

3. 1991 SARs

3.1 Introduction

Two files are available from the 1991 Census – an individual SAR which maximises geographical detail and has only limited information about the household in which the individual lives and a Household file which contains detailed information about all individuals in the household but has no geography below standard region.

The Individual SAR is a 2 per cent sample of the population. It includes 1.1 million visitors and residents in private households and communal establishments. There are over 278 SAR areas, based on local authorities or groups of local authorities of at least 120K population. The Individual SAR file includes a full range of census topics on individuals and summary information about households. Details of all variables, including geography, are available in the Codebook. An Individual SAR for Northern Ireland was released as a separate file.

The Household SAR is a 1 per cent sample of the population. It includes 216 thousand households amounting to around half a million persons within households. The data allows linkage between household and family members. In terms of geography the data includes Standard regions plus inner-London and outer-London. The Household SAR file includes a full range of census topics on individuals and derived household and family level variables. A Household SAR for Northern Ireland was released as a separate file.

3.2 Sampling the 1991 SARs

The coding of the 1991 Census for Great Britain was divided into two stages. Easy to code information, such as sex, date of birth, marital status and country of birth, was processed for all forms and then a 10 per cent sample was selected and the remaining 'hard to code' questions, mainly those relating to occupation, industry and qualification, were coded. This 10 per cent sample was then used as the base from which to draw the SARs.

The sampling for the SARs was divided into two stages, with the one per cent Household file selected first. All fully-coded household forms were ordered geographically with the lowest level the enumeration area (about 200 households). Households were then grouped into batches of 10 and one household selected at random from each batch. All sampled records were then scrambled before release to prevent households being traced by their geographical ordering.

The two per cent Individual sample was then drawn from the remaining households, hence there is no overlap between the two samples. Individuals in the remaining households were stratified into groups of nine, and two individuals selected from each group at random. It is therefore possible that more than one individual may be selected from the same household.

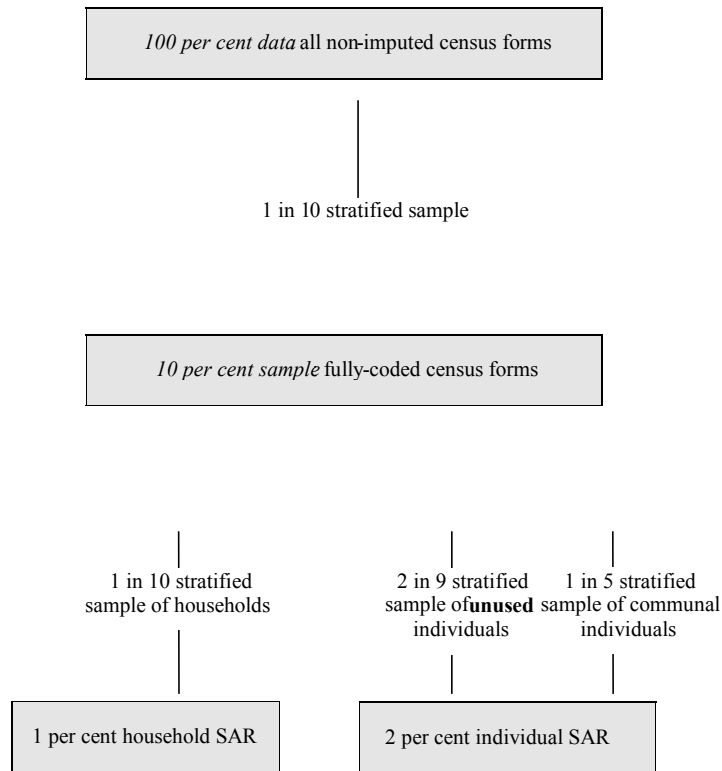
For the final stage of the sample design, individuals in communal establishments were stratified into groups of five and one individual selected at random from each group. Once again, the records were scrambled within each SAR area before being released.

In Northern Ireland all records were 100 per cent coded. The NI 1 per cent Household sample was selected first by stratifying households within enumeration districts and District Councils into groups of 100 households and then selecting one household at random from each group. The hierarchical household SAR contains 20,833 records in total: 5,255 households and 15,578 persons within those households. The variables are broadly similar to those within the GB household file (bearing in mind the differences which exist between the two Censuses) whilst, as in the GB file, there are no individual records released for households containing 12 or more individuals. The Northern Ireland census database does not hold information as to which family an individual belongs, so unlike the three-tiered structure of the GB Household SAR, analysis is only possible at the level of the household and the individual.

The NI 2 per cent Individual sample was selected by stratifying the remaining individuals into groups of 99 and by choosing two individuals at random from each group. Individuals in communal establishments were stratified geographically into groups of 50 people and one person was chosen at random from each group. There are 31,967 individual records on 10 area files, with information provided on an individual's resident status (present resident, absent resident or visitor). As with the Household file, the variables contained within the Northern Ireland 2 per cent Individual file are broadly similar to those in the GB 2 per cent file.

As with the GB SARs, to prevent any possible geographical tracing within a SAR area, the files were scrambled before release.

Summary of Sampling Method



Variables in the SAR files show the effects of both stratification and clustering. Attributes that tend to be common across areas will be affected by stratification (for example, local authority housing tenure) and will therefore have a lower sampling error than that for a simple random sample. Other variables, where values tend to be the same for all household members, will be affected by clustering, leading to larger than expected sampling errors. For example individuals within the same household are likely to have the same ethnic group and social class. The effect of clustering is more pronounced for individual level variables in the household file, as all individuals in each household are selected for this sample.

3.3 Differences across UK Countries

The 1991 Census form was very similar in England, Wales and Scotland. In Northern Ireland there were more differences in questions – e.g. religion was asked and fertility of married women, and differences in coding procedures – for example data were 100 per cent coded. As for GB, two non-overlapping samples are available. As far as possible, similar derived variables have been added to the Northern Ireland SARs. The main difference is that family level derived variables cannot be added because of the way in which families are identified in the Northern Ireland Census.

The 1 per cent Household file for Northern Ireland has no additional geographical identification whilst the 2 per cent Individual file identifies seven districts.

1991 census forms are available in PDF format for England and Wales (Communal Establishments and Private Households), Scotland and Northern Ireland.

3.4 Disclosure Control Measures in the 1991 SARs

There is a legal requirement for everyone to complete a census schedule. However, the Census Offices also have an obligation, under the 1920 Census Act, not to disclose any identifiable information that has been provided in the census. It is therefore of the paramount importance that samples of anonymised records from the census do not pose any threat to this confidentiality undertaking.

In order to address the risk of disclosure, the Economic and Social Research Council Working Party set out a four-stage process through which disclosure could occur. The process was premised on the assumption that, before disclosure could occur, an individual or household would have to be correctly identified. It was assumed that the most likely way for this to occur was by matching variables in the SAR with the same information on another, external file which held the identification of the individual or household.

The four steps necessary for disclosure were:

- (1) Key variables in the microdata sample would have to be recorded in a compatible way on an outside file. If key variables are not recorded in the same way, or contain errors, then correct matches are unlikely.
- (2) The individual in an outside file would have to be selected in the microdata sample before a match is possible.
- (3) The individual's combination of values on the key variables must be unique in the population - otherwise an apparent match with a member of the microdata sample could, in fact, be with a 'statistical twin'.
- (4) The person attempting to make the match would need to be able to verify uniqueness in the population - for example by having a list of the entire population on the key variables and thereby being sure that the match is, indeed, correct.

Rough estimates of the size of risk at each stage were made; when cumulated, the risks of disclosure appeared very low; multiplying the various probabilities together, the working party concluded that the risk of anyone in the population being identifiable from their SAR record was

negligible. Details of the calculations made at each stage are given in Marsh, Skinner et al. (1991). The arguments put forward were important in persuading the Census Offices to release the SARs, suitably modified to protect anonymity where this was felt to be at risk. In the next section the various disclosure protection measures taken are described.

Sampling as protection

The low sampling fractions of the SARs offer a strong source of disclosure protection for sensitive data. It not only reduces the actual risk that a particular individual can be found in the census output but it also reduces the chances that anyone would make an attempt at identification by this means. The two SARs (totalling 3 per cent together) are sufficiently small to offer a great deal of protection; the samples do not overlap so that the detailed household or occupational information available on the household file cannot be matched with the detailed geographical information available on the individual file.

Some alterations have been made to the data to reduce the number of rare and possibly unique cases. Information which is unique in itself, such as names and addresses, is, of course, omitted altogether (neither is it included on the ONS Census database in the first place), whilst the precise date of birth of individuals has been suppressed (age is based on the number of completed years of age at the time of the Census).

Restricting geographical information

One of the key considerations which may affect the risk of identifying an individual or household is the geographical level at which data is released. Empirical work and comparisons with SARs released in other countries showed that a sensible level for release would be areas with a population size of at least 120,000 in the individual (2 per cent) SAR. This level of geography allowed the majority of British local authorities to be separately identified. Smaller local authority districts (under 120,000 population) were grouped to form areas over 120,000. Only one geographical scheme was permitted in order to avoid overlaps where the difference between two areas could lead to the identification of sub-threshold area.

The one per cent household SAR, because of its hierarchical nature (i.e. records for the household and all its members), is more of a disclosure risk. For this reason it was decided that, for this SAR, the lowest geographical level would be the Registrar General's Standard Regions, plus Wales and Scotland. The only exception is that the South East is split into Inner London, Outer London, and the Rest of the South East Region. The smallest region, East Anglia, has a population of about 2 million.

The 1 per cent Household file for Northern Ireland has no additional geographical identification whilst the 2 per cent Individual file identifies seven districts. The geography of the Northern Ireland 2 per cent Individual SAR is based on combinations of District Councils into ten geographical areas, each area again having a minimum resident population of 120,000. Areas have been defined in order to combine district councils with similar characteristics.

Suppression of data and grouping of categories

Some alterations were made to the data to reduce the number of rare and possibly unique cases. With some variables, small categories have been grouped, either across the entire range of the variable or at the extremes (for example, for those aged over 90). The rule used to decide the level of detail to be released was that, on average, the expected sample count would be at least one for each category of each variable at the lowest geographical area permitted on each SAR. This rule was applied to each census variable.

When expected frequency counts fell below the threshold, categories were grouped. With some variables, grouping was only required at one end of the distribution: thus 'rooms' were topcoded above 14 and the number of persons in the household was topcoded above 12. With age, 91 and 92 were grouped, 93 and 94 were grouped and 95 and over was topcoded.

Generally, less detail was released on the two per cent individual SAR, because of the lower level of geography. For example, occupation was reduced to 73 categories whereas in the household file it was coded to 358 categories.

Some additional restrictions were applied to certain occupational groups which were considered a particular risk because of being in the public eye - for example actors, professional sportsmen and women and politicians. Large households were also seen as a disclosure risk in the household sample. Therefore for households containing 12 or more persons no information about the individuals in the household is given. In fact, only 28 households in the sample contained 12 or more persons. The other area of special concern was geographical information on workplace and address one year before census; this was heavily grouped before release.

The geographical ordering of records in both SARs do not reflect their geographical ordering within the ONS database. Although sampled with households grouped by county and enumeration district in England and Wales and by region and output area in Scotland, once selected, the records have been scrambled. This prevents any possibility of tracing individuals or households back through a region or district.

Finally, SAR users have to give an undertaking not to obtain or derive information relating specifically to an identified individual or household, nor claim to have derived such information. Due to the uniqueness of the SARs in British census history, it is extremely important that these conditions are met. Any breaches of the undertaking will result in the recall of the data.

3.5 Accuracy of estimates from the 1991 SARs

3.5.1 Introduction

Estimates prepared from the Samples of Anonymised Records are based on a sample of the 1991 Census data. They are estimates of the actual figures that would have been obtained from a complete enumeration of all residents. These estimates are expected to be different from complete figures because they are subject both to sampling errors and non-sampling errors. They will not necessarily be the same as those published in census reports. This section of the user guide discusses sampling and non-sampling errors in some detail and suggests how the user should assess these errors in practice. The advice can be summarised as follows:

- users should calculate confidence intervals to reflect the sampling error attached to estimates calculated from the SARs
- users should be aware of the most likely deficiencies in the quality of census responses and include relevant cautions in reports based on SARs data
- users should adjust figures from the SARs to take into account the incomplete coverage of the population, particularly where totals of population categories have been estimates rather than ratios or percentages.

3.5.2 Sampling error

Because the SARs are based on a random sample, estimates based on them may differ somewhat from the figures that would be obtained from processing all the census records; they may also differ from the estimate that would have been obtained from processing a different sample of the same size drawn in the same way from the census records.

Comparisons of SARs and population data

For a limited number of variables it has been possible to compare the estimate from the SARs and the value that is given by all census records that the SARs are drawn from. For Great Britain these comparisons are given in the table below. Throughout the table, the population base

includes present and absent residents but excludes visitors, imputed absent households and residents in imputed absent households. As one might expect with such large samples, the SARs closely represent the population from which they were drawn. Conversely, the smaller the size of the sample the greater the tendency of estimates to differ from corresponding values for the entire population. Consequently, estimates derived from the SARs for sub-groups of the population or single SAR areas will tend to deviate more from the 100 per cent statistics.

The deviation of a sample estimate from the census value is called the sampling error. The standard error of a sample estimate is a measure of the variation (the standard deviation) of the sampling error across all the possible samples and thus is a measure of the precision with which an estimate from a particular sample approximates the census value.

Table 1: Characteristics of the SARs and the population from which they were drawn Great Britain, percent.

Individual Characteristics

	% OF ALL RESIDENTS		COMMUNAL ESTABLISHMENTS	
	Individual SAR	Census population	Individual SAR	Census population
Male	48.4	48.4	41.7	41.2
Female	51.6	51.6	58.3	58.8
Age 0-15	20.2	20.2	3.4	3.7
16-17	2.5	2.5	1.2	1.3
18-29	18.1	18.1	21.6	21.3
30-44	21.3	21.2	9.9	10.1
45 up to pensionable age	19.3	19.3	8.6	8.9
Pensionable age	18.7	18.7	55.2	54.7
Single	41.0	41.1	47.3	48.0
Married	46.9	46.8	12.5	12.7
Widowed/Divorced	12.1	12.1	40.1	39.3
With Ill illness	13.1	13.1	63.5	63.3
In employment	44.1	44.3	21.6	22.1
Unemployed	4.6	4.5	3.7	3.6
Economically inactive	31.2	31.1	71.3	70.6
White	94.6	94.6	94.3	94.5
Other ethnic groups	5.4	5.4	5.7	5.5

Household Characteristics

	% OF RESIDENTS IN HOUSEHOLDS		% OF HOUSEHOLDS	
	Individual SAR	Census population	Household SAR	Census population
One person in household	10.6	10.6	26.3	26.3
Owner occupied	69.9	70.0	66.4	66.7
Rented privately (exc with job)	5.5	5.5	7.2	6.9
Rented from a housing association	2.4	2.4	3.2	3.1
Rented from a local authority, new town or Scottish Homes	20.0	20.0	21.3	21.4
Lacking or sharing use of a bath/shower and/or inside WC	0.74	0.75	1.3	1.2
No central heating	16.8	16.8	18.8	18.8
No car	24.9	24.9	33.3	33.1
Lone parent	n/a	4.15	3.7	3.7

Sources. Individual SAR, Household SAR, LBS (Tables 18 and 19 for imputed households, deducted from equivalent cells for 100 per cent data in other LBS tables). The base in each case excludes imputed households and residents in them. Crown Copyright.

The sample estimate and its estimated standard error permit the construction of interval estimates with prescribed confidence that the interval includes the true population value.

3.5.3 Non-Sampling error

In addition to the variability which arises from the sampling procedures, both sample data and the full census data are subject to non-sampling error. Non-sampling error may be introduced during any of the complex operations used to collect and process census data.

Non-sampling error may affect the data in two ways. Errors that are introduced randomly will increase the variability of the data, and should, therefore, be reflected in the standard error discussed below. Errors that tend to be consistent in one direction will make both sample and 100 per cent data biased in that direction. For example, if respondents consistently tend to under-report the number of cars available to their household then the resulting counts of households by number of cars will tend to be understated for the multi- car households and overstated for the no-car households. Such biases are not reflected in the standard error.

Sources of non-sampling error include:

a) Quality of response

Respondents to the census may misinterpret census questions or for other reasons complete the census form incorrectly. The census form requests that the head or joint heads of the household, or other adult over 16, completes the form on behalf of all members of the household. The Census Validation Survey (CVS) carried out by OPCS shortly after the 1991 Census assessed the quality of responses to the census.

b) Incomplete coverage of the census

Every census misses some people who are particularly difficult to enumerate, in spite of the thorough census field procedures designed to enumerate the entire population (evaluated in Clark 1992). The age-structure of those missed by the census has also been estimated and is significantly different from the age-structure of the population as a whole.

c) Transcription and coding errors, missing data items

During the processing of census forms, transcription and coding errors can occur. Missing items for persons on a completed census form are imputed (estimated) by the Census Offices. Corrections are made to some inconsistent data, such as persons reported married but aged under 16. Mills and Teague (1991) provide a description of the processing of census forms and the imputation of these types of missing or inconsistent data.

d) Data Modification to ensure confidentiality

In the 100 per cent tabular output of Local Base Statistics and Small Area Statistics for areas within local authorities, an additional source of error was purposefully introduced by Census Offices to provide additional protection against the identification of individuals (Census User Guide 48; Cole 1993). Counts in some cells of the tabulations are slightly adjusted; the cells that are adjusted are not known to the user. However, no such adjustment is made to the sample data in the 10 per cent tabular output or in the SARs. Other methods, described in Section 1.4, reduce the already negligible risk that individuals can be identified from records in the SARs.

3.5.4 Standard Errors and Confidence Intervals

The complex sampling design described above has implications for the estimation of sampling errors on both the individual and household file. The 1 per cent household SAR approximates to a simple stratified random sample of households, although counts of individuals in the household file are subject to the effects of clustering. In the 2 per cent individual file there are two potential sources of clustering which arise in the sampling process. First individuals are clustered into households in the selection of the 10 per cent sample and second, the removal of the household SAR

from the 10 per cent sample implies a further clustering into households (Dale and Marsh, 1993). Nonetheless, preliminary work suggests that the 2 per cent SAR approximates to a simple random sample.

Calculating standard errors and confidence intervals

The method described here for estimating standard errors of estimates from the SARs involves two simple stages. The first stage calculates the unadjusted standard error, using formulae that apply to simple random samples. The second stage multiplies the unadjusted standard error by an appropriate design factor. This is the factor by which sampling errors must be multiplied in order to compensate for the effect of clustering or stratification in the sampling process. The design factor approximates the ratio of the standard error from the actual sample design to the standard error from a simple random sample. In practice the steps are:

1. Calculate the unadjusted standard error from the appropriate formula at (ii) below.
2. Multiply the unadjusted standard error from step 1 by the design factor appropriate to the characteristic (e.g. unemployment status, or age).

The design factor that should be applied may be more or less than 1.0. If there is stratification in the sampling process the sample should be more representative than a simple random sample and the design factor will be less than one. Clustering will cause sampling errors to be larger than those found with simple random sampling and the design factor will be greater than one.

Preliminary estimations of design factors have been made using two different methods, the first using sampling point information (for the household file); the second comparing differences between expected and observed errors (for the individual file).

Design factors for household characteristics from the 1 per cent household SAR

Assumption of a design factor of 1.0 (i.e. using the unadjusted standard errors as if the sample was a simple random one) is unlikely to be far wrong when using household characteristics from the 1 per cent SAR. At worst, a slight over-estimate of the sampling error may result, as household level variables are subject to stratification effects and estimated design factors (including those relating to particular members of the household, for example the social class of the head of household) are slightly less than unity.

Design factors for individual characteristics from the 1 per cent household SAR

For analyses of individual characteristics from the 1 per cent household SAR, assuming simple random sampling may be misleading because clustering effects mean that sampling errors may be seriously underestimated. This is because this SAR includes all individuals in each sampled household and for variables such as ethnic group, country of birth, migrants, qualifications and social class, there is a tendency for individuals in the same household to have similar characteristics. The effect of household clustering could probably be ignored however for estimates of subgroups of which there is usually no more than one person per household, such as women aged over 80.

The largest effects are for ethnic group. Preliminary estimates are as follows:

Ethnic Group

White	1.84
Black Caribbean	1.60
Black African	1.83
Black Other	1.51
Indian	1.99
Pakistani	2.27
Bangladeshi	2.37
Chinese	1.87
Other Asian	1.83
Other Other	1.60

Design Factors in the 2 per cent Individual SAR

Design factors estimated for the individual SAR are based on a comparison of the difference between the SARs and 100 per cent Census data across the 278 SAR areas (having subtracted residents in wholly imputed households) and the sampling errors which would be expected from simple random sampling. The method is described in more detail in CMU Occasional Paper 2.

Individual and household level variables on the individual file are less likely to be subject to clustering and may benefit from stratification. Most design factors deviated very little from unity, many being less than one. Again the largest design factors are for ethnic group, though the effects are much smaller than on the household file.

Ethnic Origin

White	1.15
Black Caribbean	1.00
Black African	1.06
Black Other	1.04
Indian	1.26
Pakistani	1.20
Bangladeshi	1.04
Chinese	1.19
Other Asian	1.02
Other Other	1.30

Having calculated the standard error for a SAR estimate, it will often be appropriate to go on to calculate a confidence interval for the estimate. These are discussed below.

Users should also read the notes below which give further advice on the use of standard errors. Worked examples are given throughout this discussion of standard errors and their use. More details on estimated design factors are available on request from the CMU.

Generally, use of the 2 per cent individual SAR will minimise sampling errors for individual level analyses whilst the household file is the most appropriate for the analysis of household characteristics.

Calculation of standard errors

The means of calculating the unadjusted standard errors for four common statistics are given here. The derivations can be found in many statistics textbooks and most statistical software will calculate them as part of their standard output.

Statistic	Value	Approximate standard error
Sample cell count	c	$SE(c) = \sqrt{c(N-c)/N}$
Scaled cell count	$C = f * c$	$SE(C) = f * SE(c)$
Sample cell proportion	$Pr = c/n$	$SE(Pr) = \sqrt{Pr(1-Pr)/n}$
Sample cell percentage	$Pe = 100 * c/n$	$SE(Pe) = \sqrt{Pe(100-Pe)/n}$

Examples of the statistics, and definitions:

c the number of non-white textile workers in Yorkshire and Humberside region.

$C=f*c$ that number scaled to the total census enumerated population. In this case f is 50 or 100 for the individual or household SAR respectively.

N the total number of records in the SAR in the Yorkshire and Humberside region, irrespective of industry or ethnic group. In general, N is the total number of records in the SAR for the area concerned; where a characteristic of the population in communal establishments is being counted in the individual SAR, N is the total number of records from communal establishments in the area concerned. Where N is very large compared to c (N more than 30 times c), the formula can be replaced by the approximation $SE(c)=\sqrt{c}$ and $SE(C)=f*\sqrt{c}$.

P_r The number of non-white textile workers in the region (c) as a proportion of all non-whites in employment in the region (n).

P_e The number of non-white textile workers in the region (c) as a percentage of all non-whites in employment in the region (n).

The standard error of the SAR statistic is then derived by multiplying the unadjusted standard error from these formulae by the appropriate design factor.

Examples of calculation of standard errors

(a) The percentage of the population of Newham who are of Indian ethnic origin. The percentage of the sample who are Indian in Newham is 13.8 per cent. The unadjusted standard error is

$$\text{Unadjusted } SE(P_e) = \sqrt{13.80*(100-13.8)/1000} = 1.19$$

The estimated design factor for Indians on the individual file is 1.26. The standard error for this SAR percentage is therefore

$$\text{Standard error } (P_e) = 1.19*1.26 = 1.50$$

(b) The number of renting households in Britain with a person under pensionable age having a limiting long-term illness, from the household SAR, scaled to a total for all households enumerated in the census.

If the total number of such households in the household SAR is 289, it is scaled by 100 (the household SAR sampling fraction) to estimate a total in Britain of 28,900 such households. There are 215,789 household records in the household SAR in all, so the unadjusted standard error of the estimate of 28,900 is

$$\text{Unadjusted } SE(C) = 100*\sqrt{289*(215,789-289)/215,789} = 1,699$$

Note that the number $c=289$ is very small compared to the overall number of records $N=215,789$, so a very similar result would be achieved using the approximation referred to on the previous page

$$\text{Unadjusted SE}(C) = 100 \cdot \sqrt{289} = 1,700$$

From the discussion in the previous section, the design factor for household characteristics from the household SAR may be taken to be 1.0, so in this case the standard error requires no further adjustment.

Confidence intervals and inferences based on the SARs

A sample estimate and its estimated standard error may be used to construct confidence intervals around the estimate. These intervals are ranges that will contain the true population value of the estimated characteristic, with a known probability.

For example:

1. With approximately 68 per cent probability, the interval from one standard error below the estimate to one standard error above the estimate contains the true value.
2. With approximately 90 per cent probability, the interval from 1.6 standard errors below the estimate to 1.6 standard errors above the estimate contains the true value.
3. With approximately 95 per cent probability, the interval from two standard errors below the estimate to two standard errors above the estimate contains the true value.

The intervals are referred to as 68 per cent, 90 per cent, and 95 per cent confidence intervals, respectively.

Example

Using the earlier example, the standard error of the 28,900 households in Britain with someone below pensionable age with a limiting long-term illness was estimated to be 1,698. Thus a 95 per cent confidence interval for this estimated total is estimated as:

$$(28,900 - 2 \cdot 1,698) \text{ to } (28,900 + 2 \cdot 1,698), \text{ or } 25,504 \text{ to } 32,296$$

Other notes on standard errors

A standard sampling theory text or the explanatory guide to the user's statistical software should be helpful if the user needs more information

about confidence intervals and non-sampling errors. These should be consulted for details of standard errors for sums, differences and ratios of estimates from the SARs.

Zero estimates

When the proportion, percentage, or cell count is zero, the formulae in section (ii) above give estimated standard errors of zero. While the magnitude of the error is difficult to quantify, estimated percentages and totals of zero are still subject to error.

The effect of non-sampling error on the standard errors and inference using confidence intervals. The estimated standard errors given above do not include the variation due to non-sampling error that may be present in the data. The standard errors reflect the effect of simple response variability, but not the effect of systematic errors introduced by enumerators, coders, or other field processes. As a result, confidence intervals formed using these estimated standard errors may not meet the stated levels of confidence in estimating the true population value of a characteristic. One of the most important sources of error that might additionally affect the accuracy of confidence intervals is bias arising from missing records, discussed below.

3.5.5 Quality Of Census Responses

One characteristic of the SARs is that the accuracy of responses contained in them is determined almost wholly by the accuracy of the responses given by residents themselves. There has been no data modification or perturbation and imputed records for wholly absent households have not been included in the SARs; only data that was missing from a returned form has been imputed, by the 'hot deck' procedures described in Mills and Teague (1991), for the 100 per cent coded variables.

A check on the quality of responses in the Census was one of the aims of the Census Validation Survey (CVS) carried out very soon after the 1991 Census. The CVS, which seeks to establish the quality of both responses to, and coverage of, the Census, is based on a sample of around 6000 households in over 1200 enumeration districts, and was administered by means of individual interviews held between six weeks and three months after Census day (Wiggins, 1993).

1991 Census Validation Survey: quality report (Heady et al, 1996) HMSO is available online at www.statistics.gov.uk/about/data/methodology/specific/population/LS/sources/cvs.asp. This contains full details of the quality of responses. One example of the kind of information contained in the report is given below:

3.5.6 Ethnic group

The ethnic group question was newly introduced in the 1991 Census. After extensive testing in the field, it was decided to use a question which gave form fillers nine possible categories from which to choose, two of which asked for more detailed information to be supplied. The number of ethnic minority households in the CVS sample was not sufficient to justify individual analysis of all nine categories and so four aggregate codes were created: white, black (combining black Caribbean, black African and black 'other'), Indian sub-continent (Indian, Pakistani and Bangladeshi) and other (Chinese and 'any other ethnic group'). The gross error rate was only 0.8 per cent. However, this figure should be treated with caution given that the vast majority of answers were in just one category (white). If those who answered 'white' in both the Census and CVS are excluded, the gross error rate was 13.2 per cent. It was found that 21 per cent of those coded as 'other' in the Census either described themselves as white or in one of the black categories in the CVS. Conversely, 9 per cent of those coded 'black' in the Census described themselves as 'other' in the CVS. Overall, 6.1 per cent of people in households who replied in both the Census and CVS were in the non-white ethnic group in the CVS, compared with 5.8 per cent of the same people according to their replies in the Census (Heady et al, 1996).

3.6 Incomplete Coverage

Two types of resident are missing from the SARs. This section is not concerned with the effects of sampling, which have been described earlier, but with the completeness of the census itself. The SARs were drawn from the fully coded set of census records returned by households and institutions. There are two main categories of residents missing from this set of records:

3.6.1 Imputed absent households

Census records representing 869,000 residents in Great Britain (1.6 per cent) have been imputed by the census offices by using the enumerators' estimate of the number of residents in absent households and then copying characteristics from geographically adjacent households who returned late census forms. These records were used in compiling the 100 per cent tabular census output, but are excluded from the 1991 SARs, which are drawn from the 10 per cent sample. This imputation procedure was followed wherever an enumerator felt that a housing space contained residents but a) a household absent on census night did not return a form under the voluntary arrangements; or b) the enumerator could make no contact at all; or c) residents refused to complete and return a form.

3.6.2 Other residents missed by the census

Imputation of absent residents 'captured' less than half of the number estimated not to have been enumerated in the census. In some cases residents were not included on the census forms that were returned. In other cases whole households were missed by enumerators who had difficulty enumerating, for example, those living in converted and multi-occupied properties. The number of residents not included in the 100 per cent output, that is neither enumerated nor imputed, was estimated to have been 1.2 million in Great Britain (2.1 per cent).

(Since the 2001 census the assumptions above have been challenged. The following text is taken from the ONS web site at www.statistics.gov.uk/census2001/implications.asp:

'The UK has 800,000 fewer young men than previously thought. This pattern was originally identified in the 1991 census, but given a lack of confidence in the follow up survey for that census, the numbers were revised to restore the predicted pattern. In 2001, the pattern has been confirmed and validated by the one number census. We will revise population estimates back to 1982. The critical factor appears to be emigration. The International Passenger Survey works well, but it captures travellers' intentions at the time of departure. These may be prone to change once people are abroad, particularly among young men with few ties at home.'

A small number of individuals are excluded from the household SAR for confidentiality reasons: those where the number of persons in the household is twelve or more. These comprise 28 (0.013 per cent) households, approximately 0.06 per cent of residents in households.

3.6.3 Allowing for incomplete census coverage in analysis using the SARs

Often, the user will wish to make an inference from the SARs about conditions of the full population on Census night in 1991. Those missing from the Census data from which the SARs were drawn may have distinctive characteristics. To enable users to compensate for those missing in the SARs (wholly absent imputed households) plus those missed from the Census, weights are being added to the SARs which allow adjustment to the mid 1991 population estimates. These are specific to age, sex and SAR area and are available as a derived variable (POPWGHT). It is important to note that population estimates are based on residents and therefore visitors should be excluded from analysis when applying population weights.

For example, to get an accurate age/sex profile of every SAR area, variables should first be weighted by POPWGHT. The population weights could be used in a similar way to obtain estimates of individuals in other sub- populations such as ethnic groups. However, when using the population weight only age, sex and SAR area are taken into consideration. Weighting assumes that the characteristics of the imputed and missing population are the same as those of the sampled population. In general weights should be used with caution when looking at variables which have small cell sizes because very small groups may be disproportionately boosted by a large weighting factor. The method of making such adjustments, and the magnitude of the impact of census undercount on some simple census indicators is described more fully in Simpson (1993).