DATA COLLECTION

Background

The working party consisted of John Ryan (Chairman), Harold Clarke, Brian Hudson, Graham Lyons, Harry Reid and David Sanders.

It is not the intention and neither should it be the intention of this working party to provide draft material for a statistical plan for certain lines. Any such plan should be produced by groups of companies, the BIA and/or individual companies and in some parts of the world independent rationalising bodies such as ISO collect data for groups of companies. The purpose of this working party is to consider general principles and highlight possible areas of difficulty. We were mainly concerned with data for ratemaking purposes but it is obviously important to integrate these needs with data requirements for general management and underwriting administration purposes.

In terms of collection of data it is important not to be too ambitious, but on the other hand endeavour to arrange things so that items required can be collected at a later stage if so wished. Coding too many items creates problems with ensuring accuracy of data. The clerical staff who are prepared to code and punch substantial quantities of data for personal lines are limited in number and quality. If too stringent validity tests are incorporated into the company software systems, it may be found that either there are very high failure rates or alternatively staff find ways of bypassing the validity tests. In general, data input will be carried out by relatively low level grades of staff who, if they can find a way of short circuiting onerous validity tests, will do so. On the other hand, it is important to ensure the accuracy of the data. There is an important area of judgment to compromise between data accuracy, cost as well as staff acceptability of the tests.

In setting up a system for collecting and storing data, it is not usually possible to foresee all of the information which ideally could be required or the use which could be made of that data during the lifetime of the system. For example, if strict liability were to be introduced for certain classes of business, it might be very useful if historic claims had additional information recorded - e.g. the extent of the policyholders' liability. It is not feasible to collect all the information which might possibly be of use at some stage, but it may be possible to design the system with some degree of flexibility incorporated, so that additional data items may be collected

and handled at a later stage. It is also important not to jeopardise the accuracy of other data in order to obtain data that only may be required in the future. Staff can resent having to collect what they feel to be irrelevant items and this can affect their attitude to accuracy. Flexibility in system design can be a costly option however and may prove to be inadequate when needed. An alternative approach is to obtain supplemental data by way of special investigations, while ad hoc enquiries can prove to be expensive and also be a burden on staff who have already routine work to complete unless alternative resources are available.

Cost of collection of data is an important factor. Large volume and relatively high administrative expense ratio lines can justify very much more work in this area than can the smaller volume lines. This may well be in inverse proportion to the need for data collection. Products liability would, in an ideal world, have a much more complex data base than motor insurance but the expense loadings are less, though the potential for loss may well be very substantial. It is important to realise that the loss potential of a class of business is not necessarily directly related to its premium volume.

In general it is desirable to integrate data collection for ratemaking with management information systems supplemented by special calls and investigations. These principles apply particularly to company information systems but overambitious plans to collect data on a group basis can often mean that nothing is achieved whereas a plan to collect a limited amount of data with the flexibility to be upgraded at a later stage can often be more easily implemented especially if there is not general industry agreement, although the remarks made earlier should be borne in mind and the use of one-off reports is also a factor to be considered. However, upgrading at a later stage may be expensive, time consuming and have a lower priority than otherwise would be the case if everything was incorporated initially. In most cases it is more important to ensure the flexibility of the output and to be able to retrieve items from an existing data base than to be able to put extra items into the data base.

The coding system is an extremely important part of a data system. A hierarchal system such as the BIA uses for trade codes in its employers liability statistics scheme can be very flexible, although such a system may not be suited to every situation. However, the coding system also depends on the technology available.

For example, different considerations apply to the organisation of data when stored on disc as opposed to tape. Different approaches may be taken when inputting

data through VDUs as opposed to card or paper tape punching. For example, different considerations apply to the organisation of data when stored on disc as opposed to tape. Different approaches may be taken when inputting data through VDUs as opposed to card or proper tape punching.

Exposure

In devising any statistical plan for ratemaking, considerable care must be taken over defining exposure. A good definition of exposure is given by Dorweiler [1] as follows:

> "When critical conditions and injurable objects exist in such relationship that accidents may result there is said to be exposure."

In some lines it may be fairly obvious but even then caution is required. While in UK motor the car year is the usual measure, it is almost certainly not the best in terms actually measuring exposure. Some combination of mileage and mileage in urban areas is probably more accurate though much more difficult to administer.

For many lines it is usual to use the measure of exposure in order to calculate a definition of claim frequency. However, in some lines, such as treaty reinsurance, this will not be the case. Often then proxies such as premium income can be used which include measures of severity as well.

In some lines it may be desirable to use more than one measure of exposure, e.g. in household business both the number of houses and the sums insured need to be considered. Care also needs to be taken where more than one peril is covered in a policy; an example would be commercial package policies providing both property and liability coverages.

For example, while vehicle years is largely used for UK private car business, it is possible that mileage could be a better measure but it would be more difficult to administer. Another example could be man hours worked instead of payroll but man hours worked is not often readily recorded. These examples are not meant to imply that we believe that these other measures are better or worse as a measure of exposure. Clearly, this is an area where there is considerable scope for judgment but in many cases a specific research project can be carried out to see how much of the variance of the rating structure is accounted for by a less efficient proxy

and how much is accounted for by a more complex structure. For example, it may be possible to collect on a sample basis information concerning man hours worked rather than payroll for Employers Liability. It would then be possible to analyse the data to see if this explained more of the rating variation than payroll. There would be sufficient information available to come to an informed decision as to the appropriateness or otherwise of modifying the rating structure. This needs to be considered in connection with the classification system. This is an area in which further work needs to be carried out in respect of the theoretical principles of judging the adequacy of a classification system. One approach is to consider the variance explained by the rating system by comparison with the unexplained variation. Problems can arise with separating severity and frequency if they are not independent. However, a minority of members of the working party were sceptical as to whether this approach would produce results of any value. However, unless some analysis is made on these lines it is not possible to judge the adequacy or otherwise of a rating system.

For personal lines it will normally be necessary to collect the following items: policy issue data, lapse dates. There can often be considerable delays in processing these items and adjustment may have to be made to the data in order to calculate exposure accurately. Audit dates in employer's liability will need to be recorded to estimate retrospective premium adjustments.

Excesses are an area where there can be significant difficulty. Most broad groupings of excesses should be coded, though it appears that this is not necessarily the industry practice. Clearly, claims will not arise below the excess level. There are problems with claims padding by the insured should he feel the application of the excess to be unjust. There are also problems of heterogeneity and selection. There is, for example, evidence to suggest that that group of UK motorists with high voluntary excesses are a significantly different cross section of the portfolio from that of the remaining body of motorists.

Where different excesses apply to the same policy, it may not be sensible to code all excesses, e.g. windscreen damage excesses.

The above is not meant to be an exhaustive list of items to be coded and they will vary considerably from class to class and even company to company.

It is necessary to retain a history of exposure. Firstly to calculate exposure during a period at the end of that period - for example, on a domestic household policy,

the dates when the sum insured changed and the values of the sum insured both before and after the change will be needed. Further, a history of exposure is required for rating purposes. The length of the history and the detail retained are matters of judgment, though as electronic storage and microfilm facilities become cheaper in real terms, greater emphasis is going to be laid on storing past data. However, there will always be a need to compromise between the cost of storing all past data and grouping.

Note care must be taken when changes are made in recording and coding of exposure to check that past exposure is analysed correctly. Codes can be changed by staff interpretation as well as formal office edict and this needs to be considered. As an example, geographical codes may be reinterpreted by the staff when there is a change in the boundaries covered by regional managers.

Claim data

It is imperative for ratemaking purposes that claims data collected should be able to be related to the appropriate exposure. Consequently all claims information must contain sufficient information to relate it to the corresponding exposure items.

In respect of claim data, it is essential that the definition of a claim be clear. In an accident is each injured person to be considered the subject of a separate claim, in which case an occurrence coding may be required? Care must be taken in respect of the treatment of nil claims, especially in employer's liability. It is not known, of course, whether a claim is a nil claim or not until it is finally settled. However, individual companies' practices vary significantly in this area. For identical lines of business it is possible that one company will have a negligible number of nil claims whereas for another some 60% of claims may be nil claims, simply due to differing office practices. This problem is probably more acute in Employer's Liability than in other lines. This is particularly important in respect of grouped data from several companies where variation in the numbers of nil claims is substantial.

The general consensus of the working party was that usually case estimates provided much useful information and therefore usually should be collected and analysed. However, there is considerable variation in company practice in updating case estimates and in particular in "closing" a case estimate. It is not necessary and generally, on grounds of expense, not desirable to produce case estimates for all claims, espec-

ially small claims and those settled within a short period of notification. Much depends on the line of business, cost as well as day to day claims department administration. Some companies update their case estimates on a systematic basis with a formal programme though the frequency of update will vary company by company. Other companies only adjust estimates when there is some movement on the file. Thus considerable care needs to be taken when comparing aggregated case estimates.

An important sub-division is by type of claim - e.g. accidental damage or injury and possible claim settlement expenses. Consideration should also be given to whether only cumulative claim payments (by type) should be held or whether there is any advantage in holding details of each individual payment together with the dates on which it was made.

For many lines data should be collected in a form so that claims severity bands can be analysed. Exact claim amounts should usually be recorded so maximising flexibility in choice of bands. When grouping bands, it is important to allow for the "law of round numbers", i.e. if an estimate or settlement of a claim is more likely to be for £1,000 rather than £975 or £1,025. Hence when choosing bands £850 and £1,150 are much better band boundaries than £1,000.

For an individual company analysis of claims by size is largely a matter of data processing. For grouped data care needs to be taken to make sure that the limits for each company coincide. Care also needs to be taken when analysing grouped data due to possible heterogeneity. Duration to payment is also a factor to be considered.

There are special factors that apply to collective schemes such as those operated by the BIA. It is essential to have consistent definitions although there may be considerable heterogeneity amongst the data.

In many cases it will be desired to obtain additional data to refine classification systems whereas any one company's data may not have sufficient credibility, grouped data may well have. It is one of the fundamental actuarial problems to decide between credibility and heterogeneity.

In analysing data it is important to note correlations and in particular note whether they are reasonable or not. It is necessary to monitor correlations to see whether they continue to hold or not. In general, provided the correlations continue to hold, it does not matter from the insurance company point of view whether the

relations are causal or not, e.g. whether drivers of new cars as a class have more expensive claims or whether new cars are more expensive to repair. However, experience in the USA has indicated that causality can be a political issue.

Some information produced by collective schemes seems to be ignored by the industry. In particular, the age of the car appears to be negatively correlated with claim cost but does not appear to have any material acceptance as a rating criterion.

It is worth considering whether claims should be coded by peril. However, where the same peril is covered under several different policies, a per coverage coding should be considered as opposed to a per peril coding due to overlap problems and also reinstatement problems. It is important to realise that the combining average of different perils in one policy will not in general mean that the claims arising under that poolicy will be identical to the claims arising under separate policies covering the perils separately. This arises for a number of reasons such as (a) anti-selection, (b) a building that has been burned to the ground cannot be damaged by flood, (c) the existence of a single policy limit. As far as the UK is concerned, this is most likely to be a problem in small and medium sized industrial risks.

It is conceivable that more elaborate schemes than that currently used by the BIA could be derived for employers liability. However, experience rating modifications are significant and so reduce the need for class ratemaking criteria. Possibly also the lack of maturity of the data is a factor. However, the existence of experience rating as a market factor has a significant impact on industry data.

It should perhaps be asked, is the industry already collecting too much data in these lines? Does the industry need educating as to how it might use such data? Should experience rating techniques be more widely used? These seem to be much more important areas of principle than mere data collection techniques.

Ways of handling grouped data

Essentially there are two ways of handling grouped data:

(a) By merely aggregating, making some minor adjustments for differences in excesses, etc, i.e. adjusting individual claims by reducing to a common (the highest) excess level.

(b) By using an adjustment factor to convert to a common level, i.e. company A's experience is normally 90% of average.

Method (a) seems the more appropriate though the results need much greater care in interpretation, and the adjustments on some covers may be complicated, particularly different levels of excess. Adjustment should, where possible, be made for changes in contributors, at least by showing comparable data, i.e. for the least common denominator of contributors for both years. However, it may well be that the effects of differing portfolios more than offsets the impact of the adjustments under (a). Much depends on the type of business and the territory in which it is written and if care is not taken misleading results can arise.

Method (b) provides some interesting potential for further work and avoids many of the problems of method (a). The BIA has noted a considerable stability in the relationship of individual company data to the grouped data. This is probably an encouraging sign that grouped data trends are a valid prediction of company trends. Since even the largest companies would appear to have some uses for grouped data, albeit only for small parts of their business, this is an encouraging sign. Further work on the stability of such relationships, both from a theoretic viewpoint and from a pragmatic viewpoint, would seem desirable.

The above has been designed for personal lines. There are very different problems for individual risk rating and reinsurance. The efficiency and social justice of personal lines classification systems are factors that we have not considered. It is fairly obvious that a company which is otherwise "average" but does not write business in flood prone or subsidence prone areas will produce above average results. Should a large company orientate their efforts in this direction there would be market problems which are beyond the scope of this working party to consider.

Areas for future research

An area where further work is being done is in the products liability area where the BIA is developing a scheme. This is likely to be an area of considerable importance in the event of either the imposition of "reciprocal judgments" or of strict liability. It is an area of immense difficulty. Class codings are extremely difficult to interpret, e.g. a valve manufacturer may be selling to oil rig operators or to bicycle pump manufacturers.

Disease codings in employers liability are a somewhat troublesome area in relation to the year of occurrence. Clearly some method needs to be developed to allocate this correctly to ratemaking data and this will break the demand for the appropriate data. It is believed that market practice is to treat disease claims as "occurring" when they are first observed though there are problems with retired and redundant employees, particularly when there has been a change of insurer.

The paper to the Institute entitled "Compensation for Personal Injury" by G B Hey et al suggested that GISG should "examine the possibility of collecting some reliable data in case the insurance industry is faced with legislation. It might be best for a few large companies to get together, rather than for an attempt to be made to produce an industry-wide scheme." We considered this issue and decided it was a situation that the industry would be better reacting to at the time rather than collecting data in advance which may or may not be required. It was felt in this case that the possibility of Pearson being implemented was receding and that there were more urgent issues.

Further it could well be that simulation techniques were more appropriate in many similar circumstances. An assumption is made as to frequency and severity distributions and the convolution determined by simulation. Sensitivity should be tested due to parameter variation and distribution variation.

Furthermore, the industry showed that it was able to collect data for Pearson on a one-off basis indicating that information could be obtained should it become necessary. However, we would not wish to discourage individual companies pooling data and publishing results.

Experience rating produces problems in interpreting data but is an area where further work is required. Where a risk is experience rated less effort is going to be put into ascertaining the correct class rate.

Conclusion

Most data collection problems are going to involve the conflicting problems of credibility and heterogeneity. The counterplay of these two factors involves considerable judgment. It is in this counterplay that the actuary is uniquely qualified to advise. However, each case needs to be treated on its merits and this working party can

only advise on broad principles. After all, if this were not the case, it would be a simple matter to produce a manual to solve all problems.

We have only considered broad principles and, while there is no doubt as to the importance of a good data base, there are areas where special techniques can supplement shortage of data.

APPENDIX 1

The following is a list of some of the more obvious data sources in the UK. Additional suggestions are extremely welcome.

(i) BIA (contributors to schemes only)

(ii) Company schemes

(iii) Data supplied by reinsurers from their own portfolios

(iv) Large loss data published by the ROA

(v) Air claims

(vi) Fire Prevention Association

(vii) Institute of Actuaries : Mortality and Health

(vii) Government statistics

(ix) Fire Offices Committee (Fire Tariff members only)

APPENDIX 2

Data Collection - Reinsurance

A reinsurer ideally requires sufficient information for assessment of rates and for the determination of his overall exposure. In practice he has very limited information available for either of these purposes. The primary consideration is whether obtaining additional information would be cost-effective.

Rating

The reinsurer has had to develop techniques to use the information he has got and thus avoid the data collection problem. Rating methods are necessarily completely different from those of a direct office.

For proportional treaty business the reinsurer has a large number of treaties and a world-wide account. He would not have the time or resources available to consider the data in the same way as the direct office. Furthermore, the underwriting characteristics of the ceding office vary considerably and this is an important fact in the rating. The treaty is therefore considered as the unit of risk and the reinsurer requires a history of treaty results for the last few statistical years for each treaty. A reinsurer can always request more information than has usually been made available in the past. This is an increasing trend, though the current soft markets are retarding it. The only options are to continue or cancel a treaty and, if continuing, to negotiate on treaty share and commission and profit commission terms.

For most classes of business, some business is offered on a facultative basis and often a substantial number of reinsurers is involved. Apart from the leader the reinsurers can only agree to follow or decline. This must be done promptly with relatively little information. The prime consideration is often the reputation of the leader and additional information would be of no use as the business would be lost if time were taken to consider the information.

Different considerations apply to non-proportional business. Rates are usually assessed on a burning cost basis for excess of loss business where the burning cost is defined as the reinsurer's claims paid and outstanding, divided by the ceding company's premium income for the class of business considered. For some classes of business such as UK motor or UK third party where the reinsurer has available a large reason-

ably homogeneous portfolio guide rates can be produced by fitting curves to the claim payment or incurred claims data (log-normal and pareto curves are typical).

The effects of stability clause, inflation and claims development can be allowed for. A bigger problem is the lack of data for claims below the ceding companies' retentions which necessitates operating with truncated distributions and problems with assessing rates for low retentions. Though here also there is a slow trend towards reinsurers being given more information, which is also being affected by the soft markets.

Exposure

A major problem for a reinsurer is knowing the extent of his exposure. In each of Marine, Aviation and Fire Treaty business a number of treaties may include cover for the same risk (ship, plane or building, respectively). Limited information is available as to exposure to individual risks or to accumulations.

For example, for Fire in areas subject to natural (or man-made) catastrophes. This is one area where additional information on risks under a treaty could be beneficial, not least because it should enable a reduction in the rate charged for the reinsurer's own catastrophe protection if additional information on exposure was passed to the retrocessionaire. However, it is doubtful whether all ceding companies could be persuaded to supply this additional information, or even if it would be possible for them to do so. A good example is that of the Darwin windstorm where many reinsurers were affected in several layers where they were previously unaware of the potential exposure.

APPENDIX 3

ANALYSIS OF DATA

1.1 The reasons for collecting data are numerous, but the most important are:-

1) Filing of statutory returns to D.O.T.
2) Assessing adequacy of premium rates.
3) Internal management control.

1.2 The first question that could be raised is "Do we need data to assess levels of risk and reserve bases?" The answer to this question is, somewhat surprisingly NO! In many cases of reinsurance, for example, or coinsurance the office is not aware of level of risk it is underwriting, and assesses the rate on the goodwill and underwriting experience of the leading office. The only "data" it has is the fact that the principal office has generally got its rates correct in the past. The other example is the passing of the slip at Lloyds. This lack of data in assessing the adequacy of premium rates and for internal management control can be described reasonably adequately by the Theory of Games in Statistics.

1.3 Although decisions can be arrived at with little or no information via the Theory of Games, the more important decisions will occur when there is more statistical information available. The techniques then involve:-

a) Statistical inference, and
b) Multivariate analysis.

The basic problems with any form of statistical analysis is:-

1) Too little information; when little credibility can be given to the statistics provides, and
2) Too much information; when a general conclusion can be reached without observing underlying trends in the su strata of data, the analysis of the total data is expensive and time consuming and the information available for too large a sample is not that much different from that obtained by a substantially smaller sample.

tends to unitv independently of the actual population parameter, and the regression coefficients becoming increasingly unstable. The above assumes the population is Multinormal.

1.8 It can be seen that the statistical methods underlying the analysis of data is long winded and can lead to incorrect or inadequate conclusions The importance of powerful statistical techniques should not be under-estimated, nor should the significance attached to a result be over-estimated. An example of how statistical techniques could be applied in analysis of, for example, motor data is given below.

1.9 (i) First analyse the whole portfolio and derive estimates of the parameters which represent the whole class of business.

(ii) Subdivide portfolio into two (say). Do further analysis on data to obtain new estimates of parameters.

(iii) Question whether parameters differ.

(a) within two groups

(b) from initial portfolio.

(iv) Should any parameter be removed, or possibly a new one inserted. If the answer is yes revise (i) (ii) and (iii).

(v) Subdivide portfolio again and do analysis (ii) to (iv) again until portfolio becomes to small, or doubts arise to the significance of any result.

Finally check if overall models differ significantly from previous years models, or alternative models used in premium rating (for example).

1.10 It should be realised that the above process is a combined computatio /decision process and is not a task that can be solely undertaken by a computer program. Computers are useful in arriving at the rapid answer for a series of parameters, often with a best fit answer, but cannot tell what actual parameters are essential modelling the underlying portfolio experience.

1.11 It is clear that besides data collection we should be concerned with the use of the information for analysis, and the methods of analysing data. The mathematical approach of obtaining a bestfit model with no appreciation of underlying correlations and regressions is an imcomplete model. The statical techniques that could be used are to numerous to mention in this note, and a detailed description with

1.4 In General Insurance there is a fundamental problem underlying and statistical analysis; that is the uncertainty of the distribution or distributions underlying the experience. Indeed, the split of the distribution into Number of Claims (assumed in most classical textbooks as Poisson) and Size of Claims (Exponential, Gamma, Pareto) is often too simple, because subanalysis will lead to alternative compound Poisson parameters or distributions. The first principle underlying any initial statistical analysis is to make NO ASSUMPTION REGARDING THE DISTRIBUTION This leads automatically to the use of non parametric tests; in particular the classical theory of Multivariate Statistical Methods.

1.5 There are many problems associated with Multivariate Statistical Methods; the most important are

1) Incomplete analysis resulting in incorrect conclusions.
2) If too many parameters are sought, the correlations coefficients tend to unity (even when populations value to zero) and regression coefficient becomes unstable.

1.6 An example of the first case is using the statistic Actual / Expected as opposed to the Chi-squared parameter $\frac{(\text{Actual} - \text{Expected})^2}{\text{Expected}}$

The Continuous Mortality Investigations use the former parameter. Examples can be given whereby the Actual/Expected ratio looks reasonable, but the deviations about each side of the mean means that the Chi-Squared parameter gives a significant result. This is because a lot of informatic leads to laziness in completing statistical analysis.

1.7 The second case underlines the importance of not trying to analyse too many bits of information with to little information. For example, if the covenant coefficient is zero, then as the number of "independent" parameters approaches the sample size, the multiple correlation coefficient, expressed by:

$$E(R^2) = \frac{9}{N-1}$$

$$\text{and var } (R^2) = \frac{2(N-q-1)q}{(N^2-1)(N-1)}$$

where q = number of independent variates

N = sample size

applications to general insurance will take up several volumes. This note highlights a few of the problems of general techniques, and it is hoped that more detailed analysis of actual portfolios from the data collected will lead to a deeper understanding of mathematical/statistical models in general insurance.

David Sanders.
July, 1980.