SAMPLING AND WEIGHTING TECHNIQUES FOR A LARGE COMPLEX SAMPLE OF ACCIDENT DATA.

Susan Stearn
Institute for Consumer Ergonomics
75 Swingbridge Road,
Loughborough,
Leicestershire, LE11 OJB.
England.

ABSTRACT

The Accident Research Unit at the Institute for Consumer Ergonomics has been collecting data on crashed cars and their occupants since 1983. This paper describes the population from which the sample is drawn, and the techniques used to obtain the sample elements; problems in identification of the complete sample frame are highlighted.

The accident population comprises accidents where at least one vehicle is a car less than six years old which has been towed away from the scene of the accident. The sample is stratified by accident severity, geographical location and time period, and weighting factors are used to obtain estimates of parameters of interest in the population. A description is given of how the weighting factors are obtained. Weighting factors need to be re-calculated for carrying out analyses on sub-populations. Tolerance limits for estimates can also be calculated, and examples of these are given, describing the methods used to obtain the results. The methods are shown to be inexact because of the complex structure of the database and the limitations of the computer analysis package in use. Guidelines are given for the type of statistical analyses that can be validly carried out on this type of data.

1. INTRODUCTION

The Accident Research Unit (ARU) at the Institute for Consumer Ergonomics in Loughborough was established in 1983. It was set up as a result of being awarded a large contract to examine crashed cars in detail and collect injury information about their occupants, with a view to obtaining a sizeable database which can be used for analysing vehicle performance and occupant injuries, and the interaction between the two in car accidents. Setting up a large system such as this requires a great deal of work; the ARU had to set up an accident notification system, establish links with hospital Accident and Emergency consultants in order to collect injury information, devise forms for the collection of vehicle and injury data, and develop a computer system for the punching, organisation and analysis of the data. Some of this background work has already been described in other papers (6,7)[1]; this paper attempts to explain the complex structure of the sample, and the methods used to weight data in order to obtain realistic statistics.

2. THE ACCIDENT SAMPLE

2.1 Sample Frame

The ARU collects data on accidents which occur in the counties of Leicestershire, Derbyshire and Nottinghamshire. It also co-ordinates a team of accident investigators who collect data in County Durham.

In addition the Accident Research Unit at Birmingham University collect identical data in the West Midlands, and they co-ordinate a team working in the counties of Avon and Somerset. This paper does not attempt to describe the sampling techniques used by the Birmingham team, which are slightly different from those used by Loughborough; however it is possible to combine data and carry out analyses using all geographical areas.

The criteria for an accident to be included in the sample frame are that there must be a car, or specified type of light van, under six years old involved, and it must have been towed away from the scene of the accident to a recovery garage. Thèse vehicles are called case vehicles, and are examined in detail. Any other vehicles involved are called non-case vehicles, and are examined cursorily, if at all. Detailed injury information is obtained for all occupants of case vehicles, but no injury information is obtained for occupants of non-case vehicles.

Police classify the injury of each occupant as either fatal, serious, slight or not injured. The definition of each injury level is:

Fatal: Death occurring in less than 30 days as a result of the accident.

Serious: Detention in hospital as an in-patient as a result of injuries; death occurring on or after 30 days as a result of the accident; injuries such as fractures, internal injuries, severe cuts and lacerations, crushing, concussion, severe general shock requiring hospital treatment.

Slight: Injuries such as sprains, bruises, cuts judged not be be severe, slight shock requiring roadside attention.

[1] Numbers in brackets refer to the references at the end of the paper.

The accident severity is defined by the ARU as the injury severity of the most seriously injured occupant in any case vehicle. This is equal to or less than the police accident severity for any given accident, as the police accident severity is defined as the injury severity of the most seriously injured occupant in any vehicle in the accident.

2.2 Notification of accidents

When the ARU was established in 1983, it had to set up accident notification systems with local police forces. Different police forces have different ways of recording accident reports; some amass reports centrally at headquarters, and others keep records at divisional or subdivisional level. Complete computerised records are not available until approximately six months after accident dates, so ways had to be found of tapping into police manual records. The main problem encountered was that police officers do not consistently state whether vehicles have been towed away from the accident scene. This was a necessary criterion to have for the accident population, as cars could only be inspected in the required detail at recovery garages; also it tends to eliminate the very low energy accidents where no injuries have occurred. In some police forces it was necessary to ask police officers to provide additional information for the use of the ARU only. This is not an ideal situation because inevitably some officers forget this information, resulting in some relevant accidents being omitted from the population. The notification rates range from 50% to nearly 100% in different counties.

2.3 Sampling rates

Accidents are sampled at different rates, depending on the accident severity. Fatal accidents are sampled at 100%, serious at 80%, slight at 15% and non-injury at 10%.

The effective sampling rate is rather less than these figures because not all accidents are notified. Sampling rates are adjusted to take account of the notification rate, prior to entering data in to the computer. This method would cause biases in analyses of the data if the accidents which were not notified were themselves biased towards a particular type (for example, predominantly low speed urban accidents). From retrospective investigation of the police accident booklets, it appears that accidents which are not notified to the ARU occur randomly in the population; there is no noticeable bias towards any features. It therefore seemed justifiable to adjust the sampling rates, otherwise population estimates of occurrences of events would be underestimates.

Sampling rates are retrospectively calculated on a quarterly basis for each county and each accident severity.

2.4 Non-injury accidents

There is a particular problem in obtaining correct notification rates in the case of non-injury accidents. This is because police non-injury accidents are not computerised, and for many non-injury accidents only very brief details of the accident are recorded, which often does not include details of recovery arrangements of vehicles. It is extremely time-consuming to manually go through every single non-injury accident card, so an alternative method is currently on trial, whereby a sample of one in ten non-injury accident records is examined in order to estimate the total non-injury accident population falling within the ARU criteria. This work is in the process of being carried out, so most of the non-injury sample strata do not yet have estimated sampling rates, and therefore most of the non-injury accident data are effectively unavailable for analysis. It is hoped that this situation will be rectified in the near future so that more analysis can be carried out, particularly calculations of probability of injury in given situations.

3. ANALYSIS OF THE DATA

3.1 Sample structure

It has already been mentioned that the sample is stratified by accident severity. It is also stratified by geographical area, because the accident investigation team based in Durham do not investigate non-injury accidents, and investigate a smaller proportion of slight accidents. There is also some evidence to suggest that the accidents occurring in the three local counties are dissimilar because of the different mix of road types.

Effectively, because of the way the sampling rates are calculated, the sample is also stratified by quarterly time periods. There is no particular intention to vary the sampling rates over time, but it happens because the notification rate varies over time. It would be difficult to justify calculating sampling rates to cover the whole 5 year (and growing) period of data collection when rates have fluctuated considerably within that period.

The sample is not only stratified, but also clustered. When analysing data at vehicle or occupant level, it must be remembered that the vehicle (or occupant) was not individually sampled, but was selected because the accident was sampled. This can cause extra variance in the data because of homogeneity of vehicle and occupant characteristics within one accident. The clustering effect complicates the calculation of statistics, and many can be calculated only approximately.

3.2 Weighting factors

Much of the analyses carried out on the database are simple descriptive frequencies and cross-tabulations of occurrences of events. In order to produce meaningful statistics, it is necessary to weight the sample so that it resembles the population from which it came. There is little value, in most cases, in producing a frequency distribution directly from the raw sample figures, because any events occurring in fatal or serious accidents will be over represented because of the higher sampling rates. Table 1 gives an example of the difference that can occur between raw and weighted sample data.

Table 1: Frequency distribution of delta-V: Comparison of weighted and unweighted data

Delta-V(kph)	Unweighted		Weighted	
	N	%	N	%
0-20	101	21.0	557	29.9
21-40	242	50.2	941	50.5
41-60	101	21.0	285	15.3
61-80	28	5.8	67	3.6
81+	10	2.1	15	0.5
TOTAL	482		1863	

It can be seen that low energy accidents are under represented in the unweighted sample, with the percentage of vehicles with a change in velocity of less than 20kph increasing from 21.0% to 29.9% when weighting factors are applied. High energy accidents with a change in velocity of over 40kph decrease correspondingly from 28.9% to 19.4%.

The weights which are applied to the data are calculated as 1/(sampling frequency) for each stratum; that is, there is a different weighting factor for each accident severity, county and quarter. Sample frequency counts can then be multiplied by the weighting factors to obtain a population estimate.

It should be emphasised that the estimate is of the number of towaway accidents involving cars of less than 6 years old in the sampled counties, not in Great Britain as a whole. Research is currently in progress on finding a valid method of estimating national figures from this sample; a probability based estimate would not be valid, because the sampled areas were not selected at random, but were chosen to represent a mix of urban and rural areas.

The weighting factors described above differ from weighting factors as described in statistics text books, and can only be used as a multiplicative factor for estimating population values. A different set of weighting factors are needed for calculating error estimates such as variances and confidence intervals. These weights are defined as

$$W_j = \frac{N_j}{N} \qquad (1)$$

where N_j is the population of stratum j
and N is the total population.

It can be seen that

$$\sum_{j=1}^{h} W_j = 1 \qquad (2)$$

where h = number of strata in the sample.

3.3 Estimates of sample variances

It is always beneficial to calculate tolerance limits for population estimates; this gives some indication of how reliable the estimate is. The formulae for variances (and therefore standard errors and confidence intervals) are much more complicated in a complex sample such as the ARU sample than in a sample obtained by simple random sampling. The estimates of variance tend to be approximate, as they rely on certain assumptions about the data; deficiencies in the sample frame such as the retrospective calculation of sampling frequencies, and the estimation of non-injury stratum populations, ought to be considered when calculating errors, but are not currently taken into account because of the complexity of the theory.
For a stratified population, the variance of a population estimation X is defined as,

$$\begin{aligned} \mathrm{var}(X) &= \mathrm{var}(Np) \\ &= N^2 \sum_{i=1}^{h} W_i^2 \, \frac{(1 - f_i)p_i\,(1 - p_i)}{n_i - 1} \qquad (3) \end{aligned}$$

where N = total number of members of population
p = proportion of population possessing characteristic of interest
h = number of strata in sample
W_i = weight of stratum
p_i = proportion of population in stratum possessing characteristic of interest
f_i = sampling frequency in stratum
n_i = number of sample elements in stratum

This formula is sufficient to give variance estimates for variables at the accident level (that is, where there is no additional variance caused by sampling clusters). However there are few variables at this level, and they are not of intrinsic interest apart from comparing them to national statistics. The variables available include time and date of accident and speed limit and road class at the accident scene.

Most population estimates of interest would be at the vehicle and occupant level; the variance equation for a population estimate takes into account the intracorrelations of elements within clusters. The variance equation for a population estimate in a clustered sample (not stratified) is given by :

$$\mathrm{var}(X) = \frac{N^2(1 - f)}{Z^2} \quad \frac{a}{(a - 1)} [\sum_{\alpha=1}^{a} p_\alpha Z_\alpha + p^2 \sum_{\alpha=1}^{a} Z_\alpha^2 \overset{a}{-} 2p \sum_{\alpha=1}^{a} p_\alpha Z_\alpha{}^2] \qquad (4)$$

where a = number of clusters
Z_α = number of elements in cluster
p_α = proportion of population in cluster possessing characteristic of interest
$Z = \sum_{\alpha=1}^{a} Z_\alpha$ = total number of sample elements
p = proportion of population possessing characteristic of interest
f = sampling frequency

To obtain the variance of a population estimate in a clustered stratified sample, the two equations (3 and 4) have to be combined such that,

$$\mathrm{var}(X) = \sum_{i=1}^{h} W_i{}^2 \mathrm{var}\,(X_i\ ) \qquad (5)$$

where var (X_i) is the variance of the clustered population estimate of stratum i

By contrast, the variance of a population estimate in a simple random sample is:

$$\mathrm{var}(X) = \frac{N^2(1 - f)p(1 - p)}{(n - 1)} \qquad (6)$$

3.4 Computing sample variances

It has been shown that the equation for calculating the variance of a population estimate is extremely complicated. The statistical package used by the ARU for data analysis is SPSS. It is not an easy task to calculate the variances using this package, as the main use of the statistical procedures is to aggregate over all cases; it has not been designed to easily manipulate numbers over strata or clusters. In order to calculate the standard errors of the population estimates for one variable taking five possible values, over 60 lines of computer code were needed. This variable was at the accident level so no clustering effects were present; variance calculations at the vehicle and occupant level would require even more code. The results of the calculations, demonstrating the magnitude of standard errors are given in Table 2. There are some computer packages available which are designed specifically for the analysis of complex survey data (see 10). The ARU were not aware of their existence when the SPSS computer system was developed; however, they may well prove to be superior tools for the analysis of the ARU data in the future.

Table 2: Distribution of speed limit in force at accident scene

Speed limit (mph)	N	Standard error	95% Confidence Interval
30	1606	75.4	1458-1754
40	552	48.6	457- 647
50	130	29.2	73- 187
60	938	66.5	808-1068
70	470	49.3	373- 567

It can be seen that the more frequent the occurrence of an event, the lower the standard error of the population estimate is, and the tighter the confidence interval. The standard errors range from about 5% to as much as 20% of the population estimate for the less frequent occurrences. This implies that interpretation of results from this database should be carried out with caution, particularly when analysing infrequent events. Analysing sub-samples of the data introduces further sampling errors; if the sub-sample does not coincide exactly with strata boundaries, weighting factors must be re-estimated to take account of the structure of the sub-population (see 5 for details of the methodology).

Table 3 shows the magnitude of standard errors for the same variable as the previous example, but for the sub-population of cars which rolled over in the accident.

Table 3: Distribution of speed limit for sub-population of vehicles which rolled over

Speed limit (mph)	N	Standard error	95% Confidence Interval
30	32	31.1	0- 93
40	41	12.5	16- 66
50	8	11.3	0- 30
60	132	41.0	52- 212
70	165	50.2	67- 263

The sub-population contains approximately 10% of the total population, with a very different distribution of speed limits; many more rollovers occur on roads with higher speed limits. The lowest standard error is approximately 30% of the population estimate, the highest is more than the estimate. It would appear from this that it is of little value to give population estimates at all for infrequent events in small sub-populations. The best that can be said is that the event does occur. The only exception to this is when analysing sub-populations of exclusively fatal cases, where the sampling frequency is one. In these cases, there is no sampling error, so the results can be interpreted with confidence.

It is not practical to carry out variance calculations for every population estimate made because of the time taken; error estimates are made only for important or controversial results.

To simplify the calculations, the ARU have decided to treat the vehicle level variables as though they were not clustered. The average number of vehicles per accident is only 1.13, and there seems to be some precedent for ignoring clustering effects when the average cluster size is less than about 1.2 (5, section 11.3A).

There are several methods of estimating various measures for both linear and non-linear estimates of the population parameter of interest in complex samples. These vary from the conceptually simple method of random groups, which involves taking several samples from the population and computing the sample variance amongst the several estimates of the parameter of interest, to highly complex mathematical estimates of variance for non-linear estimates based on approximating the non-linear function by a linear function using Taylor series. Six methods are described and explained clearly in Wolter (10). The drawback with the methods as far as the ARU database is concerned is that those relevent to the ARU sample structure tend to rely on the population being well defined. This is not the case with the ARU, as has been described in section 2.

3.5 Statistical inference

Analytical statistics measure relationships between variables; examples 3 are regression analysis, analysis of variance and discriminant assumption of simple random sampling, although this is often not explicitly stated. Large samples are almost invariably not simple random, but are complex designs such as clustered or stratified. Statistical analyses are frequently carried out ignoring the requirements of simple random sampling, because there are no other techniques available. The theory has not been developed, mainly because of the severe problems of distribution theory in complex samples. Statistics such as means and regression coefficients of probability samples are likely to be good estimates of the population values, but the standard formulae for error terms derived under the assumption of simple random sampling are likely to result in considerable underestimates. As in illustration, one US survey (2) estimated that variances in their complex sample were 2 to 3 times larger than those calculated under the assumption of simple random sampling.

The differences do not appear to be as large in the ARU's dataset; calculations have been carried out to estimate the standard error term for the population estimates for a variable under the assumptions of simple random sampling, and of stratified sampling. The results are shown in table 4.

Table 4. Comparison of standard error terms under the assumption of stratified and simple random sampling

Speed limit (mph)	N	Standard error (stratified)	Standard error (SRS)
30	1606	75.4	47.9
40	552	48.6	37.4
50	130	29.2	16.0
60	938	66.5	40.2
70	470	49.3	30.2

The variable used is the same as in previous examples: the estimate for number of vehicles involved in accidents in differing speed restrictions. The 'correct' standard error terms are between 1.3 and 1.8 times greater than those calculated under the assumption that the sample was simple random. The implication is that for more complex analyses of the data, where the statistical theory assumes a simple random sample, any error terms calculated ought to be at least doubled in order to take into account the complex nature of the sample; again caution is urged in the interpretation of the results.

Estimates of variance can be made using the methods described in the previous section; however, as already stated, because the population is not well defined, it is difficult to apply these methods to the ARU sample.

It is not recommended in the literature to carry out significance testing on large samples, especially when the null hypothesis is of zero difference, i.e. of no relationship. If the sample is large enough, then the weakest relationship can appear statistically significant if tested. It is more useful to concentrate instead on measuring the magnitude of relationships, together with measures of the variance of the estimates.

In the ARU database, the best that can be done is to use standard statistical theory, but to be aware that variances calculated are likely to be underestimates, because of the nature of the sample.

4. OTHER PUBLISHED SURVEYS

The National Crash Severity Study collected data on crashed cars and vans and their occupants between 1977 and 1979. The structure of the sample was very similar to the ARU sample - it was stratified by accident severity, and data at the vehicle and occupant levels were clustered.

Two reports were published (8, 9) giving frequency distributions of the variables investigated; there was no attempt at analysis, and the size of the sampling errors was not stated. A third report (2) gave very comprehensive details of all the assumptions behind the statistical theory, and did give estimates of sampling errors, and demonstrated how they were calculated.

Many other papers describing accident investigation results give no such background detail, and it is therefore difficult to appreciate how relevant the results given are to the accident population as a whole. Several papers exist (for example,3) which give details of case studies of a particular occurrence. While these are of intrinsic interest and value, it would be even more interesting to state how the case studies were selected, and approximately how often and under what circumstances the event would again occur. Some other papers (for example, 4) give details of the results of a sample survey, but give no information on the structure or representativeness of the sample. Again, it would be of more value to give some detail so that deductions could be made on the relevance of the results to other situations.

5. CONCLUSIONS

This paper has described the structure of the database on crashed car and occupant injury details; it is stratified by three variables and also clustered. The identification of the sample frame is imprecise, mainly because of difficulties in identifying towaway accidents. However, a good estimate of the complete population can be obtained retrospectively, so weighting factors can be calculated, in order to estimate the frequency of occurrence of accident events in the population. Research is currently in progress to identify techniques to estimate results for Great Britain from the population estimates, which currently represent only seven out of a total of 66 counties.

Variances, standard errors and confidence limits can be calculated for population estimates, but the formulae are complicated and only approximate. Error calculations are carried out only when results are marginal or controversial. It is not strictly correct to use analytical techniques such as regression analysis on the data because these techniques were developed using an assumption of simple random sampling. However, because there is no theory available to cope with complex samples, these methods are used with the understanding that error terms are likely to be underestimates.

Some published research papers in the accident field give details of sample structures, and how representative of a given population results may be, but others give little or no information. It is urged that all papers giving results of sample surveys publish this information so that readers can infer the relevance of results.

6. ACKNOWLEDGEMENTS

The author would like to thank all those who have made this study possible, in particular:

The sponsoring consortium:

Department of Transport
Vehicle Standards and Engineering
Vehicle Inspectorate
Transport and Road Research Laboratory
Rover Group
Ford Motor Company

Staff of the Accident Research Units at Birmingham and Loughborough
Department of Transport Vehicle Examiners
Chief Constables and Staff at the Police Forces in whose areas the study has taken place
The Consultants and Staff of Participating Hospitals
HM Coroners and their Officers

7. REFERENCES

1. BARNETT, V., Elements of Sampling Theory. **English Universities' Press Ltd**, 1974.

2. GIMOTTY, P.A. et al, Statistical analysis of the National Crash Severity Study Data. **Highway Safety Research Institute, Unversity of Michigan Report No.DOT-HS-805-561**, 1980.

3. GREEN, R.N. et al, Misuse of three-point occupant restraints in real-world collisions, **Proc. 1987 International IRCOBI Conference on the Biomechanics of Impacts**, pp.103-112.

4. HARTEMANN, F. et al, The characteristics of frontal impacts in real-world accidents. **Proc.Tenth International Technical Conference on experimental safety vehicles**, pp.424-431, 1985

5. KISH, L., Survey sampling. **John Wiley & Sons**, 1965.

6. MACKAY, G.M. et al, The methodology of in-depthd studies of car crashes in Britain. **SAE paper 850556**, 1985.

7. OTUBUSHIN, A., & GALER, M.D., Crashed vehicle examination techniques. **SAE paper 860372**, 1986.

8. RICCI, L.L., NCSS Statistics: passenger cars. **Highway Safety Research Institute, University of Michigan, Report No.DOT-HS-805-831**, 1980.

9. RICCI, L.L., NCSS Statistics: light trucks and vans. **Highway Safety Research Institute, University of Michigan, Report No. DOT-HS-805-...**, 1980.

10. WOLTER,K,M., Introduction to variance estimation. **Springer-Verlag**, 1985

11. YATES, F., Sampling methods for censuses and surveys. **Griffin**, 1971.