 27.83 Max = 146.0
Correlation input & adjusted wts:       0.992

                   Target    Initial    % error      Final    % error
    Constraint
 osr Part-time      41743      54446      30.43      41743       0.00
     osr Small      79512     105800      33.06      79512       0.00
    osr Medium      85002     102521      20.61      85002       0.00
     osr Large      82771      97384      17.65      82771       0.00
osr Very large     174907     224134      28.14     174907       0.00
```

Notice how the table is able to specify five separate constraints, one for each level of farmsize.

4.7 Practical

In Chapter 3 we used the calibration weights in analysing the Farm Business Survey, treating them as if they were ordinary survey weights. When the correlation between the response variable is weak this will be a reasonable, and slightly conservative, assumption. However, when the correlation is stronger it can lead to a serious over-estimation of the variance. To illustrate this, reanalyse the June survey wheat data, (Section 2.2) using the previous wheat area, xa1 as a calibration variable. The file June_calibration.gwb contains the data, with the new holdings strata removed, since it lacks any data for xa1.

First carry out a linear calibration, with A1_wheat as the data variable. Sheet totals contains a table with the totals for each stratum, which should be used for the constraints. Save the fitted values in a variable called whfit. Then analyse A1_wheat using the **General Survey Analysis** menu, using the calibration weights. Compare the standard error from this analysis with an analysis allowing for the impact of calibration by entering whfit in the **Fitted Values** box on the **Save Options** menu.

4.8 Hot-deck imputation for missing values

In the earlier sections of this chapter we saw how weights may be modified to allow for missing values in the data. An alternative solution when data are missing for just some of the survey variables (*item non-response*) is to use imputation to replace the missing value with a plausible non-missing value. This approach involves the need for different sets of weights for different variables and, if used sensibly, may also help to reduce bias when data are not missing at random.

We shall first consider *hot-deck* imputation. The precise definition of this term varies but we shall use it in the most general sense, referring to the class of imputation methods where a missing value in one *receptor* unit is replaced by a value from a *donor* unit. To illustrate the technique we will use column subsidy20mv from FBS_England_merged.gsh; this is a copy of column subsidy but with the first 20 values replaced, for illustrative purposes, with missing values.

The simplest way to impute for these values is simply to take the value from another farm totally at random. To do this select the sub-option **Hot-deck Imputation** from the **Survey Analysis** option on the **Stats** menu. The variable requiring imputation is clicked across to the box at the top left hand corner and a suitable name for the new variable, including the imputed values, is supplied in the right hand box (Figure 4.8). Clicking the **Add to imputation list** button moves the pair to the lower boxes, allowing further pairs to be added, if required.

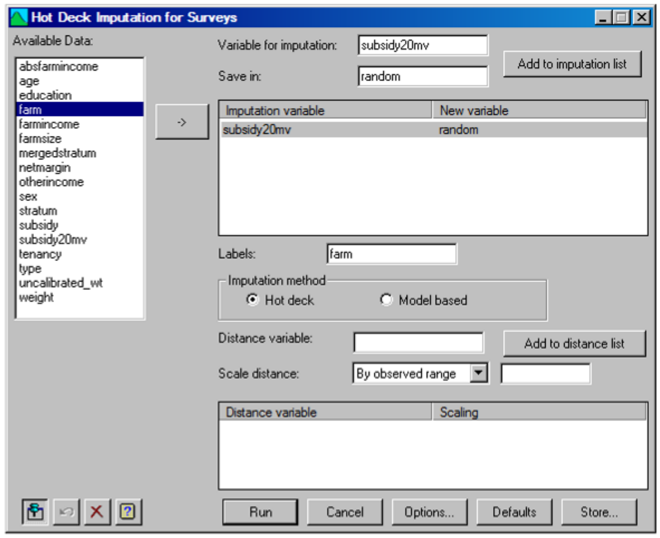
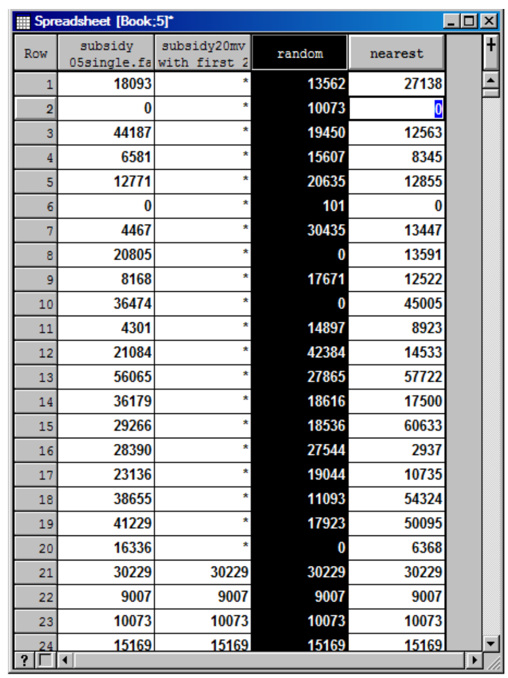


Figure 4.8

The results of the imputation can be seen most easily by putting the complete variable `subsidy`, the version with missing values and the imputed version in a new spreadsheet (Figure 4.9). Note how the new variable `random` has taken the values from `subsidy20mv`, but with the missing values replaced by values from other rows; for example the imputed value in row 2 is taken from row 23.


Spreadsheet [Book;5]*

Row	subsidy 05single.fa	subsidy20mv with first 2	random	nearest
1	18093	*	13562	27138
2	0	*	10073	0
3	44187	*	19450	12563
4	6581	*	15607	8345
5	12771	*	20635	12855
6	0	*	101	0
7	4467	*	30435	13447
8	20805	*	0	13591
9	8168	*	17671	12522
10	36474	*	0	45005
11	4301	*	14897	8923
12	21084	*	42384	14533
13	56065	*	27865	57722
14	36179	*	18616	17500
15	29266	*	18536	60633
16	28390	*	27544	2937
17	23136	*	19044	10735
18	38655	*	11093	54324
19	41229	*	17923	50095
20	16336	*	0	6368
21	30229	30229	30229	30229
22	9007	9007	9007	9007
23	10073	10073	10073	10073
24	15169	15169	15169	15169

Figure 4.9

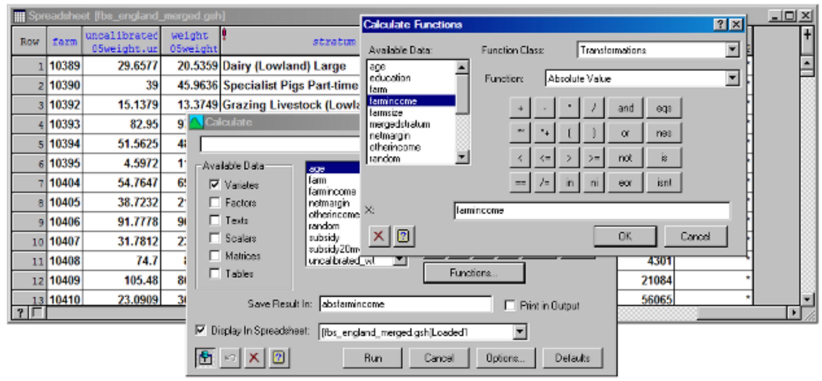


Figure 4.10

Unsurprisingly, although imputation at random avoids any bias, it is not an effective approach, giving large differences between the real values and the imputed ones. The subsidies received differ between different types of farms, and so it is sensible to take account of this in the imputation process. Subsidy also tends to be correlated with the economic size of the farm, and the variable `farmincome` provides a measure of this. There are however some negative values, so a better approach is to calculate a new variable containing the absolute values. This can be achieved by selecting **Column** from the **Calculate** option on the **Spread** menu (Figure 4.10). The imputation can then be rerun, but with variables `type` and `absfarmincome` clicked across to the **Distance variable** box. The output is shown below.

```
Hot-deck imputation
===================

Imputation method: hotdeck
Distance method: minimax
Percent threshold for matches: 0.0%
Threshold for matches: 0.0 relative to minimum
No. of potential donors: 1756
Rows imputed: 20 using 20 donors
Distance range: Min = 0, Median = 0, Max = 0

Histogram of distance
---------------------


            - 0.0003   18 ******************
     0.0003 - 0.0006    1 *
     0.0006 - 0.0009    0
     0.0009 -           1 *

Scale:  1 asterisk represents 1 unit.
```

```
Variables used to calculate distances
-------------------------------------


Variable      Scaling factor
type          *
absfarmincome 3273039


List of donors and recipients
-----------------------------


   Recipient       Donor    Distance
           1         899   0.0000257
           2         716   0.0000009
           3        1398   0.0000128
           4        1536   0.0000510
           5          47   0.0000070
           6         649   0.0003449
           7        1085   0.0000226
           8        1373   0.0000675
           9         358   0.0000098
          10         254   0.0000183
          11        1212   0.0000113
          12        1293   0.0000058
          13         319   0.0011194
          14        1632   0.0000205
          15         525   0.0000419
          16        1250   0.0000425
          17         701   0.0001130
          18        1299   0.0000458
          19         265   0.0000354
          20         863   0.0000180
```

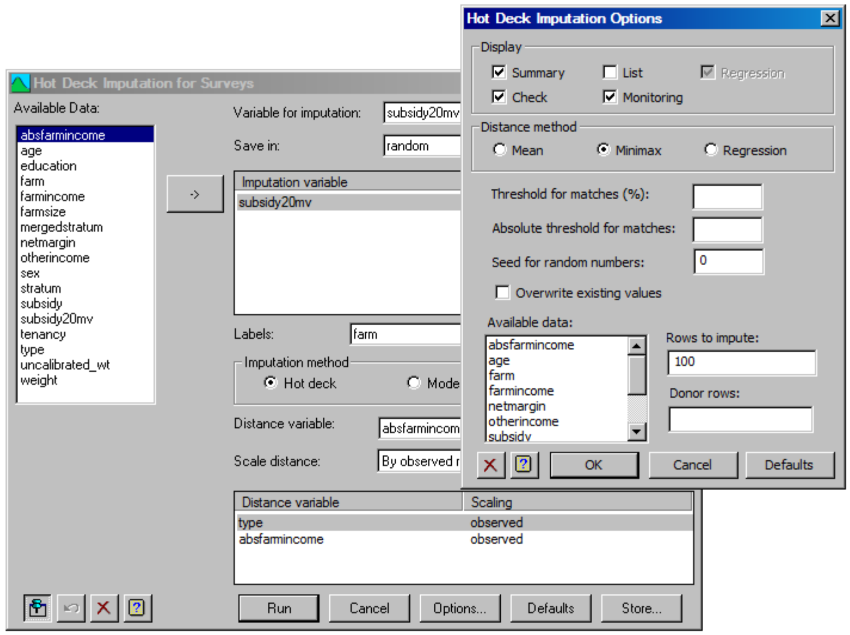


Figure 4.11

To interpret this output, we need to understand how Genstat determines the best match. Let us take row number 1 as an example. For each of the x-variables, a distance is calculated between row 1 (the receptor row) and all the potential donor rows, that is all rows with no missing values (unless otherwise specified). Since `type` is a factor the distance is calculated by an exact matching criterion, with the distance equalling zero if the types match or one if they do not. For variates such as `absfarmincome` the difference between each pair of rows is calculated. By default this is scaled by the observed range of the data; since the minimum value is 23 and the maximum is 3273062 this is 3273039 as shown above. In the case of the selected match between row 1 and row 899, the `absfarmincome` values are 14,699 and 14615 respectively, giving a distance of (14699-14615)/ 3273039 = 0.0000257. Both these rows relate to dairy farms, so the distance with respect to `type` is 0. The default *minimax* method takes the maximum of the differences relating to each potential donor row (i.e. the maximum of 0.0000257 and 0 in the example), and then selects the donor row with the lowest maximum value. The results from using this method are shown in the column `nearest` in Figure 4.9; a quick comparison suggests that it is better than the random allocation, particularly in terms of predicting zero subsidy claims.

In this artificial example, the effectiveness of the imputation process can be judged by comparing the imputed values with the real ones. In real situations a similar comparison can be achieved by setting the options as shown in Figure 4.11. The **Check** box is ticked and the **Rows to impute** is set to 100 to indicate that imputation should be carried out for 100 rows selected at random from the full dataset. The correlation between the real values and the imputed ones can then be used to assess the effectiveness of the procedure.

The resulting graph (Figure 4.12) shows that the imputation based on `type` and `absfarmincome` is reasonable; the correlation between the imputed and actual values is 0.413.

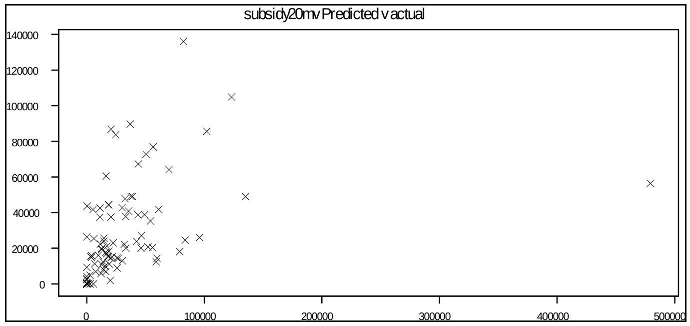


Figure 4.12

In the above imputation, the imputed value for each receptor row was taken from the best matching donor, with random selection used only to decide ties. In other situations it may be preferred to select a match at random from donors within a certain distance of the receptor row. This can be achieved by setting the thresholds in the **Options** menu, either in absolute or percentage terms. One use of this approach is in *multiple imputation*, where the variability between different randomizations of the imputation process is used to gauge the impact on the final results.

4.9 Model-based imputation for missing values

Another method of imputation that is sometimes used is *mean imputation*, where a missing value is replaced by the mean of the appropriate category. Thus in the FBS example of the previous section, we could replace missing subsidy values for dairy farms by the mean level of subsidy for those dairy farms with valid data. A natural extension of this is to use other linear regression models to predict missing values. For example we might use a regression with `subsidy` as the dependent variable and `absfarmincome` as the independent (predictor) variable. Missing values in `subsidy` could then be produced by predicting the value that would be expected for the appropriate `absfarmincome` value. No special facilities exist for doing mean imputation in Genstat, but it can easily be achieved by fitting the regression model to the full dataset (including the missing values) and saving the fitted values (**Linear Models** sub-option of the **Regression Analysis** option on the **Stats** menu).

There is, however, a disadvantage with mean imputation. Although it leads to good estimates of means and totals, it causes a downward bias in estimates of variances because the imputed values are homogeneous, without the random variation about the mean found in the real data. This leads to standard errors and confidence limits that give a misleading picture of the real precision of the estimates. To avoid this it may be helpful to add random variation to the fitted values, thus ensuring that they mimic the real data in terms of variability. The hot-deck imputation menu can be used to achieve this, adding a residual from a donor unit to the fitted value from the receptor (missing) unit to form the model based imputations. This is sometimes referred to as a *semi-parametric* imputation method, since it is midway between the non-parametric approach of the previous section and the fully parametric approach in which artificial residual values are selected from a Normal distribution of appropriate variance.

To illustrate the method, we will model `subsidy20mv` by fitting separate linear slopes against `absfarmincome` for each farm type. This can be done by **Linear Models** sub-option of the **Regression Analysis** option on the **Stats** menu as shown in Figure 4.13. It should be noted that examination of the residuals (e.g. by clicking **Further Output** and then **Model Checking**) provides strong evidence of non-Normality and so significance tests will not be valid. Nevertheless, the model can be used for imputation, provided that residuals are randomized within relatively homogeneous groups. The alternative is a model based on log-transformed subsidy; this would be more appropriate for most purposes, but may produce some implausibly large imputed values when back-transformed if the residuals show any departure from a homogeneous Normal distribution. To check the fitted model it is useful to produce a graph of the fitted relationship; this can be achieved by clicking **Further Output** and then **Fitted Model**.

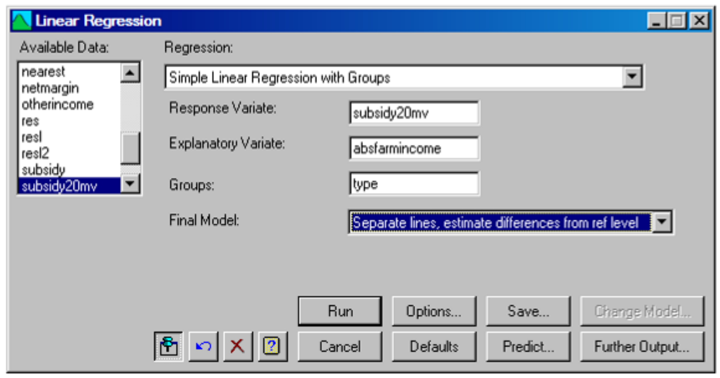


Figure 4.13

The resulting graph, after some editing to make the range of the axes more appropriate, is shown in Figure 4.14 (note that a few very large points lie beyond the maxima of the axes). The three lines with very shallow slopes correspond to pig, poultry and horticultural farms, which have received little subsidy in the past.

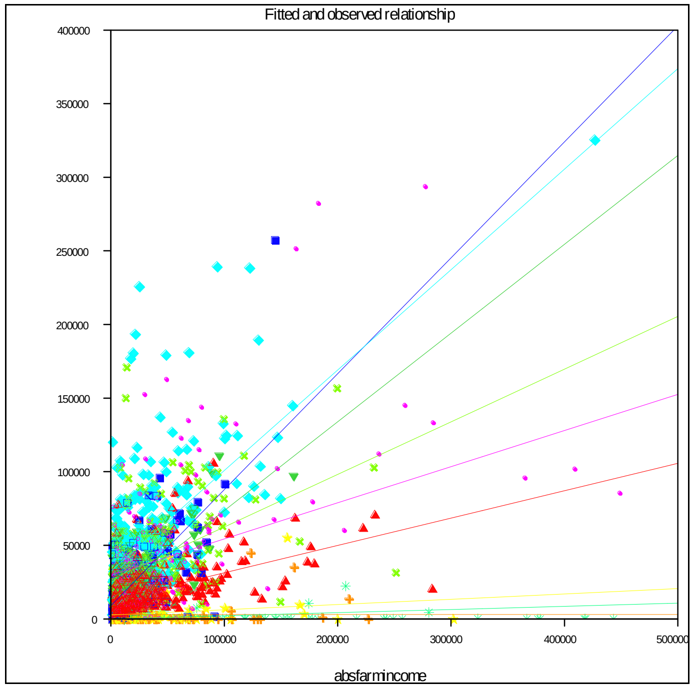


Figure 4.14

To form each imputed value we need to read off the expected level of subsidy for the appropriate level of `absfarmincome`, using the line for the correct farm `type`. The vertical distance (residual) from another, real, data point is then added to this fitted value to form the imputed value. Figure 4.15 shows how this is done using the **Hot-deck Imputation** menu. Specifying `type` as the distance variable ensures

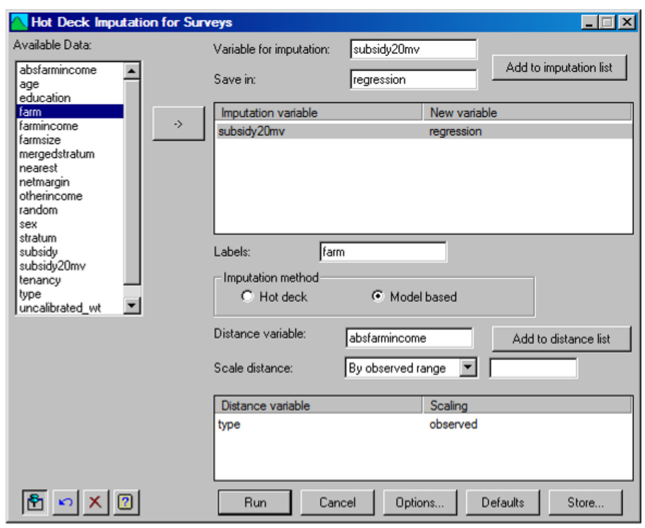


Figure 4.15

that residuals are randomized within each type (i.e. a dairy farm receives a residual only from another dairy farm). This approach was chosen because the residual variance varies substantially between farm types; there would also be a case for using `absfarmincome` in addition, but that has not been done here in order to emphasise that the distance variables used in the distance matching need not be the same as those in the fitted model.

There are other ways that regression can be used in the calculation of fitted values. One approach is to use a hot-deck approach, but with donors selected from units with similar fitted values. To do this, first fit the model as described above and then save the fitted values with a suitable name by clicking on the **Save** button on the **Linear Regression** menu. The imputation step is then exactly as described in Section 4.7 above, but with the fitted values specified as the distance variable. A variant on this is to use the estimated slopes from the regression as weights for the calculation of distances; for example, if the slope of `x1` is 0.24 and two units have `x1` values of 10 and 20, the distance is (20-10)×0.24=2.4. In the case of factors, the predicted value for each level is used as the basis for the distance calculation; thus if group 1 has a predicted value of 150 and group 2 has a predicted value of 175, the distance between a receptor unit of group 1 and a donor of group 2 is 25. The maximum distance from the different variables is then taken for each pair of units as in the minimax method. This variant can be selected using the **Regression** option for **Distance method** in the **Hot-deck Imputation Options** menu.

5 Programming Genstat for surveys

So far all the analyses in this Guide have been completed using the menu system. This is an excellent way of learning Genstat and of exploring new datasets, but to make full use of Genstat it is helpful to master the program's in-built programming language. Using programming has two big advantages for survey work: automating repetitive tasks, and maintaining a simple audit trail of the process. In this chapter you will learn about

- saving and modifying the commands generated by the menu system
- finding more information about commands
- writing simple programs to analyse a list of questions
- defining sub-populations using restrictions

5.1 Modifying menu commands

Writing a completely new program can be a daunting task and so it is generally easier to modify existing Genstat commands, maybe from a similar survey conducted in the past. However, when learning about new commands an alternative source of code to modify is provided by the Genstat menu system. Whenever the **Run** button is pressed on a Genstat menu, the commands generated to perform the analysis are copied to the **Input** window. To illustrate this we will use the data on farm incomes from `FBS_England_Merged.gsh` which we first examined in Chapter 3.

Start by opening the file `FBS_England_Merged.gsh` and then select **General Survey Analysis** from the **Survey Analysis**

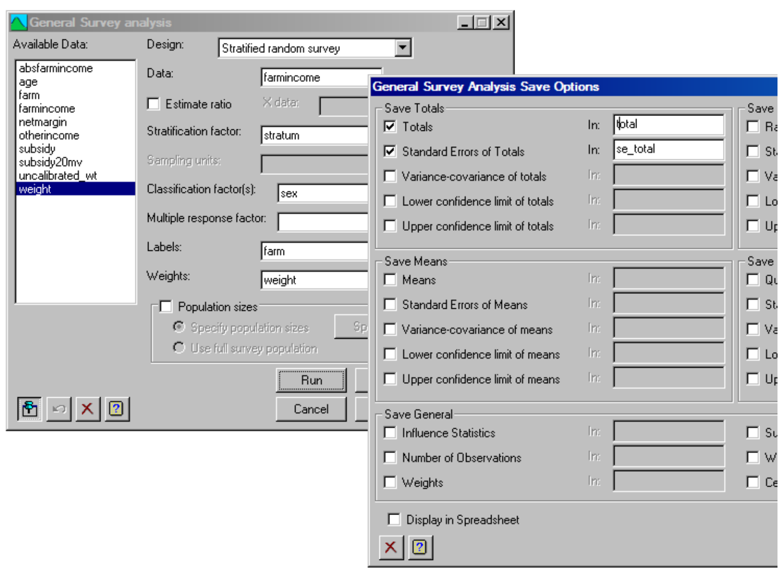


Figure 5.1

submenu on the **Stats** menu. Set the menu as shown in Figure 5.1. There is no need to alter the options menu at this stage. Now click the **Run** button and select the **Input Log**, either by clicking on it in the windows list at the left of the screen, or by selecting it from the **Window** menu.

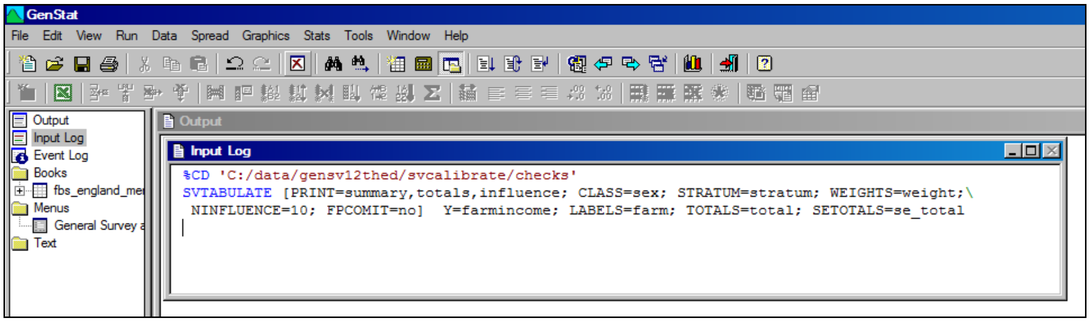


Figure 5.2

You should then see the command shown in Figure 5.2 (if necessary scroll down to the bottom of the window). Looking at the `SVTABULATE` command in more detail, it essentially consists of two parts;

1. Within the square brackets, there is a list of `options`, in this case `PRINT`, `CLASS(IFICATION)`, `STRATUM`, `WEIGHTS`, `NINFLUENCE` and `FPCOMIT`. The continuation symbol `\` is used to split the command over two lines due to its length.
2. After the square brackets there is a list of `parameters`, `Y`, `LABELS`, `TOTALS` and `SETOTALS`.

In the commands generated by the menus, the names for options, parameters and the command itself are shown in capital letters and the settings are in lower case. This is a useful convention, but either lower or upper case can be used. However, variable names must be in the correct case. Names of commands, options and parameters can all be abbreviated (to not less than four characters for commands), but we will generally show them in full in this Guide.

More detail about the syntax of commands in general can be found in the *Guide to the Genstat Command Language*, but for more information on `SVTABULATE` itself, search for it in the help facilities as shown in Figure 5.3. All possible options and parameters are shown, together with a brief description of what they do, and a list of possible settings where appropriate. If you scroll down further, you will see a more detailed description of the use of the procedure.

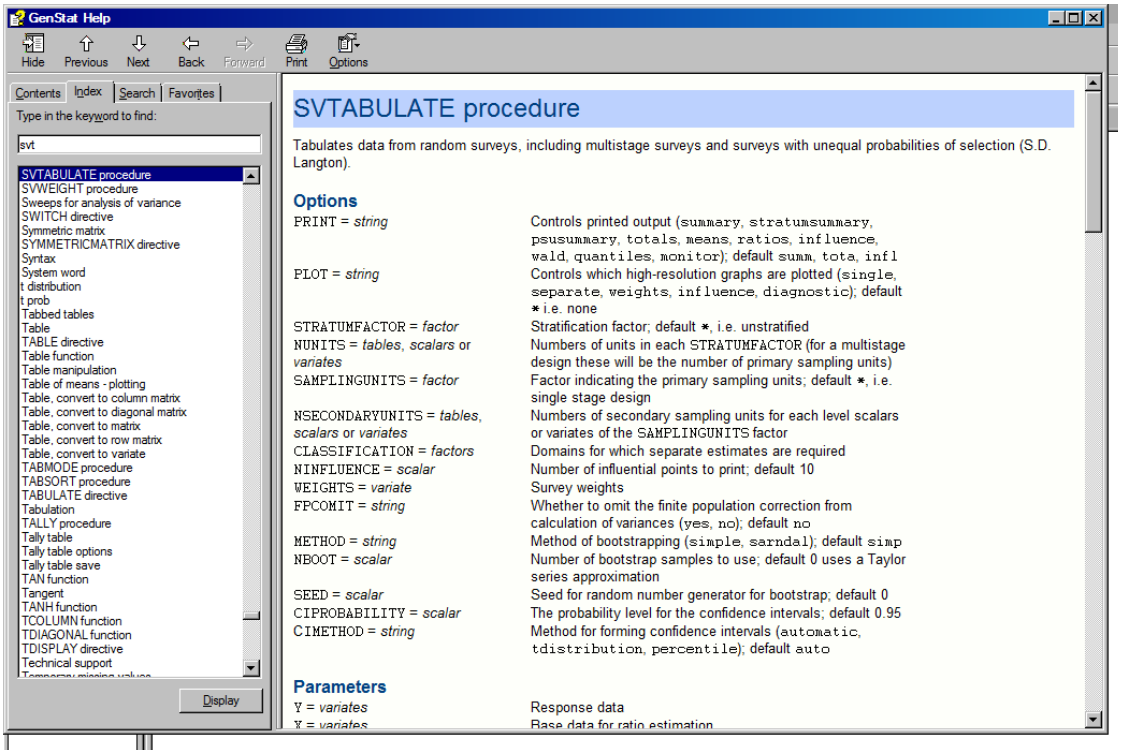


Figure 5.3

In order to make changes to the command it is necessary to copy it to a new text window, which may be created by either clicking on the button on the left of the toolbar, or by selecting **New** from the **File** menu and choosing **Text Window** from the **General** tab. You can then edit it as required. In Figure 5.4 a new variable called

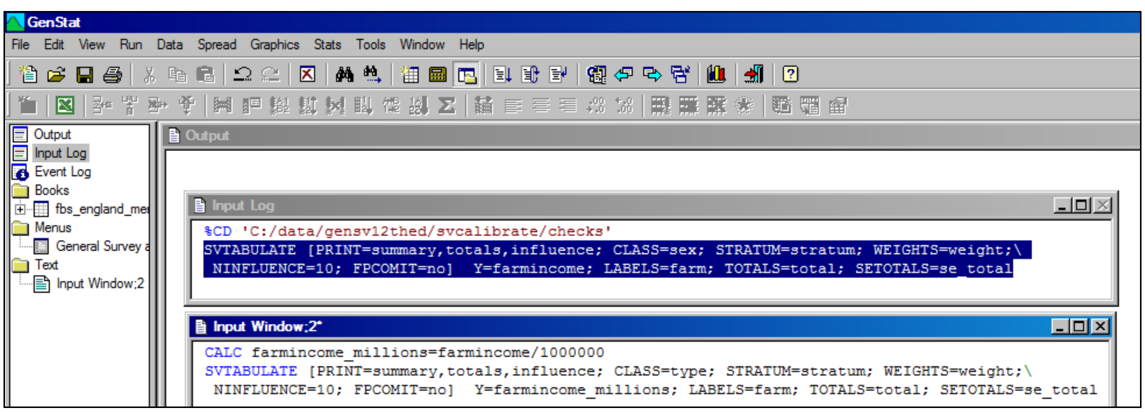


Figure 5.4

farmincome_millions has been created; this makes the output easier to read by avoiding the excessive numbers of digits in the national total. The Y parameter of SVTABULATE has been changed to this new variable and the CLASSIFICATION factor has been set to type. Once all the changes have been made, the modified command can be highlighted and results produced by selecting **Submit Selection** from the **Run** menu, or alternatively by using the button on the toolbar with a downward arrow alongside a sheet of paper.

5.2 Practical

Modify the command so that it also prints the stratum summaries and Wald test statistics. Save the test statistics in a structure called test_stats.

5.3 Analysing lists of variables

In most surveys there are many variables to analyse and programming provides a way of automating this repetitive task. When doing this, however, it is important to examine the output of each variable separately, as there may be issues, such as the treatment of outliers or the appropriate sub-population to analyse, which will vary.

The simplest way to analyse several variables is simply to list them at the Y parameter. For example, to analyse farmincome, otherincome, subsidy and netmargin, and to save the means per farm, we could type:

```
SVTABULATE [PRINT=summary,means,influence; CLASS=type; STRATUM=stratum; WEIGHTS=weight;\
  NINFLUENCE=10; FPCOMIT=no]  Y=farmincome,otherincome,subsidy,netmargin; LABELS=farm;\
  MEANS=meanfi,meanoi,meansub,meannm; SEMEANS=sefi,seoi,sesub,senm
```

This is a good moment to explain the difference between the *options* within the square brackets and the *parameters* that follow them. There are four Y variables and the parameters MEANS and SEMEANS also have four settings corresponding to them, so that the means for farmincome are stored in meanfi, etc. When the same setting is appropriate for each Y variable, as is the case for LABELS, it is sufficient to write LABELS=farm, since the values are recycled so it is treated as if it said LABELS=farm,farm,farm,farm. By contrast, options apply to all Y variables. Thus, the three settings of the PRINT option, apply to all the Y variables and so the summary, means and influence statistics are printed for each one.

This listing approach works well with small numbers of variables, but is more problematic when a survey contains very large numbers of questions. The commands then become very long, with an increasing risk of failure due to typing errors. In particular, if an item is missed off the list for a parameter like `MEANS`, the wrong means can end up in the wrong structure, which may be difficult to spot. This problem can be avoided by the use of `FOR` loops and *pointers*.

`FOR` loops are best illustrated by a simple example. Suppose we just want to print the analyses for the variables `farmincome`, `otherincome`, `subsidy` and `netmargin` without saving the results. Using an *implicit loop*, as described above, we would write:

```
SVTABULATE [PRINT=summary,means,influence; CLASS=type; STRATUM=stratum; WEIGHTS=weight;\
  NINFLUENCE=10; FPCOMIT=no]  Y=farmincome,otherincome,subsidy,netmargin; LABELS=farm
```

Exactly the same output could be achieved using a `FOR` loop as follows:

```
FOR d= farmincome,otherincome,subsidy,netmargin
  SVTABULATE [PRINT=summary,means,influence; CLASS=type; STRATUM=stratum; WEIGHTS=weight;\
    NINFLUENCE=10; FPCOMIT=no]  Y=d; LABELS=farm
ENDFOR
```

The structure `d` is known as a *dummy*. The code between the `FOR` and `ENDFOR` commands is executed four times, with the dummy representing a different variable each time. Thus the first time `d` represents `farmincome`, the second time `otherincome`, etc. More than one dummy can be set, as in the following example which saves the tables of means in suitably named structures using a dummy called `mtab`.

```
FOR d= farmincome,otherincome,subsidy,netmargin; mtab= meanfi,meanoi,meansub,meannm
  SVTABULATE [PRINT=summary,means,influence; CLASS=type; STRATUM=stratum; WEIGHTS=weight;\
    NINFLUENCE=10; FPCOMIT=no]  Y=d; LABELS=farm; MEANS=mtab
ENDFOR
```

5.4 Practical

Modify the `FOR` loop above so that it produces tables of `farmincome` cross-tabulated by a) `sex` of farmer, b) `type` of farm, and c) `tenancy` type. Note that this example cannot be achieved using an implicit loop because `CLASSIFICATION` is an option, not a parameter.

5.5 Pointers

In itself, the use of a `FOR` loop does not give much advantage over the implicit loop approach of simply listing the variables to use as `Y` parameters. However, their usefulness can be increased by the use of pointers. Pointers are lists of variables. For example, the following command defines a pointer containing the four variables analysed above:

```
POINTER [VALUES= farmincome,otherincome,subsidy,netmargin] ydata
```

Suffixes can be used to refer to individual elements of this list, as for example, `ydata[1]`, whilst two or more can be listed as `ydata[1,3]`. Most importantly, the whole list can be referred to by using empty brackets, `ydata[]`. Try the following commands which each produce descriptive statistics for one or more of the variables, as indicated by the comments in quotation marks:

```
POINTER [VALUES= farmincome,otherincome,subsidy,netmargin] ydata
DESCRIBE ydata[2]       "stats for otherincome"
DESCRIBE ydata[2,3]     "stats for otherincome and subsidy"
DESCRIBE ydata[1...3]   "stats for farmincome, otherincome and subsidy"
DESCRIBE ydata[]        "stats for all four variables"
```

Note how three dots (`...`) is used to continue a series of numbers.

Pointers can be used most easily in `FOR` loops by using the `NTIMES` option, which specifies the number of times the loop is to be executed, and the `INDEX` option, which defines a *scalar* (single valued structure) taking the value 1 the first time, 2 the second, etc. Since our pointer contains four structures, we can write:

```
POINTER [VALUES= farmincome,otherincome,subsidy,netmargin] ydata

FOR [NTIMES=4;INDEX=i]
  SVTABULATE [PRINT=summary,means,influence; CLASS=sex; STRATUM=stratum; WEIGHTS=weight;\
    NINFLUENCE=10; FPCOMIT=no]  Y=ydata[i]; LABELS=farm; MEANS=mean[i]; SEMEANS=sem[i]
ENDFOR
FSPREAD mean[],sem[]
```

This time, we have also used pointers to save both the means and their standard errors. These pointers are not defined in advance, so the variables do not have names (e.g. `meanfi` etc.), but we can still refer to them using the pointer-suffix notation. The final statement uses the `FSPREADSHEET` (form spreadsheet) command to display a spreadsheet containing the means and standard errors.

Finally, the commands below demonstrate a couple of refinements of these commands. Instead of manually telling the program to execute the loop four times, we have calculated a scalar `nvy` containing the number of structures in the pointer and set the `NTIMES` option to equal this. As a result, if we alter the variables in the pointer, no further changes are needed elsewhere in the program, because it automatically determines the number of times to execute the commands within the `FOR` loop.

```
POINTER [VALUES= farmincome,otherincome,subsidy,netmargin] ydata
CALC nvy=NVALUES(ydata)

SCALAR i;VALUE=1
FOR [NTIMES=nvy;INDEX=i]
  SVTABULATE [PRINT=summary,means,influence; CLASS=sex; STRATUM=stratum; WEIGHTS=weight;\
    NINFLUENCE=10; FPCOMIT=no]  Y=ydata[i]; LABELS=farm; MEANS=mean[i]; SEMEANS=sem[i]
ENDFOR
FSPREAD mean[],sem[]
```

The other modification is to create the scalar i before the loop and give it the initial value 1. This has no impact on the results when running the whole block of commands but it does allow the commands to be tested before running them on all the variables. When running commands in a loop, a minor typing mistake can sometimes results in large numbers of warning messages and a large volume of text in the output window. This can be confusing, so it is easier to test the loop first using just the first variable, and then go on to run it properly only after any problems have been rectified. To do this, first run the commands up to, but not including, the FOR command (see the output window in Figure 5.5). Then highlight the commands within the FOR loop, as shown in input window 1 of Figure 5.5 and run them using **Submit Selection** from the **Run** menu. Examine the output, checking it has done what you wanted it to do before running the whole section of code.

5.6 When things go wrong

Programming in any computer language is not easy. For example, a simple typing mistake can cause unexpected errors later on in a program. Even the best programmers make errors, and so understanding them and learning how to correct them is an important skill. Because there are so many types of errors it is difficult to cover all possibilities, but the list below provides some pointers that may help.

1. One error or warning message in a program often triggers further ones later on even though the later commands may be completely correct, so try to find the original problem. In particular, in the output window do not focus on the warning message at the bottom of the window, without scrolling up to check for earlier messages. The output button on the fault message

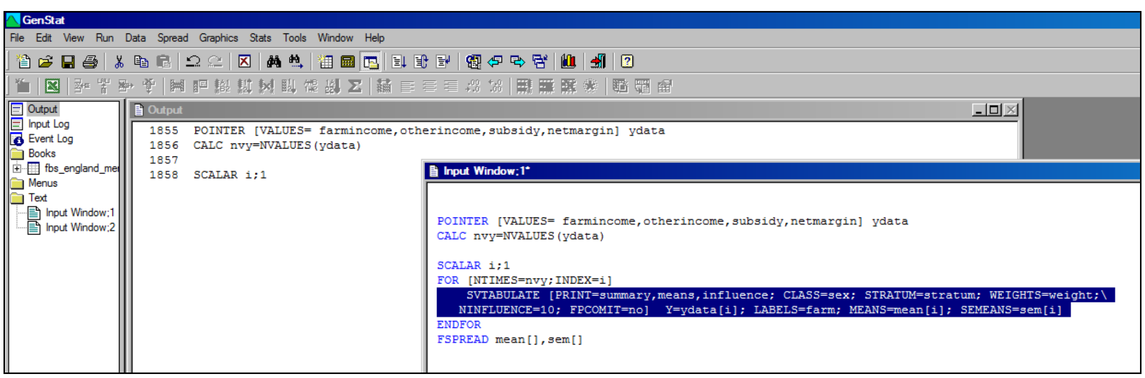


Figure 5.5

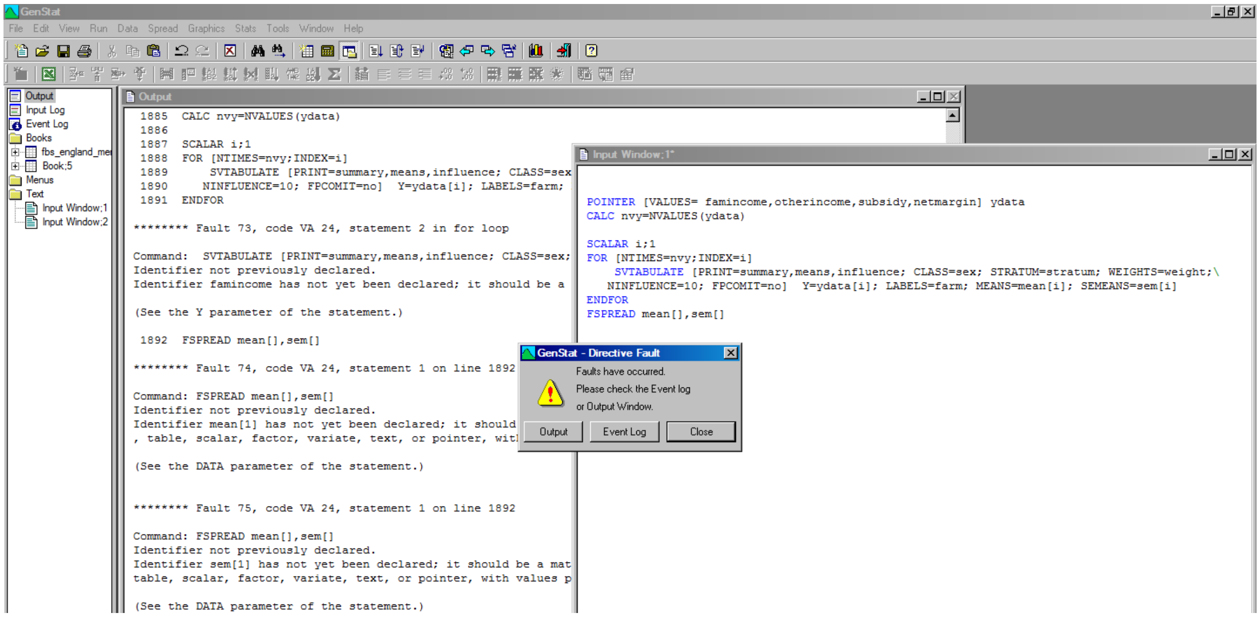


Figure 5.6

dialogue box will generally take you to the earliest message. Figure 5.6 provides an example; clicking output will highlight the first fault which includes the message `Identifier famincome has not yet been declared`. In this case the mistake was in the pointer statement where `farmincome` is mis-spelt as `famincome`, with the result that `SVTABULATE` cannot analyse this non-existent variable.

2. As the above example shows, many problems relate to variables that cannot be found, perhaps because the identifier has been wrongly typed, or because the command creating them has not worked properly. When faced with a message like this, check that the variable exists. This can be done using the **Data** tab in the left-hand pane of Genstat. Look carefully at the spelling and remember that variable names are case sensitive. Alternatively, the `DUMP` command provides information on particular variables, whilst `LIST` produces a list of all structures of a particular type; both can be run by typing them in an input window:

```
1862  dump famincome,farmincome
Dump
====

Identifier      Type  Length   Values Missing  Ref.No.

 famincome  *              *   Absent       *     -610
farmincome  Variate     1776  Present       0     -749
 1863  list variate

  Structures of type VARIATE

         identifier  number of values
               farm              1776
    uncalibrated_wt              1776
             weight              1776
                age              1776
          netmargin              1776
         farmincome              1776
        otherincome              1776
            subsidy              1776
        subsidy20mv              1776
      absfarmincome              1776
```

3. When one fault occurs, this can often lead to subsequent problems, and so it is often sensible to clear all data and start again in order to remove the risk of unexpected errors. Selecting **Clear All Data** from the **Data** menu will achieve this, although an alternative is **Restart Server** from the **Run** menu; the latter also closes all open files and is therefore better when external files are being used.

5.7 Reading from and writing to data files

So far, we have opened the spreadsheet `FBS_England_Merged.gsh` manually, but this process can also be automated using the `SPLOAD` command:

```
SPLOAD 'FBS_England_Merged.gsh';ISAVE=ipo
```

Notice the `ISAVE` parameter; this creates a pointer listing all the columns in the spreadsheet, and is particularly useful when some rows need to be excluded in the subsequent code, for example to produce estimates for a subpopulation. `SPLOAD` works only with Genstat spreadsheets, but the IMPORT command can import data from a wide variety of filetypes, including Excel spreadsheets and Genstat workbooks. The DBIMPORT command can read data from Access and other databases.

As well as reading from a variety of file types, Genstat can produce results files in various formats. In the earlier examples, we used FSPREADSHEET to create spreadsheets within Genstat, and these can be saved in a variety of formats by selecting **Save As** from the **File** menu. Alternatively, the OUTFILE option of FSPREADSHEET allows Genstat spreadsheets to be created directly, whilst EXPORT can create files in a variety of formats, including Excel files and Genstat workbooks. The example below reads the data using SPLOAD and sends the results to an Excel file, without any need to use the Genstat menus.

```
SPLOAD 'FBS_England_Merged.gsh';ISAVE=ipo

POINTER [VALUES= farmincome,otherincome,subsidy,netmargin] ydata
CALC nvy=NVALUES(ydata)

SCALAR i;VALUE=1
FOR [NTIMES=nvy;INDEX=i]
  SVTABULATE [PRINT=summary,means,influence; CLASS=sex; STRATUM=stratum; WEIGHTS=weight;\
    NINFLUENCE=10; FPCOMIT=no]  Y=ydata[i]; LABELS=farm; MEANS=mean[i]; SEMEANS=sem[i]
ENDFOR

EXPORT [OUTFILE='FBS_Results.xls'; METHOD=add; SHEET='Tables by sex'] mean[],sem[]
```

Data files can also be used to store a list of variables to be analysed. This approach can be particularly useful when there are very large number of variables and defining pointers in code may become cumbersome. It also allows staff not familiar with Genstat to set up the analysis using a spreadsheet package, without the need to understand the Genstat program.

This is illustrated below. The Excel file FBS_England_Merged.xls contains a list of variables to tabulate by sex in sheet tables by sex. Using IMPORT these lists are created as text structures in Genstat but the FPOINTER command[6] allows them to be converted to pointers. This is illustrated below:

[6] FPOINTER is not a standard feature of Genstat but is part of the Biometris library, which may be installed from http://www.vsni.co.uk/software/genstat/user-area/

```
SPLOAD 'FBS_England_Merged.gsh';ISAVE=ipo

IMPORT 'FBS_England_Merged.xls';sheet='by_sex'

FPOINTER TEXT=tdata; POINTER=ydata

"set up pointers for tables of means and standard errors"
TXCONSTRUCT [TEXT=tmean] 'mean_',tdata
FPOINTER TEXT=tmean; POINTER=mean
TXCONSTRUCT [TEXT=tsem] 'se_',tdata
FPOINTER TEXT=tsem; POINTER=sem

CALC nvy=NVALUES(ydata)

SCALAR i;VALUE=1
FOR [NTIMES=nvy;INDEX=i]
  SVTABULATE [PRINT=summary,means,influence; CLASS=sex; STRATUM=stratum; WEIGHTS=weight;\
    NINFLUENCE=10; FPCOMIT=no]  Y=ydata[i]; LABELS=farm; MEANS=mean[i]; SEMEANS=sem[i]
ENDFOR

EXPORT [OUTFILE='FBS_Results.xls'; METHOD=add; SHEET='Tables by sex'] mean[],sem[]
```

Notice that we have also used `FPOINTER` to create the pointers `mean` and `sem` explicitly. This ensures that the columns in the Excel file have informative names (e.g. `mean_farmincome` rather than `mean[1]`). The `TXCONSTRUCT` command creates these names by joining text structures together.

`TXCONSTRUCT` can also change the case of text structures and join texts to strings formed from numerical structures. This is illustrated in the example below. `TXCONSTRUCT` is used to put the list of variables into upper case, and this new text is then used to form the pointer mean. Thus the table of means formed from `farmincome` is called `FARMINCOME`.

The other complication in this example is that the sheet `crosstabs` specifies different tabulation factors for different variables. As a result a separate spreadsheet needs to be created for each loop; all the tables cannot be put into the same spreadsheet because the `CLASSIFICATION` factors vary. Names have been created for these sheets by using `TXCONSTRUCT` to combine the loop number with the variable name, producing names such as `Table 3 subsidy`. Note how the `$` symbol allows the use of individual rows of the structure `tvariate`; for example, if the scalar `i` has the value 3, then `tvariate$[i]` gives the value in the third row of the structure.

```
SPLOAD 'FBS_England_Merged.gsh';ISAVE=ipo

IMPORT 'FBS_England_Merged.xls';sheet='crosstabs'

FPOINTER TEXT=tvariate,tfactor;POINTER=pvariate,pfactor
TXCONSTRUCT [TEXT=tmean;CASE=upper] tvariate
FPOINTER tmean;mean

CALC nvy=NVALUES(pvariate)

SCALAR i;1
FOR [NTIMES=nvy;INDEX=i]
  SVTABULATE [PRINT=summary,means,influence; CLASS=pfactor[i]; STRATUM=stratum;\
    WEIGHTS=weight; NINFLUENCE=10; FPCOMIT=no]  Y=pvariate[i]; LABELS=farm;\
    MEANS=mean[i]; SEMEANS=sem[i]
  TXCONSTRUCT [TEXT=tsheet;SEPARATOR=' '] 'Table',i,tvariate$[i];DECIMALS=0
  EXPORT [OUTFILE='FBS_England_Crosstabs.xls';METHOD=add;SHEETNAME=#tsheet] mean[i],sem[i]
ENDFOR
```

5.8 Restrictions and subsets

In the earlier chapters we have seen the importance of restrictions. These were used in Section 2.4 to identify outliers, and in Section 3.3 to define sub-populations with `SVTABULATE`. In this section we shall see how to define these with commands, using the example of Section 3.3, in which we looked at income of male farmers tabulated by their educational background.

The `RESTRICT` command is very simple; it has no options and only three parameters, of which only the first two are need here. The first parameter, `VECTORS`, lists the structures to be restricted (*vectors* is a collective name for one-dimensional structures such as variates, texts and factors). Unlike the restrictions generated by the **Restrict/Filter** item on the **Spread** menu, where any restriction applies to all variables in a spreadsheet, restrictions defined in the command language can apply to any group of variables. In this case we could just restrict `farmincome`, but it is equally easy to restrict all the variables, by using the pointer formed by the `ISAVE` parameter of `SPLOAD`. The output shows this, and is identical to that of Section 3.3:

```
SPLOAD [PRINT=*] 'FBS_England_Merged.gsh';ISAVE=alldata

RESTRICT alldata[];CONDITION=sex.EQ.1

```

```
  34  SVTABULATE [PRINT=means; CLASS=education; STRATUM=stratum; WEIGHTS=weight] farmincome

Means for subpopulation defined restriction in farmincome with 95% confidence limits
------------------------------------------------------------------------------------

                      n   Sum wts      Mean      s.e.   %RSE/CV     Lower     Upper
 05farmer.education
  school only       526     19874     13807      1510     10.93     10846     16768
         GCSE       230      8536     30082     11729     38.99      7078     53087
     A levels       121      4123     20041      3081     15.37     13997     26084
      college       511     16356     20886      1680      8.04     17590     24181
       degree       222      6789     38041      5063     13.31     28110     47972
     postgrad        41      1645      9757      4682     47.98       574     18940
   apprentice        36      1323     15941      3389     21.26      9294     22587
        other        36      1094     25402      8467     33.33      8796     42008
         Mean      1723     59740     21403      1884      8.80     17708     25098

Standard errors based on Taylor series approximations. Confidence limits use t-distribution
with 1701 d.f.

  35
  36  RESTRICT alldata[]
```

Let us now look at how the restriction is defined using the `CONDITION` parameter. `CONDITION` should be set to a logical expression that takes the value 1 for the rows to be included in the analysis and 0 for those to be excluded. The `CONDITION` may be formed by calculating a suitable variate, or by reading it from a file, but, most commonly, it is specified using Genstat's *relational operators*. In this case the relational operator `.EQ.` is used to test whether the value of `sex` in each row is equal to 1, which is the value used for male. The most common simple relational operators are the following:

equality	`.EQ.`	or	`==`
non-equality	`.NE.`	or	`<>`
less than	`.LT.`	or	`<`
less than or equals	`.LE.`	or	`<=`
greater than	`.GT.`	or	`>`
greater than or equals	`.GE.`	or	`>=`

In this case, since `sex` is coded 1 for `male`, 2 for `female`, there are a variety of ways that we could have specified the restriction. Any of the following would have achieved the same restriction:

```
RESTRICT alldata[]; CONDITION=sex.LE.1
RESTRICT alldata[]; CONDITION=sex.LT.2
RESTRICT alldata[]; CONDITION=sex.NE.2
```

Restrictions can be combined by using the operators `.AND.` and `.OR.`, so we could restrict to male farmers with degrees (coded as 4) by putting:

```
RESTRICT alldata[]; CONDITION=sex.EQ.1.AND.education.EQ.4
```

Brackets can be used to avoid ambiguity. The first expression below gives male farmers in the degree or postgrad groups, whereas the second gives male farmers with degrees or farmers of either sex with postgraduate qualifications:

```
RESTRICT alldata[];\
  CONDITION=sex.EQ.1.AND.(education.EQ.4.OR.education.EQ.5)
RESTRICT alldata[];\
  CONDITION=(sex.EQ.1.AND.education.EQ.4).OR.education.EQ.5
```

Whilst these operators are very simple and straightforward, the use of numerical levels for a factor with labels can cause confusion. It is not, for example, immediately apparent that `degree` is level 4 of `education`, because the levels are numbered from 0, not from 1. The following two operators allow either numerical or textual comparisons, and permit several values to be compared at once:

inclusion `.IN.`
non-inclusion `.NI.`

For example, the following output shows the analysis for male farmers in the degree or postgrad groups:

```
SPLOAD [PRINT=*] 'FBS_England_Merged.gsh';ISAVE=alldata

TEXT [VALUES=degree,postgrad] ed2
RESTRICT alldata[];CONDITION=sex.IN.'male'.AND.education.in.ed2

SVTABULATE [PRINT=means; CLASS=education; STRATUM=stratum; WEIGHTS=weight] farmincome

Means for subpopulation defined restriction in farmincome with 95% confidence limits
------------------------------------------------------------------------------------

                      n   Sum wts      Mean       s.e.   %RSE/CV     Lower     Upper
 05farmer.education
  school only         0         0         *          *         *         *         *
         GCSE         0         0         *          *         *         *         *
     A levels         0         0         *          *         *         *         *
      college         0         0         *          *         *         *         *
       degree       222      6789     38041       5063     13.31     28110     47972
     postgrad        41      1645      9757       4682     47.98       574     18940
   apprentice         0         0         *          *         *         *         *
        other         0         0         *          *         *         *         *
         Mean       263      8435     32524       4203     12.92     24281     40767

Standard errors based on Taylor series approximations. Confidence limits use t-distribution
with 1701 d.f.


RESTRICT alldata[]
```

Notice that it is good practice to remove restrictions when they are no longer required, by giving a `RESTRICT` command with no `CONDITION` set. Otherwise unexpected results can arise when multiple restrictions are applied to the same variables.

In the above examples we want to confine the analysis temporarily to a subset of the data. Sometimes, however, there is a need to exclude part of the dataset permanently, and this may be achieved by using the `SUBSET` command. The syntax is slightly different to `RESTRICT` in that `CONDITION` is an option not a parameter. The following example shows how farms with negative incomes can be excluded from the dataset.

```
  66  DESCRIBE [SELECTION=nobs,mean,min,max] farmincome


Summary statistics for farmincome
=================================

      Number of observations = 1776
                        Mean = 30540
                     Minimum = -448626
                     Maximum = 3273062

  67  SUBSET [farmincome.GE.0] alldata[]
  68  DESCRIBE [SELECTION=nobs,mean,min,max] farmincome


Summary statistics for farmincome
=================================

      Number of observations = 1413
                        Mean = 43785
                     Minimum = 23
                     Maximum = 3273062
```

Whilst `SUBSET` is frequently useful in writing programs, it should not normally be used with survey commands such as `SVTABULATE`, except for removing unsampled units or units not forming part of the population. This is because calculation of the correct standard errors for a sub-population uses information from the whole sample, not just the units in the groups of interest. Instead `RESTRICT` should be used to define the sub-population, as described above.

6 Survey design and sampling

So far we have considered analysis with little, if any, consideration of the design of the survey. This reflects the reality that many statisticians, particularly those at the start of their careers, analyse surveys which they have not themselves designed. In this chapter we will partially redress this balance. However, in doing so we shall concentrate on practical issues; we do not have the space here to consider the full theory of survey design.

In this chapter you will therefore learn about

- selecting random samples
- stratified random samples
- sample selection for cluster and two-stage designs

6.1 Selecting random samples

To illustrate the principles of sample selection, we shall consider how to select a simple random sample from the June agricultural survey population in `Junemod.gsh` using the **Survey sampling** menu. The appropriate settings are shown in Figure 6.1 to take a 10% sample of the farms. The proportion of farms to sample is put in the **Numbers/proportion to sample** box; Genstat determines automatically whether numbers or proportions have been given, treating them as proportions if the highest value is less than 1. The **Units in population** box is set to the total number of farms in the population (19156).

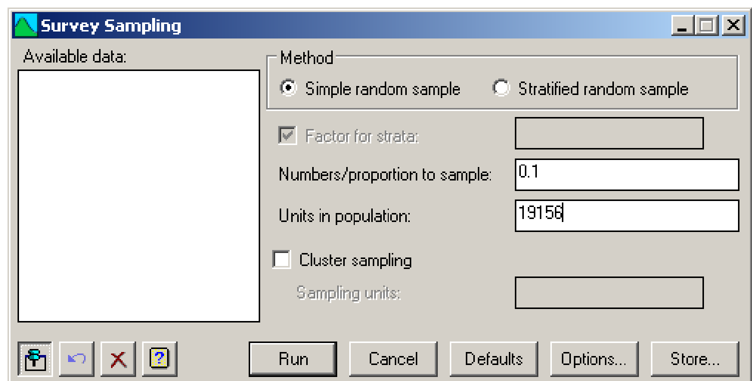


Figure 6.1

In order to save details of the units selected, it is necessary to click on the **Store** button. **Sampled units** box can be used to identify the selected units, and in Figure 6.2 this has been set to a variate called `sampno`. If the **Output data format** is set to **whole population**, then this variable will contain a 1 where a unit is sampled and a 0 where it is not selected. Alternatively, if the **Output data format** is set to **sampled units**

only, then it contains the row numbers of the selected units. In this dataset, farms are identified by a holding number stored in the variate `holding`, and so it is useful to have a list of the selected numbers. This can be achieved by placing the cursor in the **Existing variable** box and then double clicking `holding` in the **Available data list** to move it across (Figure 6.2). The name `sampled_holding`, for the list of sampled units, is then entered in the **New variable name** box before clicking the **Add to saved variables** button. Additional variables can be added to the **Currently saved variables** list if required, thus building a new dataset containing details of the selected units.

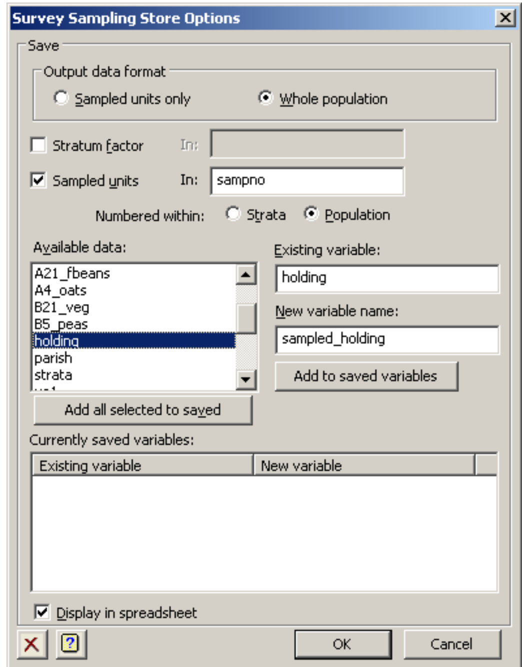


Figure 6.2

6.2 Selecting stratified random samples.

Let us now see how the above ideas can be extended to stratified random samples. Where, as in the June Survey example, a complete population dataset exists containing the stratification factor, one approach is to supply a list of numbers in the **Number/proportions to sample box** after ticking the **Factor for strata** box from the **survey sampling** menu. This is quick but carries more risk of error for designs with many strata and so an alternative is to supply the numbers in a table.

A new table can be created by selecting **create** from the **new** submenu of the **spread** menu. After clicking the Table icon, the **Create from existing factors** box

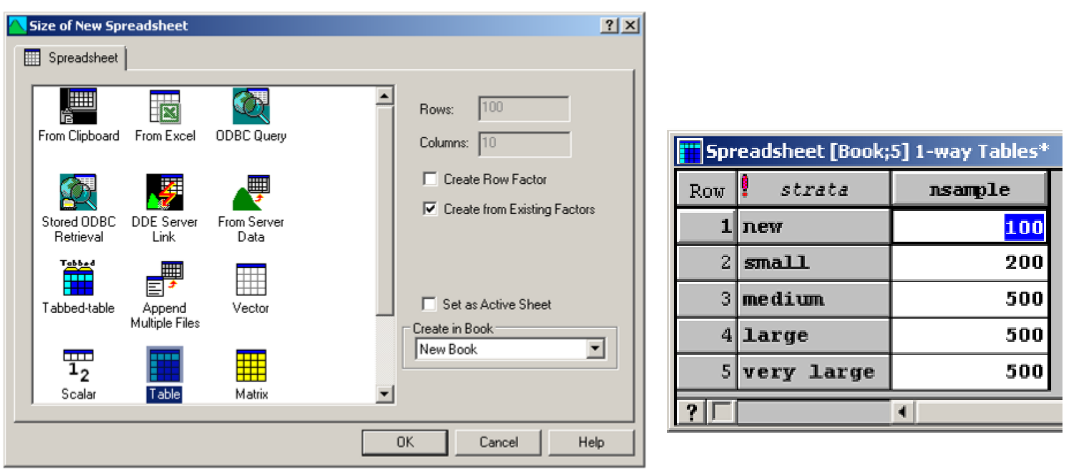


Figure 6.3

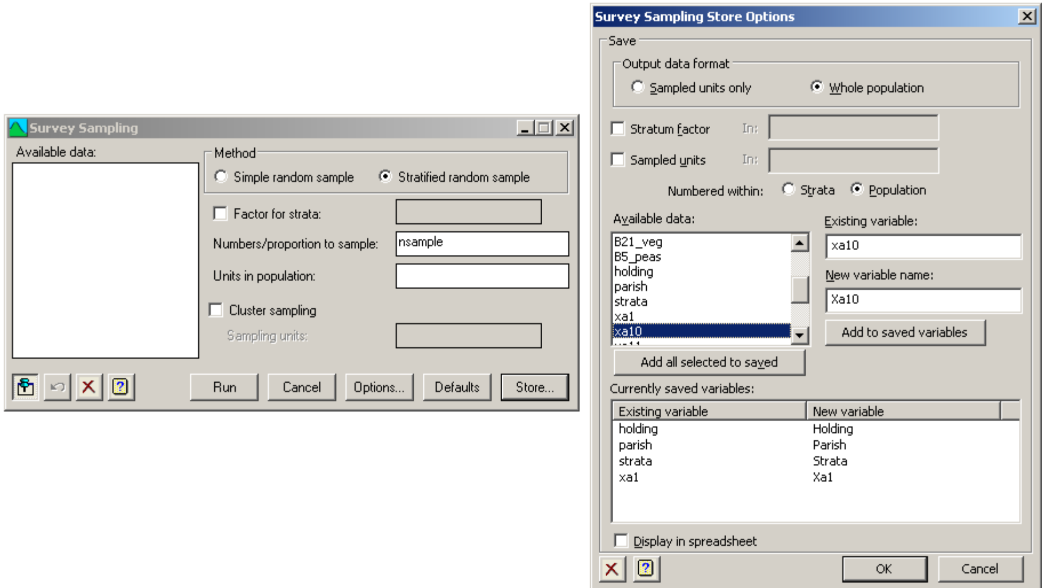
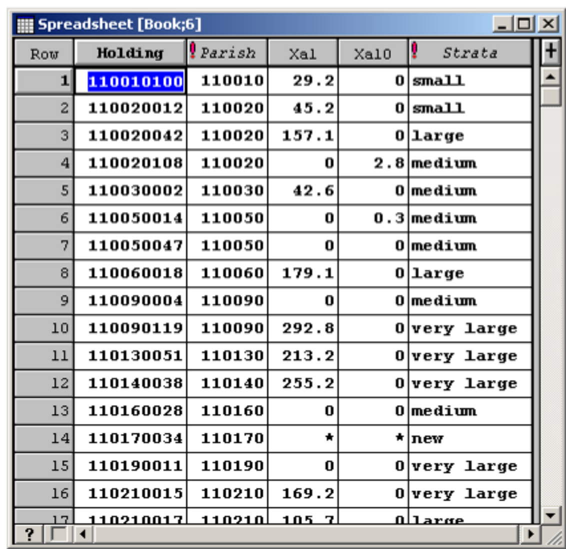


Figure 6.4

should be checked (Figure 6.3), and the factor strata selected from the list. The required numbers can then be entered in the table, as is shown on the right of Figure 6.3.

Once this has been completed, the table can be used as input for the **Survey sampling** menu, as is shown in Figure 6.4. Note that the **Units in population** box can be left empty, since this information can be deduced from the factor `strata` which classifies the `nsample` table. The right hand side of Figure 6.4 shows the settings of the **Survey Sampling Store Options** menu. Once again the **Output data format** has been set to **Sampled units only**, but this time a number of variables are shown in the **Currently saved variables** list in order to create the spreadsheet shown in Figure 6.5; this could be used for analysis once the response data is added. Note that the stratification factor for analysis is obtained by including `strata` in this list.



Row	Holding	Parish	Xa1	Xa10	Strata
1	110010100	110010	29.2	0	small
2	110020012	110020	45.2	0	small
3	110020042	110020	157.1	0	large
4	110020108	110020	0	2.8	medium
5	110030002	110030	42.6	0	medium
6	110050014	110050	0	0.3	medium
7	110050047	110050	0	0	medium
8	110060018	110060	179.1	0	large
9	110090004	110090	0	0	medium
10	110090119	110090	292.8	0	very large
11	110130051	110130	213.2	0	very large
12	110140038	110140	255.2	0	very large
13	110160028	110160	0	0	medium
14	110170034	110170	*	*	new
15	110190011	110190	0	0	very large
16	110210015	110210	169.2	0	very large

Figure 6.5

Alternatively it could be obtained by checking the **Stratum factor** box and supplying a name for the new factor in the associated box, but the approach used ensures that it appears in the same spreadsheet as the other new variables.

By default the following summary output is produced:

```
Survey sampling results
=======================


              Population      Sample   p sample
      strata
         new        2613         100      0.038
       small        5851         200      0.034
      medium        5479         500      0.091
       large        3074         500      0.163
  very large        2139         500      0.234
       Total       19156        1800      0.094
```

The above method assumes that there is an existing Genstat dataset defining each unit in the population. Sometimes this is not the case, and instead we want to create a new dataset as part of the sampling process. Figure 6.6 shows how the data should be organised in a spreadsheet (in this case a Genstat spreadsheet, but an Excel file could be used and imported using the Excel wizard).

Before this information can be used in the **Survey sampling** menu, `Strata`

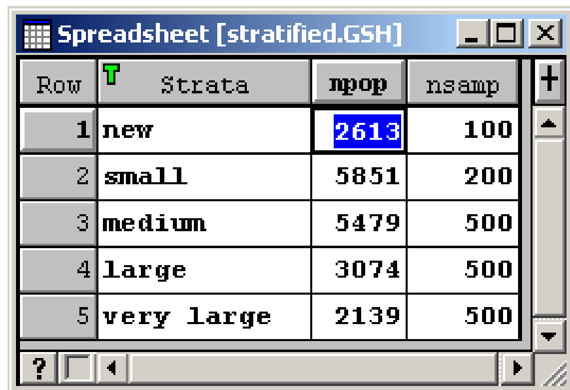


Figure 6.6

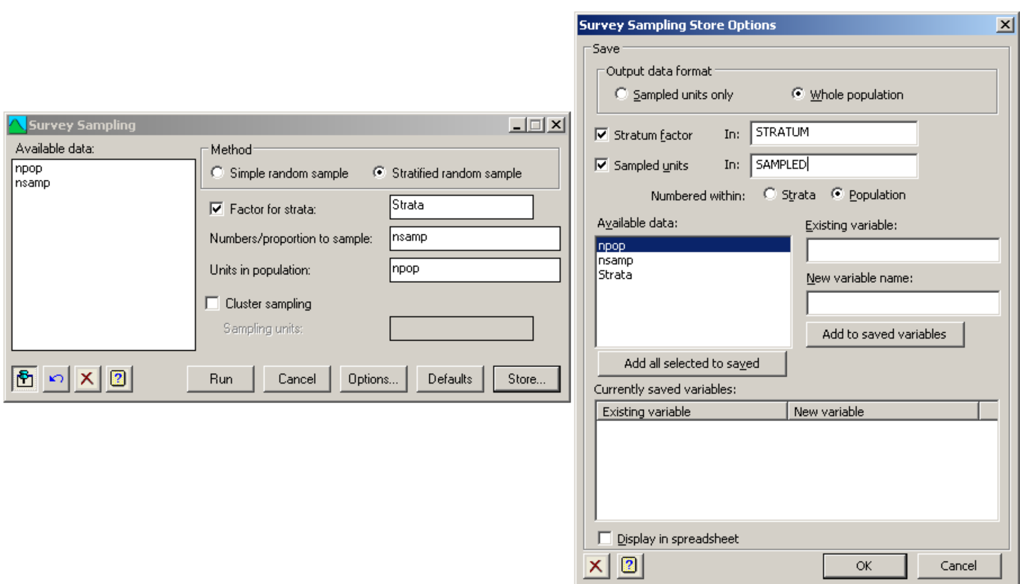


Figure 6.7

needs to be converted into a factor, for example by right mouse clicking on it and selecting **Convert to factor**. The settings for the **Survey sampling** menu are shown in Figure 6.7. Since, unlike in Figure 6.4, the structures `nsample` and `npop` are variates rather than tables, the **Factor for strata** box needs to be ticked and the factor name supplied in the box. In this example, the **Output data format** is set to **Whole population** in order to create a new dataset describing all units in the population, with variable `SAMPLED` having a value 1 where a unit is sampled (left hand side of Figure 6.8). Alternatively, the **Output data format** could be set to **Sampled units only**, in which case `SAMPLED` lists the numbers of the sampled units. With the latter format it is usually appropriate to set the **Numbered within** radio button to **Strata**; this will be useful, for example, where a numbered list of units is available for each of the strata. The format is shown on the right hand side of Figure 6.8.

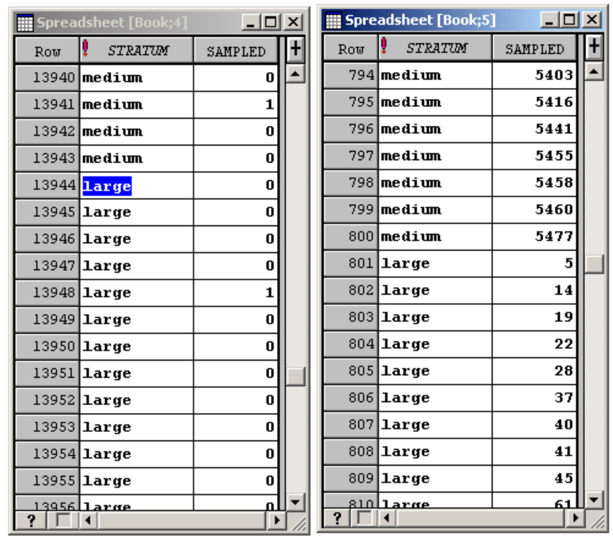


Figure 6.8

6.3 Cluster and multi-stage sampling

Sometimes, rather than sampling individual units at random, we wish to sample groups of units together; this is known as a cluster sample. For example, in the

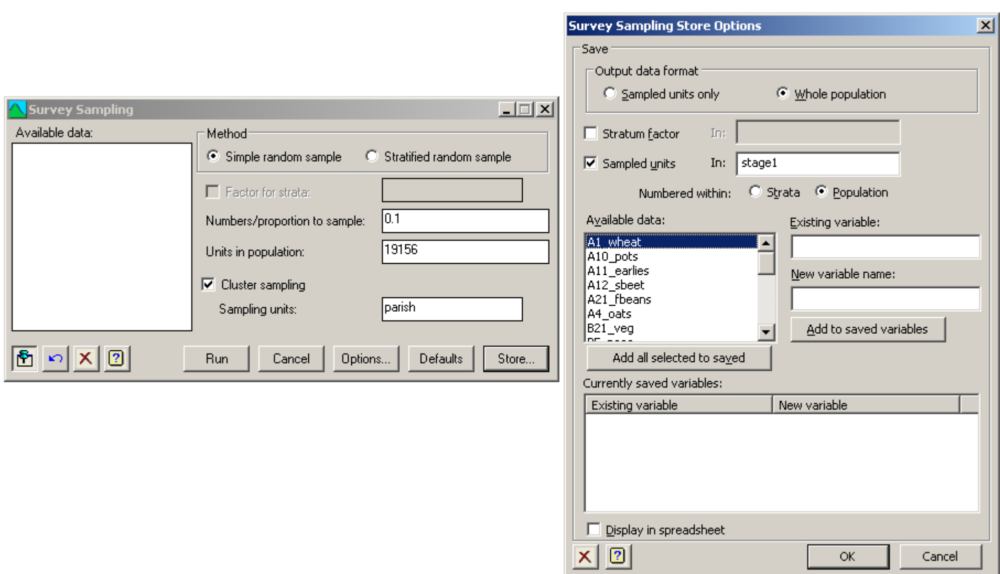


Figure 6.9

June Survey dataset, the holdings are grouped into parishes. Let us suppose that we wish to sample 10% of the parishes, collecting data from all holdings in the selected parishes. For simplicity, we will not stratify the sample, but the same approach can be extended to stratified samples, provided that the cluster units are nested within the strata.

Figure 6.9 shows the settings to achieve this; they are identical to those in Figure 6.1 except that `parish` is entered in the **Sampling units** box. The output produced is shown below; the population size is now shown in terms of the number of parishes.

```
Survey sampling results
=======================


               Population       Sample    p sample
  psu_stratum
 Unstratified        2701          270       0.100
        Total        2701          270       0.100
```

This is all that is required for a cluster sample in which data is collected from all units within the selected clusters. However, sometimes a second stage of sampling is required to select a subset of units from the clusters selected by the first stage; this is a multi-stage sample. For this exercise we will assume that it is required to sample 40% of holdings in those parishes selected in the first stage.

To achieve this with the example, the parishes are treated as if they are strata and a table is created containing the sampling proportions or numbers for each parish. (If the sampling fraction is the same for all parishes, unstratified sampling could be used, but we will not use this method since it cannot be applied to more

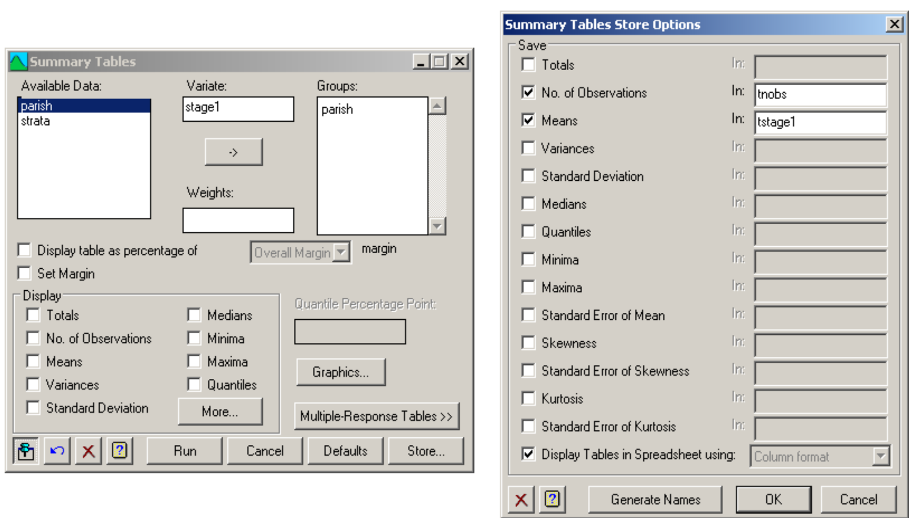


Figure 6.10

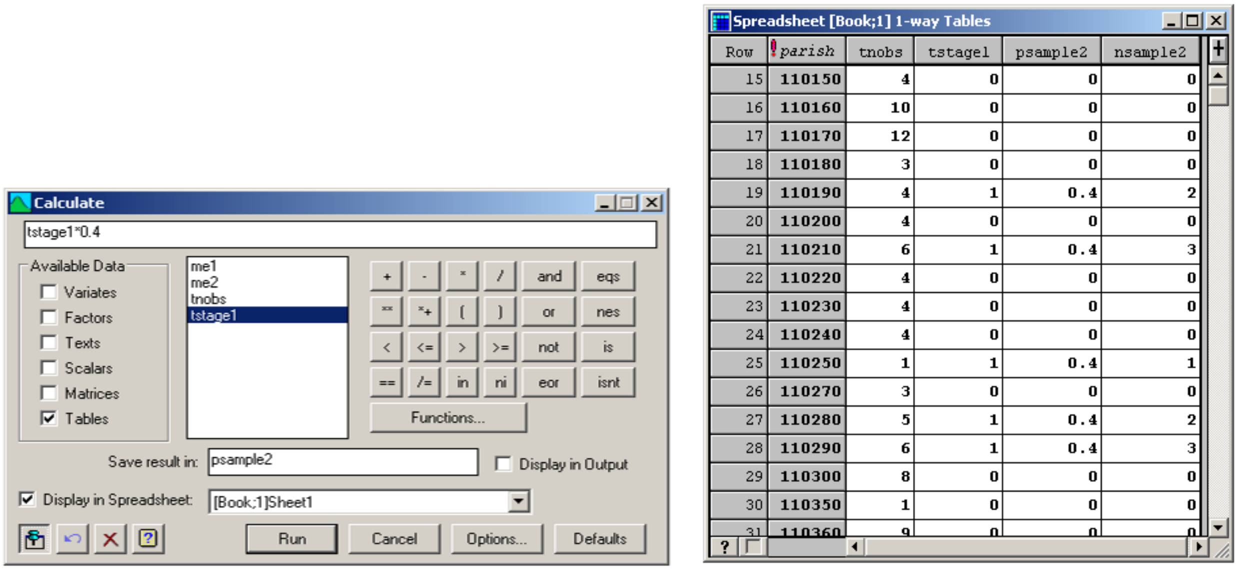


Figure 6.11

complex situations). The table can easily be created using **Summary tables** from the **Survey analysis** menu (Figure 6.10), provided that the **Whole population** option was selected in the first stage of sampling, as shown in Figure 6.9. The table of means produced in table `tstage1` will then contain the value one for holdings sampled in the first stage and a zero for those not sampled. Selecting **Calculate**, then **Column**, from the **Spread** menu, enables us to multiply this table by 0.4, as shown in Figure 6.11, to produce the required table of sampling proportions for the second stage.

An undesirable property of the sampling proportions in table `psample2` is that, because some parishes contain just a single holding, the 40% sample will result in

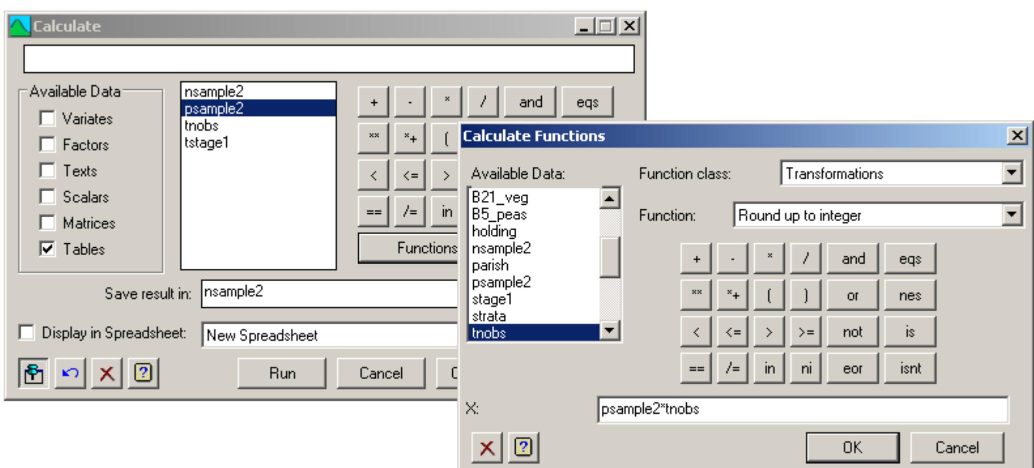


Figure 6.12

no holdings being sampled in these parishes. This problem can be solved by calculating the numbers to sample from the proportion by multiplying the sampling proportion by the number of holdings using the `CEILING` function to round up to the nearest whole number, as is shown in Figure 6.12.

Finally Figure 6.13 shows the settings to obtain the final sample, and the extract of the output corresponding to the parishes shown in the previous figure is shown below.

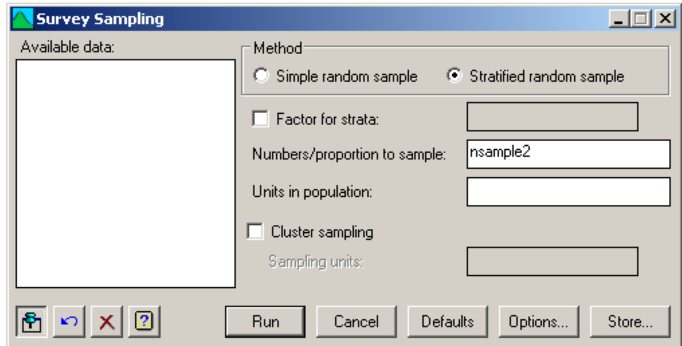


Figure 6.13

```
Survey sampling results
=======================

             Population      Sample    p sample
       parish
       110010         6           0       0.000
       110017         1           0       0.000
       110020        10           0       0.000
       110030         3           0       0.000
       110050         6           0       0.000
       110060         7           3       0.429
       110070         4           0       0.000
       110080         6           0       0.000
       110090        12           5       0.417
       110100         5           0       0.000
       110110         4           0       0.000
       110120        14           0       0.000
       110130         7           3       0.429
       110140         9           4       0.444
       110150         4           0       0.000
       110160        10           0       0.000
       110170        12           0       0.000
       110180         3           0       0.000
       110190         4           2       0.500
       110200         4           0       0.000
       110210         6           2       0.333
       110220         4           0       0.000
       110230         4           0       0.000
       110240         4           0       0.000
       110250         1           1       1.000
       110270         3           0       0.000
       110280         5           2       0.400
       110290         6           2       0.333
```

7 Regression models for survey data

As well as producing tables of means and totals, the analysis of surveys will frequently involve fitting models to explore relationships between variables. Thus in a health survey, we may want to explore the characteristics of people suffering from a particular disease, or in a wildlife survey we might relate the presence of a particular species to the characteristics of the surveyed sites.

In this chapter you will learn about

- whether a weighted model is appropriate
- how to fit weighted linear regression models with appropriate variance estimates
- using bootstrapping to obtain standard errors for more complex models
- the relationship with the methods of Chapter 3

7.1 To weight or not to weight

Survey weights are designed to produce unbiased estimates of population parameters, so it might seem logical to use them in all analyses. However, bias is not the only consideration when determining an appropriate analysis. An unbiased estimator with very wide confidence limits is, in practice, less useful than a more precise, but slightly biased one. When survey weights within a stratum are highly variable, estimates formed using those weights will be imprecise, and so there may be a case for using an unweighted estimate instead, provided there are grounds for believing the bias to be small.

The above argument applies to the estimation of any statistic but, in the case of regression, there are also other considerations. Regression may be used in a 'descriptive'[7] way, in which the objective is to produce an unbiased estimate of the relationship between two variables. Weights would generally be used for this type of analysis. However, regression is often used in a more 'analytical' way to explore relationships in the survey dataset. In this situation it is often important not to miss important relationships, and it may be sensible to accept a limited amount of bias in order to achieve this.

It is also important to consider the population to which inferences from the regression analysis apply. When using a survey to estimate a mean or a total it is

[7] See Chapter 4 of Analysis of Health Surveys by E.L. Korn and B.I.Graubard (1999, Wiley).

generally clear that we want to produce an estimate that is applicable to the particular population from which we sampled. For example, in the case of the analysis of Section 3.2 it is clear that the estimated average income applies to commercial farms in England in the year of the survey, and we would not usually expect to extrapolate this to farms in a different year or a different country.

This is also sometimes the case in regression analysis of survey data, particularly when we are using regression in a descriptive setting, maybe to improve our estimates of means or totals. Here the confidence limits of a regression slope represent the uncertainty in the estimate of the relationship in the population. Thus, if we had the full data from every unit in the population for both the dependent and independent variables in the regression, we would know the true slope and no confidence limits would be needed.

However, when regression is used in an analytical context, the relationships may have wider applicability. For example, we might model the relationship between farm income and a variety of characteristics of the farms, in order to suggest how farmers could improve their incomes. These results might be used to influence government policy to the farming sector in future years, on the ground that the underlying relationships would continue to hold, even if the incomes themselves changed, for example as a result of changes in commodity prices. In this analysis we are interested in a wider ‘super-population’ of farms, rather than just the population existing in the year of the survey, and it may therefore be more appropriate to apply conventional regression analyses for an infinite population, rather than sample survey estimators.

If it is decided to adopt a standard, unweighted regression analysis, it is still important to consider the survey design when deciding what terms to include in the model. We will discuss this later in the chapter.

7.2 Linear regression for surveys

The approach to survey regression implemented in Genstat is based on the same Taylor series approximation as in the methods of Chapter 3. The analysis produces identical parameter estimates to an ordinary regression with the appropriate weighting. However, the variances are calculated by an approximation that allows for the lack of independence that results from the structure of the survey. Also, unlike ordinary generalized linear models, the residual variance is estimated separately in each stratum; this can be important when the magnitude of the response variable differs substantially between strata, as is often the case in business surveys.

To illustrate the weighted analysis of survey data, we will use another subset of the Farm Business Survey data and investigate how the amount of Government support received by farms (`subsidy`) is related to the area of the farm (`farmarea`). The data are in `FBS_Regression.gsh`.

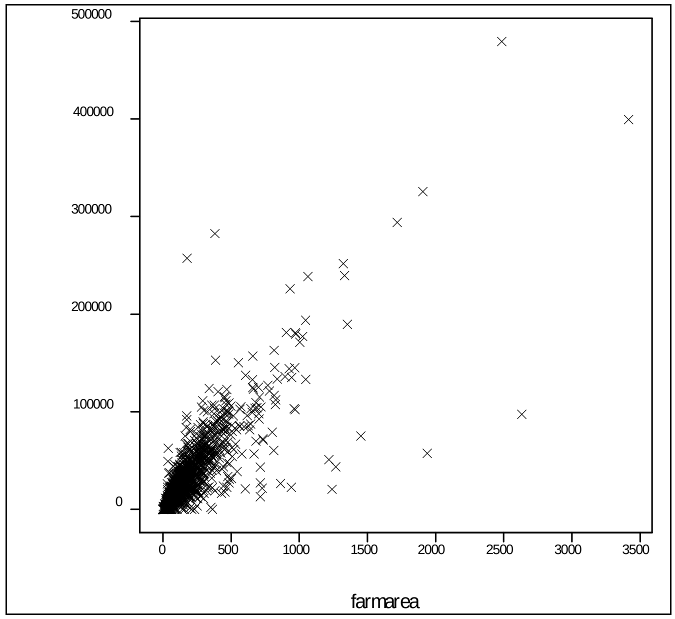


Figure 7.1

Before fitting any regression model, it is sensible to plot the relationship between the variables. This is shown in Figure 7.1 which was drawn by selecting **2-D Scatter Plot** from the **Graphics** menu. The most striking feature is that both variables show a skew distribution, with a few relatively large values, but most points in the bottom left hand corner of the plot. With an ordinary regression analysis some form of transformation, probably using logs, would be needed to meet the assumption of a Normal distribution of errors. For survey regression, as with the estimation of survey means and totals, we are not relying on Normality, and so a transformation is not absolutely necessary. However, unless there is a strong

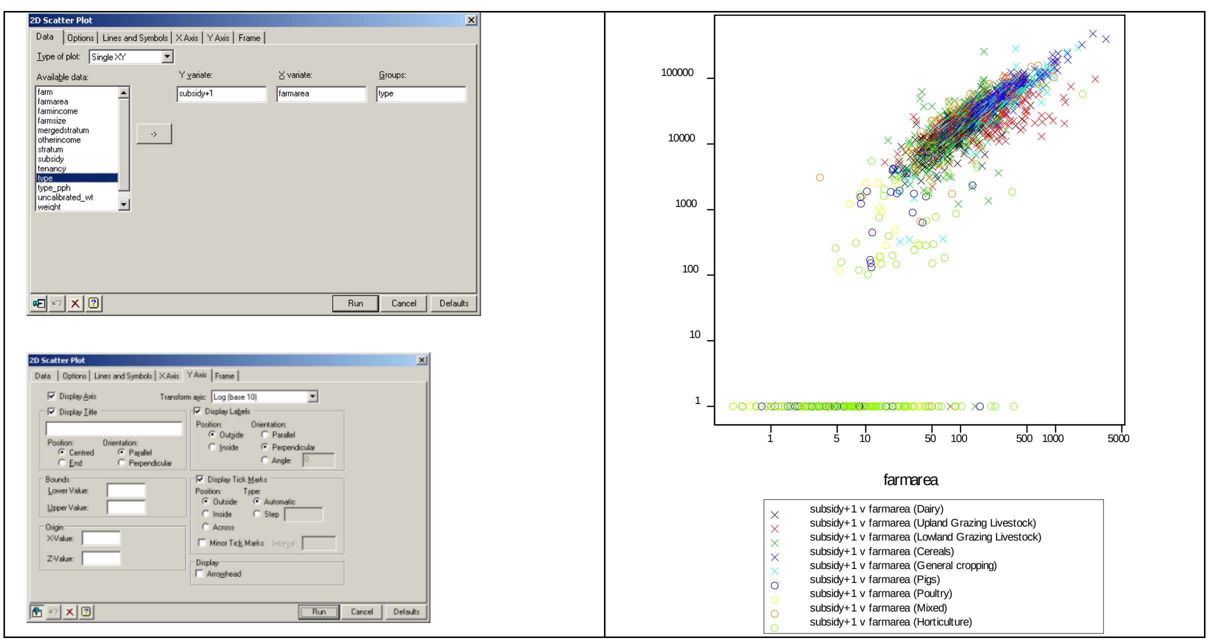


Figure 7.2

reason for wanting to work on the natural scale, it may be preferable to transform the data anyway, because otherwise the outlying high values will have high leverage and may distort the relationship.

Figure 7.2 shows the settings of the **2-D Scatter Plot** menu to plot the variables on the log scale. Note that because `subsidy` contains some zero points, the y variable is set to `subsidy` + 1 so that these can be displayed (if this is not done, Genstat will not display the y-axis on the log scale). The second step is to set the **Transform axis** box to **Log(base 10)** for both the **Y Axis** and **X Axis** tabs. The resulting graph is shown on the right of Figure 7.2. It is now clear that there is a strong approximately linear relationship, but there are a row of points along the bottom with zero `subsidy` (i.e. a value of 1 for `subsidy` + 1). The graph also shows that almost all of the points in this row represent pig, poultry or horticultural farms; these are sectors that received no subsidies in the past and have much lower rates of uptake of the current support payments. It therefore makes sense to exclude these farm types from the analysis by selecting **Restrict/Filter** from the **Spread** menu and then choosing **To Groups (factor levels)**.

Figure 7.3 shows the settings of the **Generalized Linear Models for Survey Data** menu to fit the regression model. Note the use of `mergedstratum` to avoid the problems caused where there is a single valid observation in some strata. The output is shown below.

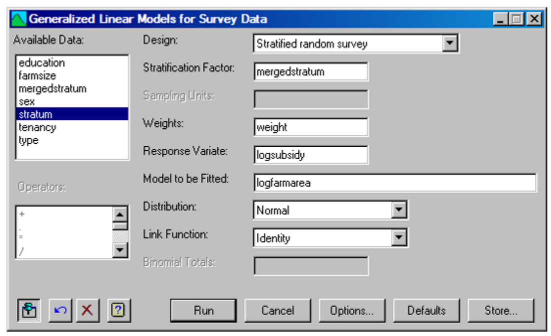


Figure 7.3

```
Regression analysis
===================

 Response variate: logsubsidy
   Weight variate: scaledwts
     Fitted terms: Constant, logfarmarea

   Supplied weights:  weight
             Strata:  mergedstratum
  Observations used:  1449
           PSU used:  1776
    Population size:  61653
  Obs in sub-population:  1449
  Subpopulation size:  52988
          CI method:  tdistribution (95% limits)
```

```
Estimates of parameters with 95% confidence limits
--------------------------------------------------

                    Estimate        s.e.       Lower       Upper
    Constant           2.289       0.072       2.148       2.431
 logfarmarea           0.961       0.032       0.898       1.024

Standard errors are based on Taylor series approximations. Confidence limits use t-
distribution with 1728 d.f.
Based on subpopulation defined by restriction in logsubsidy
```

Note that the regression slope is close to 1.0. As increasing the log value by 1.0 is equivalent to a ten-fold increase on the natural scale (remember we used logs to the base 10), this implies that a tenfold increase in `farmarea` results, on average, in a roughly tenfold increase in `subsidy`.

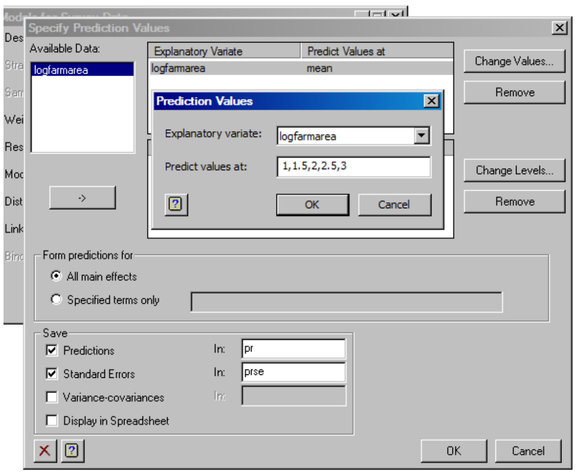


Figure 7.4

Interpreting a list of regression coefficients can be difficult, particularly in more complex models containing interaction terms. In these situations it is often helpful to examine tables of predictions from the model. Figure 7.4 shows how this may be achieved by clicking on the **Specify Prediction values** button on the **Options** menu. The variable `logfarmarea` is clicked across into the **Explanatory Variate** box. By default, values are predicted at the mean value, but by highlighting the row and clicking the **Change Values** box a list of values can be specified as shown. The output is shown below.

```
Predictions from regression with 95% confidence limits
------------------------------------------------------

Predictions for logfarmarea

                  Prediction        s.e.       Lower       Upper
 logfarmarea
         1.0           3.250     0.04065       3.170       3.330
         1.5           3.731     0.02554       3.681       3.781
         2.0           4.211     0.01308       4.185       4.237
         2.5           4.691     0.01444       4.663       4.720
         3.0           5.172     0.02766       5.118       5.226
```

```
* Note: Standard errors are based on Taylor series approximations. Confidence
limits use t-distribution with 1775 d.f.
```

In Figure 7.4 we have also saved the predictions in structure `pr`. Because predictions may be formed for more than one model term, `pr` is a pointer with one element for each requested term. In this simple case, where there is just one explanatory variate for which predictions are needed, `pr[1]` is a table containing the predictions.

7.3 Generalized linear models for surveys

In the above example, we used a log-transformation to achieve approximate Normality of the response variable. In other situations we may prefer to fit a generalized linear model (GLM) with error distribution other than the Normal distribution or with a different link function. See Chapter 3 of *A Guide to Regression, Nonlinear and Generalized Linear Models in Genstat* for more details of the range of models available.

To illustrate the use of GLMs we shall investigate the characteristics of those pig, poultry and horticultural farms that did not claim any support payments and hence appeared in the row of points at the bottom of Figure 7.2. The first step is to construct a new variable taking the value 1 for these farms and 0 for the farms where `subsidy` is greater than zero. This can be done by selecting the **Spread** menu and then **Column** from the **Calculate** sub-menu (Figure 7.5). The resulting variable is then analysed using a GLM with a binomial distribution, with the number of binomial trials set to 1 (Figure 7.6). Initially we will try using the log of the farmed area and the farm type as explanatory variables. We will restrict the analysis to the three types of farms that we are interested in, and we will use variable `type_pph`, which has levels and labels only for the three types, rather than `type`, to avoid warning messages relating to the farm types not of interest. Taylor series approximations are not available for non-Normal models

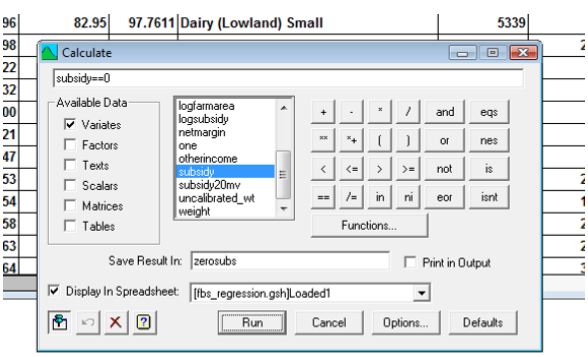


Figure 7.5

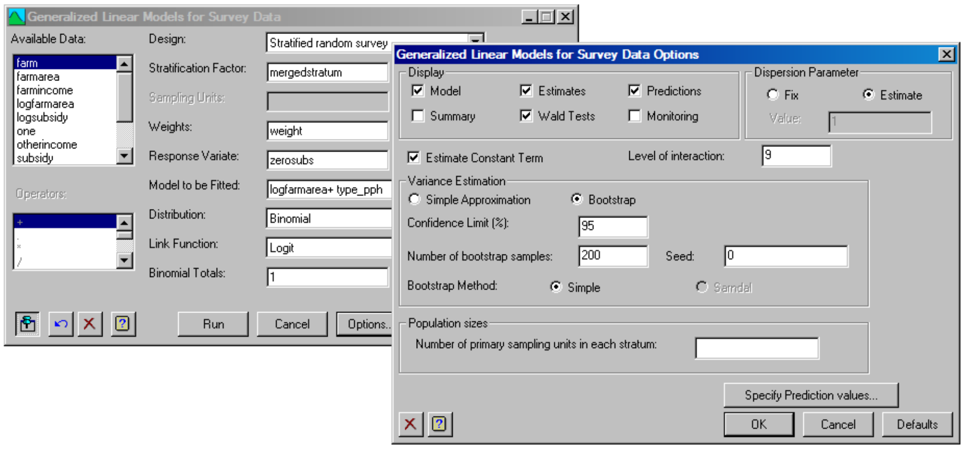


Figure 7.6

in Genstat at present, so instead we select the bootstrap variance method with two hundred bootstrap samples; this is sufficient to produce reasonably robust preliminary results without taking too long, although it is best to use several thousand for the final analysis if bootstrap confidence limits and Wald test statistics are required. The output is shown below.

```
Regression analysis
===================

 Response variate: zerosubs
  Binomial totals: 1
     Distribution: Binomial
    Link function: Logit
   Weight variate: scaledwts
     Fitted terms: Constant + logfarmarea + type_pph

   Supplied weights:  weight
             Strata:  mergedstratum
  Observations used:  327
           PSU used:  1776
    Population size:  61653
  Obs in sub-population:  327
  Subpopulation size:  8665
  Bootstrap samples:  200
   Bootstrap method:  simple
          CI method:  tdistribution (95% limits)

Estimates of parameters with 95% confidence limits
--------------------------------------------------
```

```
                              Estimate        s.e.       Lower       Upper
             Constant             3.69        0.57        2.57        4.81
          logfarmarea            -3.47        0.49       -4.42       -2.51
     type_pph Poultry             0.52        0.50       -0.46        1.49
type_pph Horticulture             0.86        0.44       -0.01        1.73

Standard errors based on 200 bootstrap samples. Confidence limits use t-
distribution with 1728 d.f.
Based on subpopulation defined by restriction in zerosubs

Wald Tests
----------

       Term       Wald          F        df1        df2          P
logfarmarea      50.51      50.51          1       1728     <0.001
   type_pph       3.82       1.91          2       1727      0.149
```

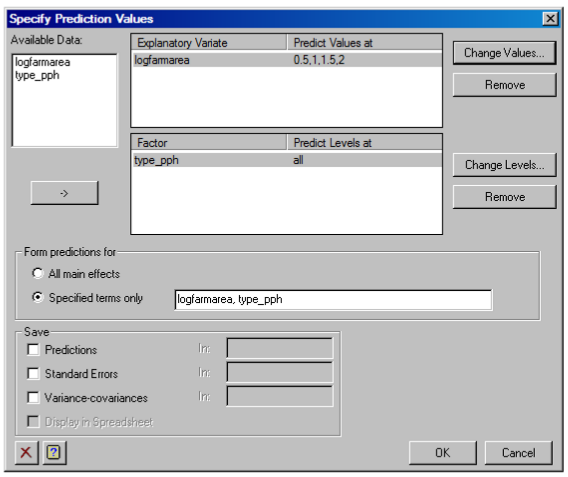


Figure 7.7

Wald tests for the significance of the fitted terms are also shown; the test statistics are calculated using a variance-covariance matrix derived from the bootstrap parameter estimates. These statistics are particularly useful for factors with more than two levels, when the statistical significance of differences cannot easily be deduced by examining the estimates and their standard errors. In this case `logfarmarea` is very highly significantly different from zero, whereas `type_pph` is well above the conventional 0.05 level of significance.

Once again, it is useful to form predicted values to give a better impression of the results. The settings for this are shown in Figure 7.7. By default, predictions are formed for all combinations of the variables, which in this case would mean a table with rows representing the different values of `logfarmarea` and the columns different levels of `type_pph`. To produce separate tables for `logfarmarea` and `type_pph` these terms are listed in the **Specified terms only** box.

```
Predictions from regression with 95% confidence limits
------------------------------------------------------

Predictions for logfarmarea

               Prediction        s.e.       Lower       Upper
  logfarmarea
          0.5      0.9302     0.02475      0.8816      0.9787
          1.0      0.7072     0.05088      0.6074      0.8070
          1.5      0.3064     0.05220      0.2040      0.4088
          2.0      0.0734     0.02597      0.0225      0.1244


Predictions for type_pph

               Prediction        s.e.       Lower       Upper
     type_pph
         Pigs      0.6523     0.07824      0.4989      0.8058
      Poultry      0.7586     0.07361      0.6142      0.9030
 Horticulture      0.8156     0.06455      0.6890      0.9422

* Note: Standard errors based on 200 bootstrap samples. Confidence limits use t-
distribution with 1728 d.f.
```

Looking at the output above, it can be seen that around 71% of farms with 10ha (i.e. `logfarmarea` = 1) do not claim support payments, but this falls to only 7% of those with 100ha (`logfarmarea` = 2). By contrast, as would be expected from the non-significant Wald test statistic, there is much less difference between the predictions for the different levels of `type_pph`, with an estimated 65% of pig farms, 76% of poultry farms and 82% of horticultural farms not claiming payments. The confidence limits shown are based on the t-distribution and the bootstrap standard error of each predicted value; this is the default for less than 400 bootstrap samples. With larger numbers of bootstrap samples, confidence limits are derived from the appropriate percentiles of the distribution of bootstrapped predicted values.

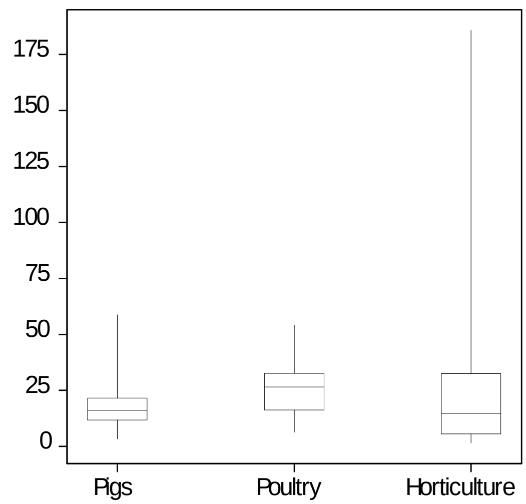


Figure 7.8

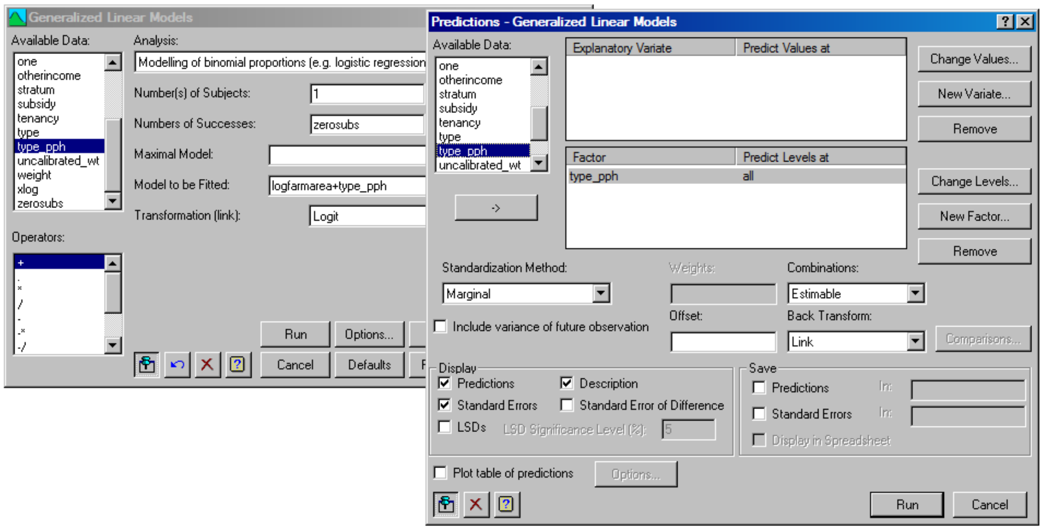


Figure 7.9

7.4 Fitting unweighted models

As discussed in Section 7.1, it may be useful to consider an unweighted model fitted by standard regression approaches, particularly when the weights are highly divergent. Figure 7.8 contains a boxplot of weights for the three farm types used in fitting the logistic regression model of Section 7.3, showing that the weights are particularly variable for horticultural farms. It is therefore sensible to compare the results above with those from an unweighted model.

The output below shows predictions for type_pph for a logistic regression regression model of zerosubs fitting explanatory variables for logfarmarea and type_pph (Figure 7.9).

```
Predictions from regression model
---------------------------------

These predictions are estimated mean proportions, formed on the scale of the
response variable, corresponding to one binomial trial.

The predictions have been formed only for those combinations of factor levels for
which means can be estimated without involving aliased parameters.

The predictions are based on fixed values of some variates:
        Variate   Fixed value   Source of value
    logfarmarea        0.9886   Mean of variate
```

```
The standard errors are appropriate for interpretation of the predictions as
summaries of the data rather than as forecasts of new observations.

Response variate: zerosubs

               Prediction        s.e.
     type_pph
         Pigs      0.5335     0.08424
      Poultry      0.6905     0.07337
 Horticulture      0.8504     0.03120

* MESSAGE: s.e's, variances and lsd's are approximate, since the model is not
linear.

* MESSAGE: s.e's are based on dispersion parameter with value 1
```

Compared to the equivalent weighted results, there are some big differences in the parameter estimates, especially for pig farms. A deviance test for adding `type_pph` to the model is highly significant (χ^2 = 7.53 with 2 d.f., P<0.001). In addition, the standard error for horticulture farms is much lower at 0.031 compared to 0.072 in the weighted analysis; the lower standard error for horticultural farms in the conventional analysis reflects their larger sample size, whereas in the weighted survey analysis this is counteracted by the variable weights for this farm type. Such differences are not unusual when sample sizes are relatively small, but do indicate that results should be treated with caution.

When fitting unweighted regression models to survey data it is good practice to include variables relating to the survey design in the model, and to check for interactions between these and the explanatory variables of interest. However this can be problematic when the design variables themselves influence the response variable. In the current example, the strata are based on a combination of farm type (`type`) and economic size (`farmsize`); thus the `mergedstratum` factor cannot be included in the model because it is aliased with `type_pph`. The factor `farmsize` can be included in the model, although it might itself have an impact on whether a farm claims subsidy and it is also correlated with the physical size of the farm. If `farmsize` is fitted, `type_pph` ceases to be significant and this may indicate that the discrepancy between the weighted and unweighted results is related to the differences in economic size between the groups of farms.

7.5 Relationship with cross-tabulations

When the explanatory variables in a weighted survey regression with Normal errors are all factors, prediction will produce the same results as the cross-tabulation methods of Chapter 3. This is illustrated in the practical of Section 7.6 below.

The equivalence between the two approaches can be useful when fitting more complex models. For example, if we wish to estimate `farmincome` by for all combinations of `type` and `tenancy` this could be done either using either the **General Survey Analysis** menu or **Generalized Linear Models for Survey Data** menu fitting the model `type*tenancy`. However, some cells are based on low numbers of observations and may be unreliable. An alternative model which avoids this problem involves fitting the main effects only by using `type+tenancy` in the **Model to be Fitted** box of the **Generalized Linear Models for Survey Data** menu.

7.6 Practical

To illustrate the equivalence of the two approaches, use the dataset in `FBS_Regression.gsh` to predict mean `farmincome` levels by farm type using the **Generalized Linear Models for Survey Data** menu with `mergedstratum` as the stratification factor. Then repeat the analysis using the **General Survey Analysis** menu.

Appendix 1: Genstat code for all examples

This appendix shows the code required to generate the analyses shown or described in the text. The code is simplified as much as possible, for example by omitting options set by the menus despite using the default values, but names of commands, parameters and options are not generally abbreviated.

1 Basic principles

1.1-1.3 Getting the data into Genstat

Note use of backslash (or double forward slash) in pathnames.

```
IMPORT 'C:/Progra~1/Gen16Ed/Data/Province.xls';\
  SHEET='simple RS full pop'; ISAVE=ipo

SVSTRATIFIED [PRINT=summary,totals,means] unemployment; LABELS=municipality

"Section 1.2 - repeat above command saving TOTALS"
SVSTRATIFIED [PRINT=summary,totals,means] unemployment; LABELS=municipality; \
  TOTALS=tot_unemploy; SETOTALS=se_tot
FSPREADSHEET tot_unemploy,se_tot

" Section 1.3 - again repeat, this time printing influence stats
  and plotting graph "
SVSTRATIFIED [PRINT=summary,totals,means,influence; PLOT=single] unemployment; \
  LABELS=municipality
```

1.4 Practical

Two alternatives are shown below to construct `unemployment2`; one requires knowledge of the row number to be replaced by a missing value, whereas the other works with the name of the municipality. The latter uses the `MVINSERT` function; the first argument is the original version of the data, the second is a logical expression indicating the rows to replace with missing values.

```
IMPORT 'C:/Progra~1/Gen16Ed/Data/Province.xls'; \
  SHEET='simple RS full pop'; ISAVE=ipo

SVSTRATIFIED [PRINT=summary,totals,means] unemployment; TOTALS=tot_unemploy

DUPLICATE unemployment;NEWSTRUCTURE=unemployment2
CALC unemployment2$[1]=CONSTANTS('missing')
" alternatively the following does the same as the above,
  but without the need to know the row to replace with a missing value"
```

```
CALC unemployment2=MVINSERT(unemployment;municipality.in.'Jyvaskyla')

SVSTRATIFIED [PRINT=summary,totals] unemployment2; TOTALS=tot_mv

PRINT (tot_unemploy-tot_mv)/tot_unemploy
```

1.5 Analysis with response data only

```
IMPORT 'C:/Progra~1/Gen16Ed/Data/Province.xls'; \
  SHEET='simple RS sample'; ISAVE=ipo
SVSTRATIFIED [PRINT=summary,totals,means] unemployment; LABELS=municipality; \
  NUNITS=32
```

1.6 Stratified random samples – factors and tables

In this example `stratum` is imported as a variate (although we could have added an exclamation mark after the column heading to force it to be a factor). It can be converted to a factor using the `GROUPS` command, with the option `REDEFINE` set to `yes`. Alternatively, a different name could have been used, i.e.:

```
GROUPS stratum; FACTOR=stratum2
```

The new factor is then used to create the table `popsize`, which specifies the population size in each stratum.

```
IMPORT 'C:/Progra~1/Gen16Ed/Data/Province.xls'; \
  SHEET='stratified sample'; ISAVE=ipo
GROUPS [REDEFINE=yes] stratum
TABLE [CLASSIFICATION=stratum; VALUES=7,25] popsize
SVSTRATIFIED [PRINT=summary,totals,means; STRATUM=stratum] unemployment; \
  LABELS=municipality; NUNITS=popsize
```

1.7 Practical

```
IMPORT 'C:/Progra~1/Gen16Ed/Data/Province.xls'; \
  SHEET='stratified full pop'; ISAVE=ipo
GROUPS [REDEFINE=yes] stratum
SVSTRATIFIED [PRINT=summary,totals,means; STRATUM=stratum] unemployment; \
  LABELS=municipality
```

2 Estimating totals in stratified random surveys

2.1 Design based estimators

To add labels to the factor, we first create them in a text structure. Note that quotation marks are only needed for the label that contains a space. Then the labels are added to the factor definition, with option `MODIFY=yes` to ensure that the existing values are retained.

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/June.gsh'; ISAVE=jpo
"set factor labels"
TEXT [VALUES=small,medium,large,'very large',new] labs
FACTOR [MODIFY=yes;LABELS=labs] strata
SVSTRATIFIED [PRINT=summary,totals; STRATUM=strata] A1_wheat; LABELS=holding
```

2.2 Ratio estimation

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/Junemod.gsh'; ISAVE=jpo
SVSTRATIFIED [PRINT=summary,totals,influence; PLOT=separate; METHOD=separate; \
  STRATUM=strata] A1_wheat; X=xa1; LABELS=holding
"and with compact output, setting the width of the output to give sufficient room"
OUTPUT [WIDTH=110] 1
SVSTRATIFIED [PRINT=summary,totals,influence; METHOD=separate; \
  STRATUM=strata; COMPACT=yes] A1_wheat; X=xa1; LABELS=holding
```

2.3-2.4 Using restrictions

In this example we could just restrict the response variable `A1_wheat`, but often easier to restrict all variables, using the pointer created by `ISAVE` parameter of `SPLOAD` or `IMPORT`. Remember to remove the restriction when no longer required, as it can lead to unexpected results in subsequent programming.

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/Junemod.gsh'; ISAVE=jpo
RESTRICT jpo[];CONDITION=holding.NE.343460118
"first, using default of excluding restricted row totally"
SVSTRATIFIED [PRINT=summary,totals; METHOD=separate; \
  STRATUM=strata] A1_wheat; X=xa1; LABELS=holding
"now adding it back in to the total"
SVSTRATIFIED [PRINT=summary,totals; METHOD=separate; \
  STRATUM=strata] A1_wheat; X=xa1; LABELS=holding
RESTRICT jpo[]  "remove restriction"
```

2.5 Practical

There are several possible ways of doing this in code. Here we use the `WHERE` function to find the row number of holding 343460118, and then use `CALCULATE` to change its stratum. Note that we reordered this factor in Section 2.1, so that its levels are not in numerical order, as would usually be the case.

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/Junemod.gsh'; ISAVE=jpo
" create new factor "
TEXT [VALUES=new,small,medium,large,'very large','outlier'] labs2
VARIATE [VALUES=99,2,3,4,5,6] levs2
FACTOR [LEVELS=levs2; LABELS=labs2] strata2;VALUES=strata
" find row number for outlier and set to outlier stratum "
CALC rowno=WHERE(holding.EQ.343460118)
CALC strata2$[rowno]=6
" use TABULATE to check everything has worked "
TABULATE [PRINT=count; CLASS=strata2,strata]
SVSTRATIFIED [PRINT=summary,totals; METHOD=separate; STRATUM=strata2] \
  A1_wheat; X=xa1; LABELS=holding
```

2.6 The combined ratio estimator

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/Junemod.gsh'; ISAVE=jpo
SVSTRATIFIED [PRINT=summary,totals,influence; PLOT=separate; METHOD=separate;\
  STRATUM=strata] A11_earlies; X=xa11; LABELS=holding
SVSTRATIFIED [PRINT=summary,totals,influence; PLOT=single; METHOD=combined;\
  STRATUM=strata; COMPACT=yes] A11_earlies; X=xa11; LABELS=holding
```

2.7 Saving and exporting results

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/Junemod.gsh'; ISAVE=jpo
SVSTRATIFIED [PRINT=summary,totals; PLOT=*; METHOD=separate; \
  STRATUM=strata] A11_earlies; X=xa11; LABELS=holding;\
  TOTALS=a11_tot; SETOTALS=a11_se; FITTED=a11_fit; INFLUENCE=a11_inf
FSPREAD holding,a11_fit,a11_inf
FSPREAD a11_tot,a11_se
```

3 General Survey Analysis

3.1 Farm Business Survey datset – merging data

Since both datasets are in farm order, and all the farms in the Genstat sheet are also in the Excel version, the easiest approach is to use `SUBSET` to remove the extra rows from the Excel data. If this were not the case, the `JOIN` command could be used instead. Note that both sheets contain a variate called `farm`, so we take a copy of the Genstat version before overwriting it by reading in the Excel data.

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England.gsh'; ISAVE=gpo
DUPLICATE farm; farmlist
IMPORT [EMETHOD=read; EXTRAROW=2] 'FBSdata.xls'; SHEET='FBS'; ISAVE=xlpo
" remove farms from excel sheet that are not in FBS_England.gsh "
SUBSET [CONDITION=farm.IN.farmlist] xlpo[]
" check that lists of farms are correct - this should always be zero "
DESCRIBE farm-farmlist
```

3.2 Cross-tabulation

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo
SVTABULATE [PRINT=summary,means; CLASS=sex; STRATUM=stratum; WEIGHTS=weight] \
  Y=farmincome; LABELS=farm
" and with wald stats and influence stats "
SVTABULATE [PRINT=summary,means,wald,influence; CLASS=sex; STRATUM=stratum; \
  WEIGHTS=weight] Y=farmincome; LABELS=farm
```

3.3 Sub-populations

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo
RESTRICT farmincome; CONDITION=sex.in.'male'
SVTABULATE [PRINT=summary,means; CLASS=education; STRATUM=stratum; \
  WEIGHTS=weight] Y=farmincome; LABELS=farm
RESTRICT farmincome
```

3.4 Practical

Note how multiple tables can be displayed together in the same spreadsheet using code, but not using the menus.

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo
RESTRICT farmincome; CONDITION=education.in.'school only'
SVTABULATE [PRINT=summary,means; CLASS=sex; STRATUM=stratum; WEIGHTS=weight] \
  Y=farmincome; LABELS=farm; MEANS=mean_sch; SEMEANS=sem_sch
RESTRICT farmincome
RESTRICT farmincome; CONDITION=education.in.'college'
SVTABULATE [PRINT=summary,means; CLASS=sex; STRATUM=stratum; WEIGHTS=weight] \
  Y=farmincome; LABELS=farm; MEANS=mean_col; SEMEANS=sem_col
RESTRICT farmincome
FSPREAD mean_sch,sem_sch,mean_col,sem_col
```

3.5 Counts and proportions

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo
SVTABULATE [PRINT=summary,means,totals; CLASS=sex; STRATUM=stratum; \
  WEIGHTS=weight] LABELS=farm
```

3.6 Ratios

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo
RESTRICT fpo[]; CONDITION=farmincome.GT.0
SVTABULATE [PRINT=summary,ratios; CLASS=farmsize; STRATUM=stratum; \
  WEIGHTS=weight; PLOT=single] Y=subsidy; X=farmincome; LABELS=farm
SVTABULATE [PRINT=summary,ratios; CLASS=farmsize; STRATUM=stratum; \
  WEIGHTS=weight; PLOT=separate] Y=subsidy; X=farmincome; LABELS=farm
RESTRICT fpo[]
```

3.7 Quartiles and bootstrapping

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo
SVTABULATE [PRINT=summary,means,quantiles; PLOT=*; CLASS=type; STRATUM=stratum; \
  WEIGHTS=weight; PERCENTQUANT=!(5,10,25,50,75,90,95)] \
  Y=farmincome; LABELS=farm
" and with bootstrap limits "
SVTABULATE [PRINT=summary,means,quantiles; PLOT=*; CLASS=type; STRATUM=stratum; \
  WEIGHTS=weight; PERCENTQUANT=!(5,10,25,50,75,90,95); NBOOT=200; METHOD=simple] \
  Y=farmincome; LABELS=farm
```

3.8 Multiple-response tables

Note that there is no separate option for multiple-response factors. Instead the pointer to the factors is listed as the `CLASSIFICATION` setting (or one of the settings for two-way tables).

```
IMPORT 'C:/Progra~1/Gen16Ed/Data/FBSmult.gwb'; SHEET='types'; ISAVE=mpo
FMFACTOR [MRESPONSE=livestock; SUFFIXNULL=0; LABELNULL='null'; CODENULL='-'] \
  an1,an2,an3
" now load the main data sheet and check the farm identifiers match "
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo
DESCRIBE farm-farm3
SVTABULATE [PRINT=summary,means; PLOT=*; CLASSIFICATION=livestock;\
  STRATUM=stratum; WEIGHTS=weight]  Y=farmincome; LABELS=farm
```

3.9 Two-stage samples

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/Malawi7.gsh'; ISAVE=mpo
SVTABULATE [PRINT=summary,totals,influence; PLOT=*; SAMPLINGUNITS=EPA; CLASS=ADD;\
  STRATUM=ADD; WEIGHTS=weight; FPCOMIT=yes]  Y=GTIS_hh
" now specifying population sizes "
TABLE [CLASS=ADD; VALUES=27,9,26,32,33,33,14] nEPA
SVTABULATE [PRINT=summary,totals; PLOT=*; SAMPLINGUNITS=EPA; CLASS=ADD;\
  STRATUM=ADD; WEIGHTS=weight; NUNITS=nEPA; FPCOMIT=no]  Y=GTIS_hh
```

4 Weights and imputation
4.1-4.3 Creating and modifying survey weights

```
IMPORT 'C:/Progra~1/Gen16Ed/Data/Juneresponse.gwb'; SHEET='responses'; ISAVE=rpo
IMPORT 'C:/Progra~1/Gen16Ed/Data/Juneresponse.gwb'; SHEET='nfarm'; ISAVE=npo

SVWEIGHT [PRINT=summary,strat,psus; STRATUM=strata; NUNITS=nfarm]
OUTWEIGHTS=weights

" 4.2 practical "
SVTABULATE [PRINT=summary,totals,influence; CLASS=strata; STRATUM=strata; \
  WEIGHTS=weights]  Y=A1_wheat; LABELS=holding

" 4.3 modifying "
SVREWEIGHT [PRINT=summary; METHOD=*; WEIGHTS=weights; OUTWEIGHTS=weightsB; \
  STRATUM=strata; LABELS=holding] berror
```

4.4 Modifying weights for outliers

```
IMPORT 'C:/Progra~1/Gen16Ed/Data/Juneresponse.gwb'; SHEET='responses'; ISAVE=rpo
IMPORT 'C:/Progra~1/Gen16Ed/Data/Juneresponse.gwb'; SHEET='nfarm'; ISAVE=npo

SVWEIGHT [PRINT=summary,strat,psus; STRATUM=strata; NUNITS=nfarm] \
  OUTWEIGHTS=weights

RESTRICT A1_wheat; strata.NI.'new'
SVTABULATE [PRINT=summary,ratios,influence; CLASS=strata; STRATUM=strata; \
  WEIGHTS=weights]  Y=A1_wheat; X=xa1; LABELS=holding
RESTRICT A1_wheat

SVREWEIGHT [PRINT=summary; METHOD=*; WEIGHTS=weights; OUTWEIGHTS=wt_exoutlier; \
  STRATUM=strata; OUTSTRATUM=strat_exoutlier; LABELS=holding] 343460118; NEW=1

RESTRICT A1_wheat; strata.NI.'new'
SVTABULATE [PRINT=summary,ratios; CLASS=strat_exoutlier; STRATUM=strat_exoutlier; \
  WEIGHTS=wt_exoutlier]  Y=A1_wheat; X=xa1; LABELS=holding
RESTRICT A1_wheat
```

4.5 Calibration weighting

```
IMPORT 'C:/Progra~1/Gen16Ed/Data/FBSmult.gwb'; SHEET='crops'; ISAVE=mpo
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo
" check farm numbers match between datasets "
DESCRIBE Farm-farm
" initial analysis "
SVTABULATE [PRINT=summary,totals; STRATUM=stratum; WEIGHTS=uncalibrated_wt] \
  Y=osr; LABELS=holding
SVCALIBRATE [PRINT=summary; WEIGHTS=uncalibrated_wt; OUTWEIGHTS=cal_wt; \
  METHOD=linear; TCONSTRAINTS=61655,463935; X=*,osr; LOWER=0.1; UPPER=10; \
  PLOT=weights]
```

4.6 Calibration by groups

```
IMPORT 'C:/Progra~1/Gen16Ed/Data/FBSmult.gwb'; SHEET='crops'; ISAVE=mpo
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo
" check farm numbers match between datasets "
DESCRIBE Farm-farm
SVCALIBRATE [PRINT=summary; WEIGHTS=uncalibrated_wt; OUTWEIGHTS=cal_wt; \
  METHOD=linear; TCONSTRAINTS=61655,463935; X=*,osr; LOWER=0.1; UPPER=10; PLOT=*]
```

4.7 Practical

```
IMPORT 'C:/Progra~1/Gen16Ed/Data/June_calibration.gwb';sheet='totals'
IMPORT 'C:/Progra~1/Gen16Ed/Data/June_calibration.gwb';sheet='response'

" ratio analysis for comparison "
SVSTRATIFIED [PRINT=summary,totals; METHOD=separate; STRATUM=strata; \
  SAVESUMMARY=no] A1_wheat; X=xa1; LABELS=holding; NUNITS=nhold; \
  XTOTALS=totxa1; TOTALS=totrat ;setot=serat

SVCALIBRATE [PRINT=summary; WEIGHTS=weights; OUTWEIGHTS=calwt; METHOD=linear;\
  TCONSTRAINTS=nhold,totxa1; X=*,xa1; STRATUM=strata] Y=A1_wheat; FITTED=a1fit

SVTABULATE [PRINT=summary,totals; CLASS=strata; STRATUM=strata; WEIGHTS=calwt] \
  Y=A1_wheat; TOTALS=totcal; SETOTALS=secal

SVTABULATE [PRINT=summary,totals; CLASS=strata; STRATUM=strata; WEIGHTS=calwt] \
  Y=A1_wheat; TOTALS=totcalfit; SETOTALS=secalfit; FIT=a1fit

PRINT totrat,totcal,totcalfit,serat,secal,secalfit
```

4.8 Hot-deck imputation for missing values

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo

SVHOTDECK [PRINT=summary,list; METHOD=hotdeck; DMETHOD=minimax; SEED=0] \
  subsidy20mv; NEWSTRUCTURE=random

CALCULATE absfarmincome=ABS(farmincome)

SVHOTDECK [PRINT=summary,list; METHOD=hotdeck; DMETHOD=minimax; SEED=0;\
  DVARIABLES=type,absfarmincome; DRANGES=*,*] subsidy20mv; NEWSTRUCTURE=nearest; \
  OVERWRITE=no

" and imputing 100 at random to check "
SVHOTDECK [PRINT=summary,check,monitoring; METHOD=hotdeck; DMETHOD=minimax;\
  SEED=0; DVARIABLES=type,absfarmincome; DRANGES=*,*; IMPUTE=100] subsidy20mv
```

4.9 Model-based imputation for missing values

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo

" fit model with separate slopes for each farm type "
MODEL subsidy20mv; RESIDUALS=res; FITTED=fits
FIT [PRINT=model,summary,estimates; CONSTANT=estimate; FPROB=yes; TPROB=yes] \
  type*absfarmincome
" check residuals "
RCHECK [RMETHOD=deviance; GRAPHICS=high] residual; composite
" plot relationships "
RGRAPH [GRAPHICS=high]
" then use to form imputed values, taking residual at random from within farm type"
SVHOTDECK [PRINT=summary,list; METHOD=modelbased; DMETHOD=minimax; SEED=0;\
  DVARIABLES=type; DRANGES=*] subsidy20mv; NEWSTRUCTURE=regression; OVERWRITE=no

" alternative method: this takes an observation at random from those with fitted
  values (see MODEL statement above) within 100 of the nearest fit. Note that
  THRESHOLD is set to -100 (a negative distance indicating it is an absolute value)
  and DRANGES is set to 1, to prevent any scaling "
SVHOTDECK [PRINT=summary,list,monitoring; METHOD=hotdeck; DMETHOD=minimax; SEED=0;\
  DVARIABLES=fits; DRANGES=1; THRESHOLD=-100] subsidy20mv; NEWSTRUCTURE=regfit;\
  OVERWRITE=no
```

5 Progamming Genstat for surveys

Since the main chapter lists the commands for most sections, only the practicals are shown here.

5.2 Practical

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo
SVTABULATE [PRINT=summary,totals,influence,psusummary,wald; CLASS=sex; \
  STRATUM=stratum; WEIGHTS=weight] Y=farmincome; LABELS=farm; TOTALS=total;\
  SETOTAL=se_total; WALD=test_stats
```

5.4 Practical

```
SPLOAD 'C:/Progra~1/Gen16Ed/Data/FBS_England_merged.gsh'; ISAVE=fpo
FOR d=sex,type,tenancy ;mtab= meansex,meantype,meantenancy
  SVTABULATE [PRINT=summary,means,influence; CLASS=d; STRATUM=stratum;\
    WEIGHTS=weight; NINFLUENCE=10; FPCOMIT=no]  Y=farmincome; LABELS=farm; \
    MEANS=mtab
ENDFOR
```

6 Survey design and sampling

6.1 Selecting random samples

```
SPLOAD '%GENDIR%/Data/Junemod.gsh'; ISAVE=jpo
SET [SEED=6510]
SVSAMPLE [PRINT=summary; SAMPLE=sampno; NUNITS=19156; NSAMPLE=0.1; METHOD=sample;\
  NUMBERING=population] OLDVECTOR=holding; NEWVECTOR=sampled_holding
FSPREADSHEET  sampno,sampled_holding
```

6.2 Selecting stratified random samples

```
SPLOAD '%GENDIR%/Data/Junemod.gsh'; ISAVE=jpo
SET [SEED=6510]
"Survey Sampling"
TABLE [CLASS=strata; VALUES=100,200,500,500,500] nsample; DECIMALS=0
SVSAMPLE [PRINT=summary; NSAMPLE=nsample; METHOD=sample; NUMBERING=population]\
 OLDVECTOR=holding,parish,xa1,xa10,strata; NEWVECTOR=Holding,Parish,Xa1,Xa10,Strata

job 'structures not defined'
TEXT [VALUES=new,small,medium,large,'very large'] Strata
VARIATE npop,nsamp; VALUES=!(2613,5851,5479,3074,2139),!(100,200,3(500))
SVSAMPLE [PRINT=sum; STRATUMFACTOR=STRATUM; SFLAB=Strata; NUNITS=npop;\
  NSAMPLE=nsamp; SEED=5642; METHOD=pop; SAMPLE=SAMPLED]
FSPREAD STRATUM,SAMPLED
"use tabulate to check"
TABULATE [PRINT=nob,total,mean; CLASS=STRATUM; MARGIN=yes] SAMPLED
```

6.3 Cluster and multistage samples

```
SPLOAD '%GENDIR%/Data/Junemod.gsh'; ISAVE=jpo
SET [SEED=6510]
SVSAMPLE [PRINT=summary; SAMPLE=stage1; NUNITS=19156; NSAMPLE=0.1;\
  METHOD=population; NUMBERING=population; CLUSTER=parish]

TABULATE [PRINT=*; CLASSIFICATION=parish; MARGINS=no] stage1; NOBS=tnobs;\
  MEANS=tstage1
CALC psample2=tstage1*0.4
CALCULATE nsample2=CEILING(psample2*tnobs)
"alternatively this sets proportions to 0.99 when tnobs equals 1"
CALC psample2b=tstage1*(0.4+0.59*(tnobs.EQ.1))
FSPREAD tnobs,tstage1,psample2,nsample2,psample2b

SVSAMPLE [PRINT=summary; SAMPLE=stage2; NSAMPLE=nsample2; METHOD=population;\
  NUMBERING=population] parish,holding; NEWVECTOR=Holding,Parish
FSPREAD holding,parish,stage1,stage2
```

7 Regression for surveys
7.2 Linear regression for surveys

```
SPLOAD 'FBS_Regression.gsh'; ISAVE=fpo

XAXIS 1;MARK=1000
DGRAPH [WINDOW=5;KEYWINDOW=0;TITLE='subsidy v farmarea'] subsidy; farmarea

YAXIS 3;TRANSFORM=log10
XAXIS 3;TRANSFORM=log10;MARK=!(1,10,100,1000)
DGRAPH [WINDOW=3;KEYWINDOW=0;TITLE='subsidy v farmarea (log scale)'; \
       SCREEN=keep] subsidy+1; farmarea; PEN=type

CALC logsubsidy=LOG10(subsidy+1)
CALC logfarmarea=LOG10(farmarea)

RESTRICT logsubsidy;CONDITION=type.ni.!t(Pigs,Poultry,Horticulture)
SVGLM [PRINT=model,estimates,wald,pred; DISTRIBUTION=normal; LINK=identity; \
 TERMS=logfarmarea; WEIGHTS=weight; CIPROB=0.95; PFACTOR=logfarmarea; \
     PLEVELS=!(1,1.5...3)] logsubsidy;PRED=pr;LOWPRED=lpr;UPPRED=upr
```

7.3 Generalized linear models for surveys

```
SPLOAD 'FBS_Regression.gsh';ISAVE=ipo

CALC zerosubs=subsidy.EQ.0
CALC logfarmarea=LOG10(farmarea)

RESTRICT zerosubs,logfarmarea;CONDITION=type.in.!t(Pigs,Poultry,Horticulture)

SVGLM [PRINT=model,estimates,wald,pred; DISTRIBUTION=binomial; LINK=logit;
FACTORIAL=9;\
 CONSTANT=estimate; DISPERSION=*; TERMS=logfarmarea+type_pph;
STRATUMFACTOR=mergedstratum;\
 WEIGHTS=weight; METHOD=simple; NBOOT=200; SEED=0; CIPROB=0.95;
PFACTORS=logfarmarea,type_pph;\
 PLEVELS=!(0.5,1...2),*; PTERM=logfarmarea,type_pph; SEED=742002] zerosubs;
NBINOMIAL=1
RESTRICT zerosubs,logfarmarea
```

7.4 Fitting unweighted models

```
SPLOAD 'FBS_Regression.gsh';ISAVE=ipo

CALC zerosubs=subsidy.EQ.0
CALC logfarmarea=LOG10(farmarea)

RESTRICT zerosubs,logfarmarea;CONDITION=type.in.!t(Pigs,Poultry,Horticulture)
MODEL [DISTRIBUTION=binomial; LINK=logit; DISPERSION=1] zerosubs; NBINOMIAL=1
FIT [PRINT=model,summary,estimates; CONSTANT=estimate; FPROB=yes; TPROB=yes; \
       FACT=9] logfarmarea
```

```
ADD [PRINT=acc;FPROB=yes] type_pph
PREDICT [PRINT=description,predictions,se; COMBINATIONS=estimable; \
       BACKTRANSFORM=link; ADJUST=marginal] type_pph; LEVELS=*
```

7.6 Practical

```
SPLOAD 'FBS_Regression.gsh';ISAVE=ipo

SVTABULATE [PRINT=summary,means,influence,wald; CLASS=type; STRATUM=mergedstratum;\
   WEIGHTS=weight]  Y=farmincome; LABELS=farm
SVGLM [PRINT=model,estimates,wald,predictions; TERMS=type; \
   STRATUMFACTOR=mergedstratum; WEIGHTS=weight; PFACTORS=type] farmincome
```