STATISTICAL METHODS IN GENERAL INSURANCE

Philip J. Boland
National University of Ireland, Dublin
philip.j.boland@ucd.ie

The Actuarial profession appeals to many with excellent quantitative skills who aspire to "make financial sense of the future." However the road to qualification as an Actuary through the Institute or Faculty of Actuaries in the UK (or the Society of Actuaries or Casualty Actuary Society in the USA) is not an easy one, and a series of very challenging exams must be passed to qualify as a Fellow Actuary. These exams test many skills, and in particular demand a good knowledge of probability and statistics. The main areas of work for actuaries are traditionally life assurance, actuarial consultancy, general insurance, and investment. Although statistical skills are required in all of these areas, they are particularly important in general insurance. In this paper we discuss the basic tools and techniques in probability and statistics that are essential for an actuary who intends to work in general insurance.

INTRODUCTION TO THE ACTUARIAL PROFESSION

The actuarial profession is a relatively small one, but is generally highly regarded worldwide. Actuaries consistently rank among the most desirable careers in ratings of professions. Actuaries are normally very well paid, but they are expected to possess diverse quantitative skills and be able to apply them with wisdom and considerable skill in "making financial sense of the future"! The road to qualification for an actuarial trainee is however often a lengthy and arduous process. In the UK and Ireland, qualification as a "Fellow" Actuary is done either through the Institute of Actuaries (FIA) in London or the Faculty of Actuaries (FFA) in Edinburgh (www.actuaries.org.uk/), although the examination process is essentially the same for both professional bodies. The high standards of the Institute and Faculty of Actuaries are well recognized worldwide, and many students from Asia and Africa also qualify as actuaries through the Institute or Faculty. Qualification as an actuary in North America is normally done through either the Society of Actuaries (www.soa.org/) or the Casualty Actuarial Society (www.casact.org/), and these highly regarded organizations also demand superior quantitative skills to qualify as an actuary. In addition to possessing excellent technical skills, an actuary needs to be able to communicate effectively with customers and clients, IT specialists, and management. With all of these skills, it is no surprise that many actuaries eventually move into senior management roles.

Most actuaries work in life assurance, actuarial consultancy, investment, or general insurance, although all are expected to have a good knowledge of investment strategy. Life assurance companies provide pensions, life assurance and other financial services to enable customers to safeguard their long-term financial security. In actuarial consultancy companies, actuaries advise organizations on all aspects of employee benefits, in particular in setting up, calculating contributions and developing investment strategies to meet payments in pension schemes. Investment actuaries may specialize in managing funds, monitoring performance and advising on investment decisions. They may work for or advise investment banks, stockbrokers, or the investment and/or personnel departments of large companies. General insurance is perhaps the fastest growing area for actuaries, and it includes health insurance, personal insurance (such as home and motor insurance), as well as large commercial risks and employer liability.

ACTUARIAL EXAMINATIONS

The Institute/Faculty of Actuaries in the UK regularly revises their examination system with the most recent changes being introduced in 2005. Under this new examination system, in order to qualify as a Fellow Actuary, one must pass 14-15 challenging exams in four stages, be at least 23 years of age and satisfy a 3-year work experience requirement. Many of the exams require a sound and good understanding of both statistics and probability. In particular in the first (Core Technical) stage of the exams, there are two subjects (CT1 – Probability and Statistics, and CT6 – Statistical Methods), which require a good foundation in both statistics and probability.

The CT1 subject covers basic topics in probability and statistics (exploratory data analysis, concepts of probability, random variables and their basic properties, various probability generating functions and moments, standard discrete and continuous distributions, concepts of independence, bivariate distributions, sums of independent random variables, the central limit theorem and its applications, basic concepts of sampling and sampling distributions, estimators and confidence interval estimation, the basics of hypothesis testing, linear relations between variables using correlation and regression, basic concepts of analysis of variance, conditional expectation, and compound distributions with applications).

The subject CT6 assumes the basic knowledge obtained in CT3, and covers many concepts in applied probability and statistics that are particularly useful for actuaries working in general or non-life insurance. The topics include: Decision theory, Loss distributions, Risk Theory, Ruin theory, Bayesian statistics and Credibility, Concepts of rating, Generalized linear models, Time series methods and Monte Carlo simulation. We elaborate more on these particular topics in the CT6 subject, and indicate why they are important for the general insurance actuary.

STATISTICAL SKILLS FOR THE GENERAL INSURANCE ACTUARY

Decision Theory

An actuary will often be called upon to either give advice on, or actually make, decisions in the face of uncertainty. The action or strategy which the decision maker ultimately takes will of course depend on the criterion adopted for making a decision. In any given situation there may be several possible criteria to consider. Some situations may be viewed as games (with an intelligent and competitive opponent like a competing insurance company), while in others the opponent may be viewed as a non-competitive opponent which we term nature (like the future state of the economy which will have an effect on whether or not a new insurance product will be affordable or acceptable to the public). Familiarity with the basics of both decision and game theory will undoubtedly help in both understanding how people make decisions and why. A good understanding of zero-sum and variable-sum games can be useful, along with the concepts of minimax, pure and mixed strategies. Examples of variable-sum games where cooperation would be useful to both parties (as in the classical Prisoners Dilemma problem) are important in understanding the benefits of compromise. The general area of decision making under risk, using both the minimax and Bayes criterion situations where one tries to cope with the unknown state of (a noncompetitive) nature is also important. In some of these situations, the degree of uncertainty in nature might be reduced by experimentation or the collection of additional sample information (although this may often be done at a cost). Finally the concept of utility as an alternative value system to a strictly monetary one is crucial to the running of a successful business, whatever it may be.

Loss Distributions

Insurance is a data-driven industry, and insurance companies employ large numbers of analysts to understand claims data. No one likes to lose, and an actuary in particular needs to model both the frequency and size of losses and claims. Techniques in exploratory data analysis such as histograms, quantile plots, and summary statistics including sample estimates of skewness and kurtosis can be very useful tools in obtaining a feeling for the *typical claim size*. Relatively large claims, which may be infrequent, are of particular concern and hence the need to find and use distributions with relatively fat tails like the Pareto, Weibull and lognormal distributions. Although the empirical distribution function can be a useful tool in understanding claims data, there is often a natural desire to "fit" a probability distribution with reasonably tractable mathematical properties to claims data. Actuaries often make use of software (e.g., "BestFit") which tries to fit a myriad of classic distributions to a data set via various criteria such as Chi-square Goodness of Fit tests, the Kolmogorov Smirnoff and Anderson Darling tests, and the AIC (Aioke Information Criteria). For example the commercial package BestFit (www.palisadeeurope.com/bestfit/) claims to be the most popular distribution fitting software, and states that it is able to find "the distribution curve which best describes your data sets."

The claims actuary will also want to consider the impact of deductibles, reinsurance arrangements and inflation on that part of a claim which will be handled by the base insurance company. This involves a good understanding of conditional probabilities and distributions. For example, if X is a typical claim this year and inflation of size i is expected next year, then what is the distribution of $(1+i)X$? If the excess of any claim over M is to be handled by a reinsurer, what is the typical claim distribution for the base insurer?

Risk Models

Harold Cramer in 1930 stated that “The Object of the Theory of Risk is to give a mathematical analysis of the random fluctuations in an insurance business, and to discuss the various means of protection against their inconvenient effects.” The general insurance actuary needs to have an understanding of various models for the risk consisting of the total or aggregate amount of claims S payable by a company over a fixed period of time. Such models will inform the company and enable it to make decisions on amongst other things: expected profits, premium loadings, reserves necessary to ensure (with high probability) profitability, and the impact of reinsurance and deductibles.

In the collective risk model for S, one uses the random variable N to indicate the number of claims made, and write $S= X_1 + ... + X_N$ where X_i represents the amount of the i^{th} claim which is actually made in the time period being considered. In the collective risk model for aggregated claims, S has what is called a compound distribution. Another common model for the aggregate claims S is the individual risk model whereby $S = Y_1 + ... + Y_n$. Here n is the number of policies (in some cases this may coincide with the number of policyholders) in the portfolio under consideration, and Y_i is the random variable representing the claim amount arising from the i^{th} policy (or policyholder). Since in a short period of time normally only a small proportion of policies give rise to claims, most of the terms Y_i will be equal to 0. This is called the individual risk model for S since there is a term in the sum for each individual policy or policyholder. The actuary is required to have a good understanding of the statistical properties of both of these models for aggregated claims, in particular how they might be approximated and in some cases recursively calculated, as well as how they are effected by frequency and severity of claims. Furthermore it is important to know how various deductibles and reinsurance arrangements (proportional, excess of loss, and stop loss) affect claims payable.

Ruin Theory

If one lets $U(t)$ represent the net value of a portfolio of risks or policies at time t, then the actuary is certainly interested in studying the possible behaviour of $U(t)$ over time. In a technical sense one may say that ruin occurs if at some point t in the future, the net value of the portfolio becomes negative. The probability of this event is often called the “probability of ruin,” and it is often used as a measure of security. $U(t)$ will take into account relatively predictable quantities such as initial reserves U and premium income up to time t, but it also must take account of (claim) payments which are more variable and random in nature, and of course much harder to predict. Hence the need to study and understand stochastic models of the so called surplus process $\{U(t)\}_t$, which represents the surplus or net value of a portfolio of policies over time. In most cases it is not possible to give an explicit expression for the probability of ruin of a surplus process, however a classic inequality of Lundberg (1909) provides a useful upper bound, and the so-called adjustment coefficient provides a useful surrogate measure of security for such a process. Simulation can be a useful tool in estimating the probability of ruin, and in many complex cases is the only real way of getting a grasp on the probability of ruin. It is important for the actuary to investigate how the probability of ruin in a surplus process (in both finite and infinite time) is affected by factors such as the premium rate, initial reserves U, the typical claim X, the claim arrival rate $\lambda(t)$, and various levels and types of reinsurance.

The Bayesian Approach to Statistics and Credibility Theory

Credibility theory in general insurance is essentially a form of experience-rating, which attempts to use the data in hand as well as the experience of others in determining rates and premiums. Often an actuary has to estimate expected future claim numbers and/or total aggregate claims for a portfolio of policies on the basis of rather limited sample or current information x, but where other collateral information is also at hand. Let us assume there is crucial parameter θ of interest, which for example may be the annual claim rate or a related expected aggregate claims total. Often there is other collateral or prior information from business or portfolios of a similar nature, which might be useful in estimating θ. Let us denote by θ_s an estimate of θ based on the sample information *x*, and by θ_c an estimate of θ based on the available collateral information. In the situation where θ is a mean, θ_s might be the sample mean of *x* and θ_c some prior estimate (say μ_0) of this mean. A key question is often "How might we combine the two (sample and collateral) sources of information to get a good estimate of θ, and in particular how much weight or credibility *Z* should our estimate put on the sample estimator θ_s?" Clearly the Bayesian framework for statistics comes into play in the combination of these sources of information. Surely the value of *Z* should both be an increasing function of the amount of sample information acquired over time, and it should take account of the relative values of the sample and collateral information available.

A credibility estimate of θ is a linear combination of the sample estimator θ_s and the collateral estimate θ_c of the form $Z\,\theta_s + (1\text{-}Z)\,\theta_c$ where *Z* is the "credibility" we put on the sample estimator θ_s. This general expression is often called the "credibility premium formula." Traditionally there has been an emphasis on only using estimates θ_s which are linear in the observations in the credibility premium formula, and although such estimates have considerable appeal there is no theoretical reason why other sample estimates cannot be used.

Generalized Linear Models

Modelling relationships between various observations (responses) and variables is the essence of most statistical research and analysis. Constructing interpretable models for connecting (or linking) such responses to variables (which may be of a nominal, ordinal or interval (continuous) nature) can often give one much added insight into the complexity of the relationship that may often be hidden in a huge amount of data. Letting *Y* be an observation where *Y* is a member of the exponential family of distributions, the general insurance actuary must have an understanding of the relationship between $\mu_i = E(Y_i)$ and other explanatory variables $x_{i1}, x_{i2}, \ldots x_{ip}$ through a linear predictor $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$, where $g(\mu_i) = \eta_i$.

Are the numbers of automobile claims made by a driver related to age, education, gender, type of vehicle, engine size and daily usage? In what way is the size of an employer liability claim related to the personal characteristics of the employee (age, gender, salary) and the working environment (safety standards, hours of work, promotional prospects)? We might want to study how the number of claims made by an insured individual depends on various explanatory variables or predictors, or how the number of accidents at an intersection depends on weather, traffic intensity, hour of the day, or day of the week. One might consider a logit model if we are interested in modelling how the probability of an individual developing a critical illness or having a car accident is related to age, gender and various other health or risk factors. Logistic regression may be used to identify variables that are predictive of how long a customer stays with a company. In mortality studies, the initial rate of mortality q_x at age *x* has been modelled by logistic regression. It is clear that generalized linear models are now a fundamental tool in the arsenal of a general insurance actuary.

Time Series Methods

Any actuary must keep abreast of investment and economic trends, and therefore should be familiar and comfortable with the basic concepts of univariate time series including stationarity, autoregressive and moving average models as well as the more general ARIMA models. A knowledge of the basic theory of random walks and co-integration of times series is also essential together with an ability to apply these concepts to investment models.

Concepts of Rating

In general insurance, claims due to physical damage (to a vehicle or building) or theft are often reported and settled reasonably quickly. However in other areas of general insurance, there may be considerable delay between the time of a claim inducing event and the determination of the actual amount the company will have to pay in settlement. When an incident leading to a claim occurs, it may not be reported for some time. In employer liability insurance, the exposure of an employee to a dangerous or toxic substance may not be discovered for a considerable amount of time. In the case of an accident the incident may be quickly reported, but it may be a considerable amount of time before it is determined actually who is liable and to what extent.

Clearly an insurance company needs to know on a regular basis how much it should be setting aside in reserves in order to handle claims arising from incidents that have already occurred, but for which it does not yet know the full extent of its liability. Claims arising from incidents which have already occurred but which have not been reported to the insurer are termed IBNR (incurred but not reported) claims. Claims which have been reported but for which a final settlement has not been determined are called Outstanding. Claims reserving is a challenging exercise in general insurance, and one should never underestimate the knowledge and intuition that an experienced claims adjuster uses in establishing reserves and estimating ultimate losses. However mathematical models and techniques can also be very useful, and give the added advantage of laying a basis for simulation.

In order to give a flavor for the type of problem one is trying to address in claims reserving, consider the triangular representation of cumulative incurred claims given in Table 1 for a household contents insurance portfolio. The origin year refers to the year in which the incident giving rise to a claim occurred, and the development year refers to the delay in reporting relative to the origin year. For example, incremental claims of 128,561-101,892 = 26,669 were made in 2004 in respect of claims originating in 2002 (and hence delayed 2 years). This type of triangular representation of the claims experience is often called a delay triangle. There are many questions which a company would like answered with respect information of this type, but certainly one would be to determine what reserves it should set-aside at the end of 2005 to handle forthcoming claim payments in respect of incidents originating in the period 2001-2005. In brief, how might one run-off this triangle?

Table 1: Cumulative Incurred Claims in a Household Contents Insurance Portfolio

	Development Year				
Origin Year	0	1	2	3	4
2001	39,890	85,160	108,465	116,910	124,588
2002	47,597	101,892	128,561	138,538	
2003	50,230	105,962	132,952		
2004	51,423	108,390			
2005	54,567				

Of course any good analyst would question the quality of the available data, and make use of any additional information at hand. For example, in this situation is it fair to assume that all claims will be settled by the end of the 4th development year for any origin year? Can we make the assumption that the way in which claims develop is roughly similar for those originating in different years? Should inflation be taken into account? Is there information at hand with respect to the number of claims reported in each of these years (is there a delay triangle for reported claim numbers)? What other knowledge have we about losses incurred in the past (for example with respect to premium payments) for this type of business? In many situations it is best to try several methods to get a reasonable overall estimate of the reserves which should be held.

One of the most frequently used techniques for estimating reserves is the chain ladder method. In this method one looks at how claims arising from different origin years have developed over subsequent years, and then use relevant ratios to predict how future claims from these years will evolve. The question of how to deal with past and future inflation in estimating reserves must be considered. The average cost per claim method is a popular tool which takes account of the numbers of claims reported. The Bornhuetter-Ferguson method uses additional information such as loss ratios (losses relative to premiums) together with the chain ladder technique to estimate necessary reserves. All of these techniques are quite deterministic in nature, but one may also consider statistical models which would allow one to evaluate fitness, variability and basic assumptions better.

No Claim Discount (NCD) systems (sometimes also called Bonus-Malus systems) are experience-rating systems which are commonly used in motor insurance. NCD schemes represent an attempt to categorize policyholders into relatively homogeneous risk groups who pay premiums relative to their claims experience. Those who have made few claims in recent years are rewarded with discounts on their initial premium, and hence are enticed to stay with the company. Depending on the rules in the scheme, new policyholders may be required to pay the full premium initially and then will obtain discounts in the future as a result of claim free years. The general insurance actuary modelling an NCD scheme would frequently use Markov chain methods to investigate how premiums and movements take place over time.

Monte Carlo Simulation

Computer simulation is commonly used in many areas of general insurance. It may be used to model claim incidence and size. It is often used to model both claim size and numbers for different models in risk theory, and to estimate the probability of ruin for different premium ratings and reserves. The simulation of time series models may be useful in estimating losses over time. Simulation can also be a useful tool in studying convergence and stability in various experience rating systems such as no claim discount schemes.

CONCLUSION

The general insurance actuary needs to know the essentials of decision and game theory to compete in the market of general insurance. An understanding of probability and statistical distributions is necessary to absorb and evaluate risk and ruin when balancing claims, reserves and premiums. In introducing and developing new products, credibility theory and Bayesian statistics play a role in evaluating sample and collateral information. Markov chains are important in predicting the success of rating methods, including NCD systems. Generalized Linear models are essential tools in finding risk factors for premiums calculations. Time series methods are used in various ways to predict trends, and simulation methods are crucial to understanding the many models considered for anything from new products to revisions in rating schemes. Is it no wonder that the general insurance actuary must be a practicing statistician!

REFERENCES

Boland, P. J. (to appear 2006). *Statistical Methods in Insurance and Actuarial Science*.

Bowers, N., Gerber, H., Hickman, J. Jones, D. and Nesbitt, C. (1986). *Actuarial Mathematics*, Schaumburg, IL: Society of Actuaries.

Buhlmann, H. and Gisler, A. (2005). *A Course in Credibility Theory and its Applications*. New York: Springer.

Cramer, H. (1930). On the mathematical theory of risk. *Skandia Jubilee Volume*, Stockholm.

Dickson, D. C. (2005). *Insurance Risk and Ruin*. Cambridge International Series on Actuarial Science.

Herzog, T. N. (1999). *Introduction to Credibility Theory*. Winsted, CT: ACTEX Publications.

Hossack, I. B., Pollard, J. H., and Zehnwirth, B. (1999). *Introductory Statistics with Applications in General Insurance* (2nd edition). Cambridge University Press.

Klugman, S. A., Panjer, H. H. and Willmot, G. E. (1998). *Loss Models: From Data to Decisions*. New York: Wiley Series in Probability and Statistics.

McMahon, D. and Boland, P. J. (to appear 2006). Qualification Time for an Actuary. *The Actuary*.