Department for Work & Pensions

Uncertainty in Family Resources Survey-based analysis

June 2014

Contents

Summary

1. Due to the complex nature of the Family Resources Survey (FRS) sample structure, estimating the uncertainty around FRS estimates is in itself a complex topic. The methodology section of the FRS annual report outlines the issues involved and provides tablulations of confidence intervals for selected variables.[1]
2. This paper outlines a number of methods which can be applied to estimate uncertainty around FRS estimates more generally, including practical advice and SAS code for derivations. It is aimed at a technical audience - those using the FRS and related datasets to conduct analysis.
3. The methods described are (or will be) used by Department of Work and Pensions (DWP) analysts to estimate uncertainty in a range of FRS-based publications.
4. The first method is an approximate technique that treats the FRS as if the sampling were done by a simple random sample and then makes an approximate modification to correct for the fact that the FRS actually uses a more complex sampling strategy. Its use is appropriate when calculating numbers or percentages of people, benefit units, or households, and when approximate results are acceptable.
5. The second method is a more complex technique that takes the design of the FRS sample directly into account. Its use is appropriate for linear estimates (such as mean values), but not more complex measures such as medians.
6. The third method is bootstrapping. Its use is appropriate even for relatively complex estimates. It is used to estimate confidence intervals estimates of the number of households below the median income in the 2012/13 Households Below Average Income annual publication[2]. However, it can be time consuming to produce estimates using this method.
7. Since the first two techniques include assumptions that are not valid for very small sample sizes, a technique for presenting uncertainty around estimates based on small sample sizes is discussed. A method for handling the case where there are no cases in the sample is also discussed.

[1] The FRS annual report is available here:

https://www.gov.uk/government/collections/family-resources-survey--2#documents

[2] The HBAI annual report is available here:

https://www.gov.uk/government/collections/households-below-average-income-hbai--2#documents

Introduction

8. The Family Resources Survey (FRS) is an annual survey of around 20,000 households across the UK. As with all surveys, there is some uncertainty around estimates derived from it[3]. This follows from the fact that not every household in the country is interviewed as part of the survey, and there is a possibility that those households that *are* interviewed contain a greater fraction of (for example) households below the poverty line than the general population.

9. There are mathematical approaches that can be used to calculate and describe this uncertainty.

10. The relatively complex design of the FRS sample and (for some indicators such as poverty levels) the relatively complex nature of the estimates made can be beyond the standard approaches. The FRS team have investigated methods of measuring uncertainty in estimates derived from the FRS. There are several options, of varying degrees of complexity, which are appropriate for different circumstances.

11. The Households Below Average Income (HBAI) series is derived from the FRS and, therefore, its measures are also subject to sampling uncertainty.

12. This document sets out the methodologies employed and the circumstances under which their use is appropriate.

The design of the FRS sample

13. The simplest possible sample design is a “simple random sample”. In this case, 20,000 households would be chosen at random from a list of all households in the UK. This is not a practical sample design for a national survey in a country as large as the UK, since the costs involved in moving interviewers about the country to match the random sample distribution would be prohibitive. Therefore, the FRS uses a more complex design.

14. Full details of the sample design are set out in the *Methodology* chapter of the FRS annual publication[4]. In brief, however, the FRS uses a stratified clustered sample design[5]. The *Primary Sampling Units* (PSU) are the postcode sectors (the first part of the postcode and the first digit of the second part – e.g. LS4 7). Each PSU is assigned to a stratum, and from each stratum a random sample of PSUs is drawn. Thus, not every postcode sector is used in each sample year. Within each selected PSU, a random sample of addresses is drawn and visited.

15. This sample structure is illustrated schematically in Figure 1, on the next page.

[3] Almost any statistics text book will cover the topic; for example Bhattacharyya, G.K. and Johnson, R.A., *Statistical Concepts and Methods*, Wiley, 1977.

[4] https://www.gov.uk/government/collections/family-resources-survey--2

[5] Chapter 10 in Cochran, W. G., *Sampling Techniques* 3rd Ed., Wiley, 1977.

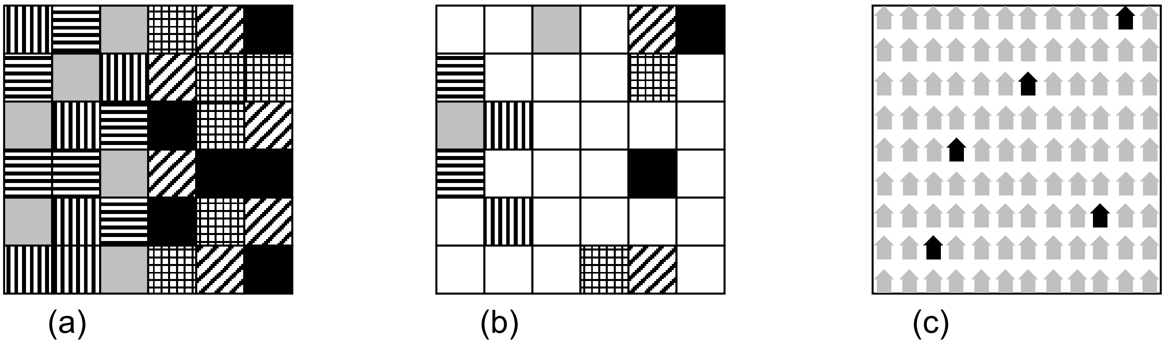


Figure 1: A schematic representation of the FRS sampling. (a) The PSUs (postcode sectors) are represented by the small squares. They are assigned to strata, represented by the shading. (b) A sample of PSUs is drawn at random from each stratum (non-shaded PSUs are not visited). (c) A single PSU is shown, containing many households. A random sample of those (shown darker) are selected to be interviewed.

16. One half of the PSUs are retained for the next year (although fresh households are drawn within them) and the other half of the PSUs are discarded and fresh PSUs are drawn.

17. This design tends to result in estimates that have slightly wider confidence intervals than would be expected from a simple random sample. This decrease in precision is necessary in order to make it possible to carry out the survey without excessive cost.

Weighting of the FRS

18. The FRS includes a weighting, or grossing, factor, which is an estimate of the number of households in the population that are represented by a given household in the sample. For example, a household in the survey containing a couple and two children which is assigned a weight of 1,000 has been estimated to represent 1,000 such households in the country. The CALMAR[6] algorithm is used to calibrate the FRS weights such that the population totals derived from the FRS match various totals derived from various sources including the census and HMRC data. Details can be found in the *Methodology* chapter of the FRS annual publication[7].

19. The methodology for calculating the grossing factors has changed for the 2012/13 FRS dataset. The details of the changes are available in *FRS grossing methodology review and 2011 Census updates*, published by the DWP[8].

[6] CALMAR implements algorithms described in Deville, J-C and Sarndal, C-E, *Calibration Estimators in Survey Sampling* Journal of the American Statistical Association Vol. 87, No 418 (Jun 1992) pp 376-382. Full text available at http://www.jstor.org/stable/2290268

[7] https://www.gov.uk/government/collections/family-resources-survey--2

[8] https://www.gov.uk/government/collections/family-resources-survey--2

FRS datasets

20. The FRS dataset is released annually. There are two versions of each dataset available, the End User License (EUL) dataset and the Secure Access File (SAF). Both datasets are anonymous, but the EUL dataset (to which it is easier for researchers to obtain access) has had some further rounding of cash values and suppression of potentially disclosive variables. Some of the analysis discussed in the following sections can only be performed with access to the SAF dataset, which includes enough information to identify the FRS sampling structure.

Methods of estimating uncertainty in the FRS

21. We have experimented with three different methods of estimating uncertainty in the FRS. These are:

    - Modified simple random sample
    - Variance estimation based on sample design
    - Bootstrapping

22. The methods all have strengths and weaknesses, and are appropriate for use in different circumstances. They are discussed in the following sections.

23. Additionally, we have considered appropriate ways to report uncertainty in FRS-based analysis, and how to handle the situation where there are no cases with a certain characteristic in the sample. These cases are also discussed below.

Modified simple random sample

24. As noted earlier, the FRS is not a simple random sample. However, it would be quite straightforward to calculate an estimate of the uncertainty in population prevalences (such as an estimate of the number of people with long term health problems in the UK) if it were a simple random sample. This method is based on calculating such estimates, then making approximate adjustments for the additional uncertainty introduced by the more complex sample design.

25. This method is appropriate where:

    - You are calculating an estimate of the prevalence of a characteristic in the population; and
    - There is at least one case in the dataset that has this characteristic (however, see the section on small sample sizes); and
    - You only need an approximate estimate of the uncertainty.

26. This method does not require access to the SAF FRS dataset because it does not use the detailed sample structure information that is in the SAF dataset and not the EUL one.

27. Treating the FRS as a simple random sample of households, the best estimate of the prevalence of a characteristic in the population is

$$\hat{p} = \frac{\sum_{i \in C} g_i}{\sum_{\forall i} g_i} \qquad (1)$$

where C is the set of cases with the characteristic and g_i is the grossing factor associated with the ith case. That is, the numerator is the sum of the grossing factors for all cases with the characteristic, and the denominator is the sum of the grossing factors for all cases.

28. The Agresti-Coull[9] estimate of the confidence limits associated with this is

$$CI = \tilde{p} \pm 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} \qquad (2)$$

where n is the number of cases in the sample, and $\tilde{p}$ is given by

$$\tilde{p} = \frac{2 + \sum_{i \in C} g_i}{4 + \sum_{\forall i} g_i} \qquad (3)$$

29. Note that this formula is an approximation, and only valid for calculating 95% confidence limits.

30. A more complex sample structure, such as that used in the FRS, produces a different standard error which may be higher or lower than the standard error calculated above. This is expressed in the design factor. The design factor is the ratio between the obtained standard error and the standard error that would have resulted from a simple random sample of the same size

$$D = \frac{\sigma_{actual}}{\hat{\sigma}} \qquad (4)$$

31. where σ_{actual} is the standard error from the FRS calculated by a method that takes into account the complex sample structure.

32. Detailed calculations (see the next section) suggest that typical design factors for the FRS vary from around 1.1 to 1.3. The SE series of tables in the annual FRS publication[10] provide some examples in specific cases.

33. This suggests a method for approximating the standard error in population prevalences:

    - Calculate an estimate of the population prevalence (eq 1).
    - Calculate the confidence limits associated with this estimate (eq 2).

[9] Agresti, A and Coull, B. A. (1998). "Approximate is better than 'exact' for interval estimation of binomial proportions". The American Statistician 52: 119–126

[10] https://www.gov.uk/government/collections/family-resources-survey--2

- Widen the confidence limits by an appropriate design factor (see the SE tables in the FRS publication, or use 1.3 which appears to be slightly wider than is typical).

34. This method is approximate and confidence limits derived using it should be treated as indicative. Design factors as low as 0.8 and as high as 3.0 have been observed in the FRS, so estimates produced with this methodology rely on the professional judgement of the producer for their accuracy. If two estimates are separated by many times the “un-broadened” confidence interval, one could confidently state that they were significantly different. If they are separated by less than the un-broadened confidence interval, one could confidently state that they were not significantly different. However, there remains a grey area between these two extremes where more complex calculation would be necessary.

Variance estimation based on sample design

35. It is possible to take the design of the FRS sample into account when calculating standard errors in a more rigorous way than simply multiplying the standard error by an assumed design factor. This method is used to generate the SE tables in the FRS publication.

36. This method is appropriate when:

    - There is at least one non-zero case in the dataset (however, see the section on small sample sizes); and
    - You are calculating linear estimates from the data, such as mean incomes or the prevalence of a characteristic in the population. It cannot be used for non-linear estimates such as median incomes or the poverty rates derived from them.

37. This method requires access to the SAF dataset.

38. This method was developed by the Office for National Statistics (ONS) Methodology team. It uses the SAS/STAT[11] PROC SURVEYFREQ and PROC SURVEYMEANS statements. These are built-in analysis functions of the SAS package, and are based on Taylor expansion methods for variance estimation[12,13].

39. The SAF dataset includes PSU and stratum identifiers (Annex 1 contains details). Having derived these, the SAS/STAT PROCs named above can be used to calculate standard errors and confidence limits which take into account the sample design.

[11] http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_surveyfreq_a0000000212.htm

[12] Woodruff, R.S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," *Journal of the American Statistical Association,* 66, 411 - 414

[13] Fuller, W.A. (1975), "Regression Analysis for Sample Survey," *Sankhya*, 37, Series C, Pt. 3, 117 - 132

40. To calculate the confidence limits around prevalence estimates (such as estimates of the number of disabled people in the country), use PROC SURVEYFREQ. To calculate the confidence limits around mean values of continuous variables (such as household income), use PROC SURVEYMEANS. Example code is given in Annex 2.

41. Note that this analysis does not take account of non-linear behaviour. For example, this method could be used to determine confidence limits around the number of people in households with income below £500/week. It could not be used to determine the number of people in households with income below 60% of the median income, since it does not take account of the fact that the median income is itself an estimate from the survey and therefore has its own uncertainty.

Small sample sizes

42. Both of the above techniques described in the last two sections can produce implausible results when only a small number of cases with the characteristic of interest are present in the sample. This is because both methods produce symmetric confidence limits, which is an approximation that is not appropriate when the number of cases is small since the actual distribution is asymmetric.

43. In this case, one option is to use the bootstrapping methodology, described in the next section. The other option is to define some threshold number of cases below which it is inappropriate to report central estimates and confidence limits, and to report the upper confidence limit and nothing else (called a one-tailed confidence limit).

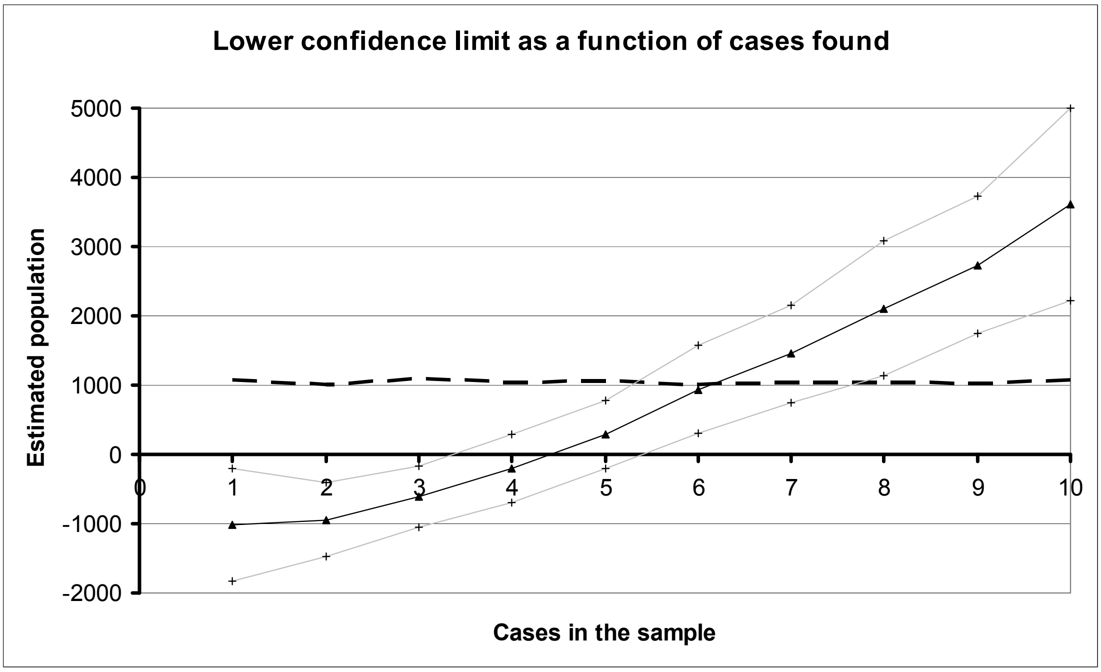
44. We considered two choices for this threshold:

    - Where the lower confidence limit, calculated with the PROC SURVEYFREQ method, implies a population of one thousand households or less (this method requires access to the SAF version of the FRS dataset).
    - Where there are ten cases in the survey (this method does not require access to the SAF version of the FRS dataset).

45. Given typical FRS grossing factors, the former is approximately equivalent to the condition that the lower confidence interval be greater than one case in the survey.

46. The second condition is a rule-of-thumb approximating the first, and is useful if you do not have access to the SAF dataset with its PSU and cluster indicators. It is derived from a simulation, where a pre-determined small number of cases were chosen at random and flagged as being members of a sub-population. Repeatedly flagging random cases and calculating the resulting lower confidence limit builds up a distribution of lower confidence limits. Varying the number of cases flagged reveals that around eight cases is sufficient to give reasonable confidence that any combination of eight cases will satisfy the first condition. Ten cases is slightly more conservative.

47. The graph below shows the mean (dark line) and symmetric 95% confidence limits (grey lines, at 1.96 times the standard deviation either side of the mean line) of the distribution of the lower confidence interval [14] as a function of the number of cases in the dataset, based on 100 repetitions. The typical estimated population represented by one case in the dataset is also shown (broken line).



Bootstrapping

48. Bootstrapping[15] is a method of estimating uncertainty through a process of resampling the dataset.

49. This method is appropriate when:

- There is at least one case with the given characteristic, or a non-zero cash value, in the dataset.

50. This method does not require access to the SAF FRS dataset.

51. Bootstrapping is quite straightforward in principle. In practice it can take minutes to hours for a computer to calculate a bootstrap estimate, and keeping track of the households and other sampling units can present a programming challenge with complex sample designs. For a simple random sample (e.g. an equi-probable sample of all addresses in the country), the process is:

- Resample the households – that is, draw households (with replacement) from the initially drawn sample
- Recalculate statistics of interest based on this resampled dataset
- Repeat these two steps a large number of times

[14] The grey lines are confidence limits on a confidence limit.

[15] Efron, B. and Tibshirani, R.J. *An Introduction to the Bootstrap*, CRC Press 1994.

- Calculate appropriate percentiles of the distribution of the statistics of interest, and use these as confidence limits.

52. This process is illustrated in figure 2, below.

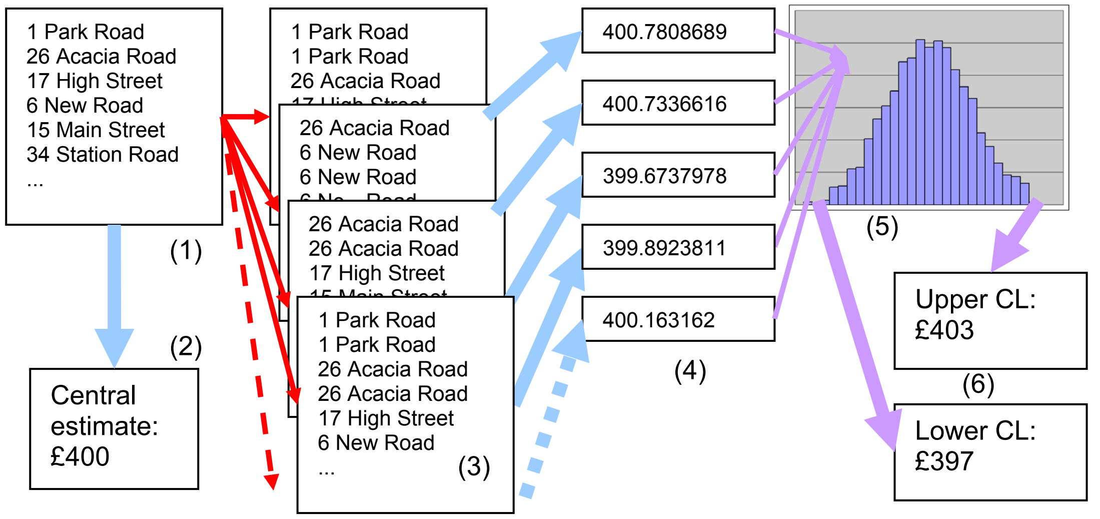


Figure 2: Bootstrapping. A sample of addresses (many more than the six shown) with associated data (1) is analysed to produce an estimate of median household income (2). The dataset is resampled to produce multiple datasets (many more than the four shown, and each the same size as the original) containing zero, one or many copies of each household in the original (3). The median household income is calculated from each resample (4), producing a distribution of values (5). The 2.5th and 97.5th percentiles of this distribution are the estimate of the 95% confidence limits (6). Please note that the numbers in this diagram are indicative and are not derived from actual data.

53. Since the FRS is a household survey, the appropriate process is to resample households from the dataset and merge on any lower-level information (such as information about individuals within the household) as appropriate. Some care must be taken in implementing the merge, since merging multiple copies of a household with data on multiple individuals in the household is a many-to-many match. Statistical packages do not all handle many-to-many matches in the same way, and not all of these ways are appropriate for this application.

54. In SAS, resampling is carried out using the PROC SURVEYSELECT command. See Annex 3 for example code.

55. This is the methodology used to generate the confidence intervals in the 2012/13 HBAI report[16]. The methodology is more flexible than that used in earlier reports, and it is easier to generate confidence limits for new variables with this method.

56. Note that, for reasons discussed in the next section, we believe that this methodology under-estimates the width of confidence intervals, which is to say, over-stating the precision of the measurement. Nevertheless, the method is flexible and appropriate for use with non-linear estimates such as poverty rates.

[16] https://www.gov.uk/government/collections/households-below-average-income-hbai--2

57. The primary advantage of the bootstrap is that it can be used to generate confidence intervals around non-linear estimates such as median incomes or poverty levels where the simpler methods cannot be used. It also naturally produces asymmetric confidence intervals where these are appropriate, including in the case of small sample sizes and income-related measures.

Future development in bootstrapping the FRS

58. The methodology described in the preceding section is perfectly suited to a survey based on a simple random sample, but there is room for improvement in the case of a more complex design, such as the stratified, clustered sample design of the FRS.

59. Bootstrapping in stratified surveys is only slightly more complicated than for a simple random sample. In such surveys, the population is divided into several strata, either for administrative convenience (geographical regions, for example) or to improve statistical efficiency (where there are expected to be different variances associated with different groups), and an independent sampling exercise is carried out in each stratum. In a sense, a stratified survey is a collection of independent surveys of spanning, non-overlapping sub-populations. The idea of bootstrapping is to simulate the process of sampling that generated the original dataset so independent resampling exercises must be carried out in each stratum.

60. The situation is rather more complicated for two-stage cluster samples, like the FRS. In a cluster sample, the sample is subdivided into groups called Primary Sampling Units (PSU), and a random sample of these is drawn. Within PSUs that were drawn, a random sample of households is drawn. This is distinct from a stratified sample, in that in the case of a stratified sample, households are visited from every stratum, whereas no households are visited in PSUs that were not drawn.

61. Thus, for a clustered sample it is necessary to resample PSUs (resulting in multiple copies of some PSUs and none of other PSUs), and then resample the households that were found in those PSUs. The resampling is carried out independently in each PSU – so a resample containing two copies of a particular PSU may contain no copies of a particular household in one of the PSU copies and several in the other.

62. This two-stage resampling process can be complex to implement, and there are some subtleties that lead to bias in the results unless they are handled correctly[17]. We have not yet been able to produce programs to produce confidence limits based on this more complex methodology, although work continues. For the time being, then, we recommend that a simple random sample of households should be used. The method is flexible enough to produce

[17] For example, bias discussed in http://www.stata-journal.com/sjpdf.html?articlenum=st0187

confidence limits for the more complex measures that are derived from the FRS and HBAI and it is relatively straightforward to implement new measures.

Multi-year estimates

63. Sometimes it is necessary to combine multiple years of FRS data in order to get a large enough sample size to generate a reliable number (for example, the ethnic group analysis in the FRS publication). Changes in values between years can also be of interest.

64. When consecutive years are in use in a calculation, the complexity of the FRS sample design must be taken into account. As noted in the section on the sample design, half of the PSUs in one sample are retained from the previous sample. This means that the samples are not completely independent (because random choices of PSUs in one sample affect the choices of PSUs in the second sample), which has an effect on the confidence intervals which must be accounted for.

65. This is not difficult in principle, and the ONS Methodology Team do produce confidence limits around a three-year average for the annual FRS publication (table SE.7). However, it is not possible to produce ad hoc analysis on this basis. The FRS dataset does not include sufficient information to do so, and the additional information needed (postcodes) is potentially disclosive.

66. Ways to work around this are under investigation. For the time being, the recommended approach is to treat successive datasets as independent random samples, and caveat any uncertainty calculations with a note to the effect that the samples are not independent, and this has not been accounted for.

67. When only non-consecutive years are under consideration (for example, the change in poverty level between 2010/11 and 2012/13), the samples are independent and the sample design does not have any special effect.

No cases in the sample

68. It is possible that no cases with a particular characteristic will be found in the FRS sample. In that case, the best estimate from the sample is that there are no cases with this characteristic in the population, and all of the techniques described above will produce a zero-width confidence interval. This is overstating our certainty, however, and information from other sources may show the statement to be outright wrong. In fact, finding no cases in the FRS implies that the characteristic is rare, not necessarily that it is non-existent (although it may be).

69. Several sources[18] cite Hanley and Lippman-Hand[19] as the source of a rule that for simple random samples where zero cases with a particular characteristic are

[18] Such as Eypasch, E., Lefering, R., Kum, C.K., Troidl, H *Probability of adverse events that have not yet occurred: a statistical reminder* BMJ 1995;311:619. Article available online at http://www.bmj.com/content/311/7005/619.full.

found, the upper 95% confidence limit for the population prevalence should be 3/*n*, where *n* is the sample size. The justification for this is that, if the prevalence of some characteristic in a population is *p* then the probability of drawing zero cases in n samples is

$$(1-p)^n \tag{5}$$

70. If this is set equal to 0.05, then p becomes the prevalence that yields a 95% chance that we would get one or more cases with a sample of size n. The 3/*n* approximation then follows, using the approximations ln(0.05)≈-3 and ln(1-*x*)≈-*x* for small *x*.

71. Since the FRS is not a simple random sample of households, this rule does not precisely apply. However, an approximation can be found by inflating this limit by the appropriate design factor. Noting that the design factors in the FRS seldom exceed 1.3, this suggests a (probably conservative) FRS-specific rule of 4/n for the upper confidence limit on the proportion of households with a given characteristic.

72. In the 2010/11 FRS there are *n*=25,356 households, representing a UK population of 26,327,621 households[20]. If the dataset contains no households with a given characteristic, then 4/*n*=0.016%, which is approximately 4,200 (≈0.016%×26,327,621) households. This could be reported in these terms:

    Given that no cases were found in the sample, we are 95% confident that the number of households in the UK with the specified characteristic lies between 0 and 4,200, or between 0 and 0.016% of all households.

73. The FRS is a household survey, so lower-level units such as benefit units[21] (of which there can be several in a household) should be treated with care. In this case, the *n* should again be the number of households in the survey, so 4/*n* would be the prevalence of households with one or more benefit units with the given characteristic. This would then be multiplied by the number of benefit units to get the final number. This is a slight underestimate since it takes no account of the variance in the number of benefit units in a household, but since 4/*n* is likely to be an overestimate of the impact of the design effects this extra variance has been ignored.

74. In the 2010/11 FRS there are 32,909,354 benefit units. Using the number of households to estimate the confidence limit for the prevalence yields 4/*n*=0.016%, which is approximately 5,200 benefit units. This could be reported in these terms:

    Given that no cases were found in the sample, we are 95% confident that the number of benefit units in the UK with the specified characteristic lies between 0 and 5,200, or between 0 and 0.016% of all benefit units.

[19] Hanley, J.A., Lippman-Hand, A, *If Nothing Goes Wrong, Is Everything All Right?: Interpreting Zero Numerators* JAMA. 1983;249(13):1743-1745. Abstract free online at http://jama.jamanetwork.com/article.aspx?articleid=385438

[20] Using the revised grossing regime

[21] A benefit unit is a single person or couple living as married, and any dependent children.

75. The same argument applies to adults, children and any other sub-unit of a household. The prevalence would be calculated based on the number of households, but be multiplied by the total number of sub-units.

76. The number of households, n, should be the number of households in the sampling frame. For example, if a figure is being derived for England only then (in 2010/11) *n* would be 18,160, the number of households surveyed in England.

Conclusion

77. Determining the uncertainty around estimates derived from the Family Resources Survey is a complex topic. This paper has presented three different methodologies, appropriate in different circumstances, and some work for future development. Currently, we believe that the variance estimation based on sample design provides the best estimates where linear estimators are required. For non-linear estimators, the bootstrapping methodology is superior. The main advantage of the modified simple random sample methodology is its technical simplicity, making it appropriate to use (with caution) when speed is required. It is also appropriate where an approximate answer is acceptable, such as where the difference between two estimates is very much larger or very much smaller than the estimates' uncertainties.

Dataset	Measure	Recommended technique
EUL	Prevalence / count	Modified simple random sample
	Mean income / median / non-linear measure	Bootstrapping
SAF	Prevalence / count / mean income	Variance estimation based on sample design
	Median / non-linear measure	Bootstrapping

Annex 1 – Deriving PSU and stratum identifiers from the FRS dataset

78. The following SAS code can be used with the SAF dataset to derive cluster and PSU identifiers. The final output is a dataset, `householdPSUIdentifiers`, containing each household identifier and its associated PSU and cluster identifier in variables named `sernum`, `psu` and `psu_pair_number`, respectively.

79. The program will not produce meaningful results if applied to the EUL dataset. In the EUL dataset, `sernum` (which contains the PSU information) is replaced with a number that is only a household identifier, without such relatively low-level geographic information.

```
%* Derive the PSU identifier and region (former GOR) in GB and sample
month in NI;
data  psu_gb(keep=sernum region psu)
            psu_ni(keep=sernum month psu);
      set frs.househol(keep=sernum);
            by sernum;
      region=floor(sernum/100000000);
      psu=floor(sernum/100000);
      if region=50 then do;   * Northern Ireland case;
            month=floor(sernum/10)-100*floor(sernum/1000);
            output psu_ni;
      end; else do;  * Great Britain case;
            output psu_gb;
      end;
run;
%* Pair up the GB PSUs to identify clusters (PSU pairs)...;
proc sort data=psu_gb;
      by region psu;

      set psu_gb(keep=region psu);
            by region psu;
      if last.psu;
data psu_pair_gb;
      set psu_pair_gb;
            by region psu;
      retain x;
      if first.region then x=0;
      x+1;
      psu_pair_number=region*1000+(x+mod(x,2))/2;

1;
run;
%* ...and combine cluster and PSU identifiers. ;
data psu_gb;
      merge       psu_gb
                  psu_pair_gb(keep=region psu psu_pair_number);
      by region psu;
run;
%* Pair up the NI PSUs to identify clusters (PSU pairs)...;
proc sort data=psu_ni;
      by month psu;
data psu_pair_ni;
```

```
      set psu_ni(keep=month psu);
      by month psu;
      if last.psu;
data psu_pair_ni;
      set psu_pair_ni;
      by month psu;
      retain x;
      if first.month then x=0;
      x+1;
      psu_pair_number=5000000+month*1000+(x+mod(x,2))/2;

run;
%* ...and combine cluster and PSU identifiers. ;
data psu_ni;
      merge       psu_ni
                  psu_pair_ni(keep=month psu psu_pair_number);
      by month psu;
run;
%* Finally, append the NI information to the GB information. ;
data householdPSUIdentifiers;
      set psu_gb(keep=sernum psu psu_pair_number)
            psu_ni(keep=sernum psu psu_pair_number);
proc sort data=householdPSUIdentifiers;
      by sernum;
run;
```

Annex 2 – Example code illustrating the use of PROC SURVEYMEANS and PROC SURVEYFREQ

80. There are two related PROCs in SAS for analysing uncertainty around categorical and continuous data, PROC SURVEYFREQ and PROC SURVEYMEANS, respectively. Before using them, it can be useful to adjust the number of significant figures that they output. This is done using PROC TEMPLATE. The following example sets the output to fifteen characters, with up to ten decimal places for the percentage, standard error and design effect fields. It need only be run once per SAS session. See the SAS documentation on the ODS TRACE statement for how to find the names for other fields.

```
proc template;
      edit Stat.SurveyFreq.OneWayFreqs;
            edit Percent;
                  format = 15.10;
            end;
            edit STDErr;
                  format = 15.10;
            end;
            edit DesignEffect;
                  format = 15.10;
            end;
      end;
      edit Stat.SurveyMeans.Ratio;
            edit Ratio;
                  format = 15.10;
            end;
            edit STDErr;
                  format = 15.10;
            end;
            edit Var;
                  format = 15.10;
            end;
      end;
run;
```

81. An example of the use of PROC SURVEYFREQ is below. The input dataset is specified in the `data=` statement. The variables of interest are listed in the `tables` statement – in this case, the output would be the number of households split by the number of dependent children, and the uncertainty associated with each one. The keywords after the slash request the design effect, and confidence limits around the percentage of households and the weighted number of households. The stratum and cluster variables, `psu_pair_number` and `psu` respectively, should be the ones produced by the program in Annex 1, and the weight variable is the appropriate FRS grossing factor.

```
proc surveyfreq   data=househol;
      tables      depchldh / deff cl clwt;
      strata      psu_pair_number;
      cluster     psu;
      weight      gross4;
run;
```

82. An example of the use of PROC SURVEYMEANS is below. The `data=`, `strata`, `cluster`, and `weight` statements function as they do in PROC SURVEYFREQ. The statistics of interest, however, are listed in the first line – this example asks for ratios, their standard errors, and confidence limits around the means. The variables for analysis are listed in the `var` statement. Since a ratio analysis was requested, the ratio statement specifies that the variable listed before the slash should be expressed as a fraction of the variable after it. This case calculates the average fraction of household income that comes from earnings, and its associated uncertainty.

```
proc surveymeans data=hh ratio stderr clm;
      var         hearns hhinc;
      ratio       hearns / hhinc;
      strata      psu_pair_number;
      cluster     psu;
      weight      gross4;
run;
```

83. Note that these PROCs do not produce the design factor, but rather the design effect. This is the ratio of the actual variance to the variance obtained from a simple random sample of the same sample size – in other words, it is the square of the design factor.

84. Full documentation of all of these PROCs is accessible online[22,23]. Searching for SAS and the name of the PROC usually produces links to the relevant SAS Institute documentation as well as examples of their use.

[22] http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#surveyfreq_toc.htm

[23] http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#surveymeans_toc.htm

Annex 3 – Example bootstrapping code

85. The following SAS code can be used with either the EUL or SAF FRS datasets to generate bootstrap estimates of confidence limits. This example uses the mean and median of benefit unit income.

86. The calculate macro is where most customisation would be expected to occur. It takes a (possibly resampled) table and reduces it to a single row containing estimates of the parameters of interest. It is called once to generate the central estimate, and repeatedly to generate the bootstrap estimates.

87. The bootstrap macro does the bootstrapping. It loops around, calling PROC SURVEYSELECT to resample the households, merging on the benefit unit level data, calling calculate to generate the bootstrap estimate, and appending its output to the output dataset.

88. Finally, a call to PROC UNIVARIATE determines the 2.5th and 97.5th percentiles of the distributions of the bootstrap results, which are the lower and upper 95% confidence limits.

89. The dataset `centralEstimate` contains the central estimates of the parameters of interest in variables named `buinc_mean` and `buinc_median`. The dataset `confidenceIntervals` contains the lower and upper confidence limits for the mean in variables named `buinc_mean2_5` and `buinc_mean97_5`, respectively, and the confidence limits around the median in similarly named variables.

```
%* Macro to delete a dataset from WORK;
%macro deleteIfExists(dataset);
      %if %sysfunc(exist(&dataset)) %then %do;
            proc datasets library=work nolist;
                  delete &dataset;
            quit;
      %end;
%mend;

%* Macro to do all of the calculations, taking a dataset and ;
%* reducing it to one row of numbers;
%macro calculate(outData,inData,bootcycle,dsetName);
      %* Calculate mean and median of the income distribution;
      proc summary data=&inData noprint;
            weight gross4;
            output out=&outData(drop=_type_ _freq_)
                                    mean(buinc)=buinc_mean
                                    median(buinc)=buinc_median;
      data &outData;
            format dataset $100. bootcycle 8.;
            dataset="&dsetName";
            bootcycle=&bootcycle;
            set &outData;
      run;
%mend;

%* Copy the relevant data into WORK and determine the size of the sample;
```

```
      set frs.househol(keep=sernum) end=lastone;
      if lastone then call symputx("NSIZE",put(_n_,best12.));
data benunit;
      set frs.benunit(keep=sernum benunit buinc gross4);
run;

%* Calculate the central estimate using the raw dataset;
%calculate(centralEstimate,benunit,.,%sysfunc(pathname(frs)));
%* Do the bootstrapping;
%macro bootstrap(outData,nCycles);
      %local i optvals;
      %* Make a note of the current values of various system ;
      %* options and then turn them all off;
      %let optvals=%sysfunc(getoption(NOTES))
                              %sysfunc(getoption(SOURCE))
                              %sysfunc(getoption(SOURCE2))
                              %sysfunc(getoption(MPRINT))
                              %sysfunc(getoption(MLOGIC))
                              COMPRESS=%sysfunc(getoption(COMPRESS));
      options NONOTES NOSOURCE NOSOURCE2 NOMPRINT
                              NOMLOGIC COMPRESS=NO;
      %* Will append to &outData, so ensure that it does not exist;
      %deleteIfExists(&outData);

            %* Resample the households...;
            proc surveyselect data=frs.househol(keep=sernum)
                              out=resampledHousehol
                              method=urs sampsize=&NSIZE noprint;
            run;
            %* ...merge on the lower level data and expand the list ;
            %* so that a household cloned N times has N records.;
            data resampledBenunit;
                  merge resampledHousehol(in=isResampled)
                              benunit;
                        by sernum;
                  if isResampled;
                  do copyNumber=1 to numberhits;
                        output;
                  end;
            run;
            %* Calculate the parameters of interest and append them ;
            %* to the results dataset;
            %calculate(oneCycle,resampledBenunit,
                                    &i,%sysfunc(pathname(frs)));
            proc append base=&outData data=oneCycle force;
            run;
      %end;
      %* Reset the system options to what the noted values;
      options &optvals;
%mend;
%bootstrap(bootResults,500);

%* Calculate the parameters of the distributions;
proc univariate data=bootResults noprint;
      var buinc_mean buinc_median;
      output out=confidenceIntervals pctlpts=2.5 97.5
                              pctlpre=buinc_mean buinc_median;
run;
```

Annex 4 – Number of cycles to use when generating bootstrap estimates of confidence limits

90. The question of the number of repetitions to use in generating bootstrap estimates of uncertainty does not have a simple answer. In general, it must be empirically determined. For analysis with the FRS, we recommend that at least 500, and preferably 1000, bootstrap cycles should be used.

91. To determine whether the number you have used is appropriate, plot the confidence limits based on the first n bootstrap cycles as a function of n. An example is shown in Figure 3. When the confidence limits have settled down, enough samples have been used. It is possible to define formal limits on what “settled down” means (for example, n=N bootstrap cycles is sufficient if the standard deviation in either confidence limit calculated from n=N-99, N-98, ..., N bootstrap cycles is less than 0.1% of the mean confidence limit calculated from the same sequence), or to accept a value from a visual inspection of the plot.

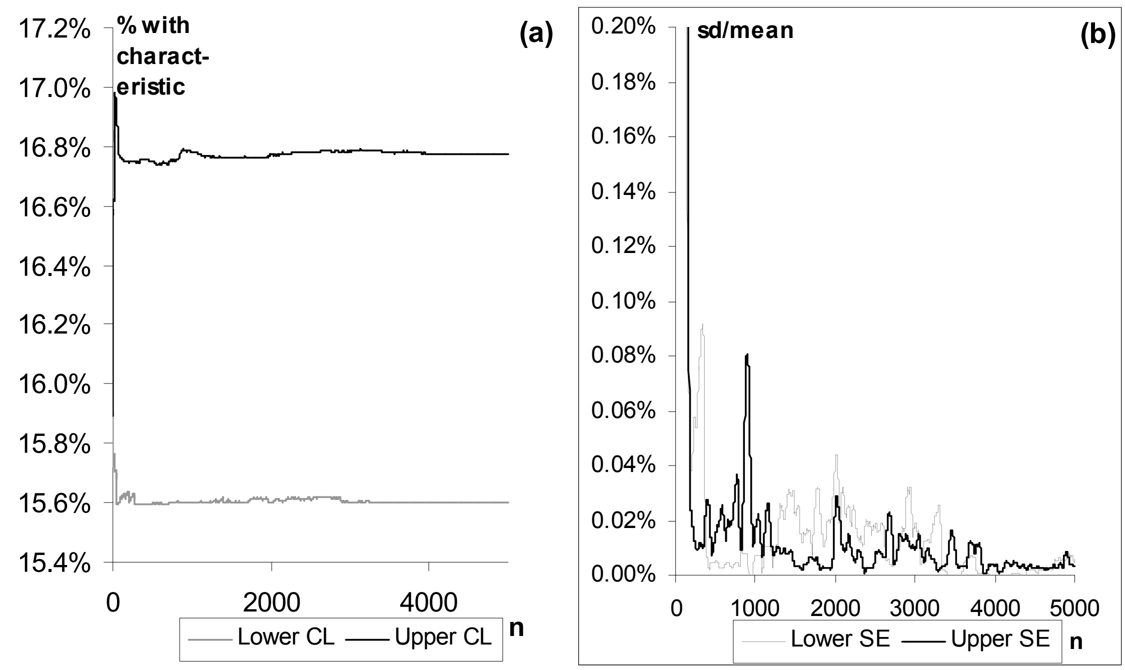


Figure 3: (a) Upper and lower confidence limit estimates for the percentage of the population with an arbitrary characteristic, derived from the FRS and expressed as a function of the number, *n*, of bootstrap cycles used. A visual inspection of the lines suggests that around 1000 cycles is sufficient for both upper and lower confidence limits to have stabilised, although variation remains. (b) The standard deviation over the mean of the confidence intervals based on (*n*-99) to *n* repetitions, expressed as a function of *n* (thus the point on the grey line at *n*=1,500 in this graph (b) is the standard deviation of the points *n*=1,401 to *n*=1,500 on the grey line in graph (a) divided by the mean over the same range). The point of diminishing returns has been reached around *n*=1000.

92. Excel, or a similar spreadsheet package, can be used to determine confidence limits as a function of the number of bootstrap cycles. The Excel PERCENTILE function can be used to calculate confidence limits from a range. If the bootstrapped estimates of a parameter (for example the contents of the dataset bootResults generated by the program in annex 3) are listed in Excel, PERCENTILE can be used to calculate the bootstrap estimate of the confidence limits based on the first n bootstrap repetitions. The STDEV and AVERAGE functions can be used to calculate a normalised estimate of the variation in the estimates based on n-99 to n cycles.

93. This is illustrated in the table overleaf. Column B is the list of the estimates of the parameter generated from the resampled datasets. Columns C and D show the lower and upper confidence limits based on column B up to that line. Columns E and F show the rolling estimate of the variation at this time.

94. The white sections of the table can be pasted into the appropriate cells of an Excel table. The graphs in Figure 3 were generated by plotting columns C and D versus A, and E and F versus A.

	A	B	C	D	E	F
1	n	Parameter	Lower CL	Upper CL	Lower SE	Upper SE
2	1	15.9623	=PERCENTILE(B$2:B2,0.025)	=PERCENTILE(B$2:B2,0.975)		
3	2	15.4369	=PERCENTILE(B$2:B3,0.025)	=PERCENTILE(B$2:B3,0.975)		
	⋮	⋮	⋮	⋮	⋮	⋮
101	100	15.7018	=PERCENTILE(B$2:B101,0.025)	=PERCENTILE(B$2:B101,0.975)	=STDEV(C2:C101)/AVERAGE(C2:C101)	=STDEV(D2:D101)/AVERAGE(D2:D101)
102	101	15.8082	=PERCENTILE(B$2:B102,0.025)	=PERCENTILE(B$2:B102,0.975)	=STDEV(C3:C102)/AVERAGE(C3:C102)	=STDEV(D3:D102)/AVERAGE(D3:D102)
	⋮	⋮	⋮	⋮	⋮	⋮