# STATISTICS IN INSURANCE

*Steve Drekic*
Department of Statistics and Actuarial Science
University of Waterloo, Ontario, Canada, N2L 3G1

In attempting to analyze insurance losses arising in connection with health coverages as well as property and casualty insurance situations involving homeowner and automobile coverages, it is imperative to understand that a portfolio of insurance business is very complicated in terms of the nature of its past and future risk-based behaviour. There are many deterministic and stochastic influences at play, and the precise prediction of the future claims experience necessitates that all such influences and their effects be identified. The role of probability and statistics is vitally important in this regard, not only in terms of providing the required statistical methodology to properly analyze any data collected by the business, but also in assessing whether a quantitative (i.e. theoretical) model is able to accurately predict the claims experience of a portfolio of insurance business.

First of all, in situations when the underlying data are very extensive and have been collected in the most appropriate form for its intended purpose, it is indeed possible to answer many of the questions which arise in general insurance using observed claim size distributions and/or observed claim counts. However, it is quite often the case that data are far from extensive and may not actually be in the most convenient form for analysis. In such circumstances, calculations are only possible if certain (mathematical) assumptions are made. In other words, a quantitative model is formulated involving the use of theoretical probability distributions. Moreover, even in situations where the data are extensive, the use of theoretical distributions may still be essential. Several reasons emphasizing the importance of their use include:

1. knowledge of their convenient and established properties, which facilitate the analysis of many problems,

2. the fact that the distribution is completely summarized by a relatively small number of parameters (which characterize its location, spread, and shape) and it is not necessary to work with a lengthy set of observed data,

3. the fact that they enable one to make statistical inferences concerning the behaviour of insurance portfolios,

4. their tractability in terms of mathematical manipulation, permitting the development of useful theoretical results.

The Central Limit Theorem justifies why normal distributions play such an important role in statistics. In particular, the well-known law of large numbers is employed in the literature on risk management and insurance to explain pooling of losses as an insurance mechanism. For most classes of general insurance, the claim size distribution is markedly skew with a long tail to the right. If an

insurer were to experience a large number of claims with respect to a particular block of business, its total payout (i.e. aggregate claims) might, however, be expected to be approximately normal distributed, being the sum of a large number of individual claims. This assumption is certainly reasonable for many purposes. There may, however, be problems associated with the extreme tails of the distribution, and these tails are particularly important for reinsurance purposes. Serious consequences could result from an insurance business basing financial risk management decisions on a model which understates the probability and scope of large losses. As a result, other parametric models, such as the gamma, log-normal, and Pareto distributions, are often much better suited to capture the positively skewed nature of the claim size distribution, and would therefore be much safer to use for estimating reinsurance premiums with regard to very large claims.

The most common and certainly best known of the claim frequency models used in practice is the Poisson distribution. In particular, the compound Poisson model for aggregate claims is far and away the most tractable analytically of all the compound models, as it is useful in a wide variety of insurance applications. It is also consistent with various theoretical considerations including the notion of infinite divisibility, which has practical implications in relation to the subdivision of insurance portfolios and business growth. On the other hand, the Poisson model inherently assumes that the individual risks within a portfolio of business are homogeneous from the point of view of risk characteristics, and this unfortunately leads to an inadequate fit to insurance data in some coverages. Consequently, perhaps the most important application of the negative binomial distribution, as far as general insurance applications are concerned, is in connection with the distribution of claim frequencies when the risks are heterogeneous, providing a significantly improved fit to that of the Poisson distribution.

In reference to the probability models above, it is also critical to realize that the parameters of a distribution are seldom known a priori. As a result, they need to be estimated from claims data before the distribution can be applied to a particular problem. Oftentimes, several different functions of the observed data will suggest themselves as possible estimators, and one needs to decide which one to use. The following criteria provide a good basis for determination:

1. the estimator should be *unbiased*, so that its expectation is equal to the true value of the parameter,

2. the estimator should be *consistent*, so that for an estimate based on a large number of observations, there is a remote probability that its value will differ seriously from the true value of the parameter,

3. the estimator should be *efficient*, so that its variance is minimal.

Statisticians have developed a variety of different procedures for obtaining point estimates of parameters, including the method of moments, least squares, and maximum likelihood. In simple situations, the various methods often produce identical results. When sample sizes are large, they all tend to provide more

or less the same answers, even in more complicated cases. In other instances, however, markedly different results can emerge, and the three criteria above are frequently used by risk practitioners in deciding which estimator to use for a given insurance application.

In conclusion, thorough treatments of these topics can be found in several reference texts including [1], [2], [3], [4], [5], [6], and [7].

# References


[1] Boland, P.J. (2007). *Statistical Methods in Insurance and Actuarial Science*. Chapman & Hall/CRC, Boca Raton.

[2] Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A. and Nesbitt, C.J. (1997). *Actuarial Mathematics, Second Edition*. Society of Actuaries, Schaumburg.

[3] Daykin, C.D., Pentikäinen, T. and Pesonen, M. (1994). *Practical Risk Theory for Actuaries*. Chapman & Hall, London.

[4] Dickson, D.C.M. (2005). *Insurance Risk and Ruin*. Cambridge University Press, Cambridge.

[5] Hossack, I.B., Pollard, J.H. and Zehnwirth, B. (1999). *Introductory Statistics with Applications in General Insurance, Second Edition*. Cambridge University Press, Cambridge.

[6] Kaas, R., Goovaerts, M.J., Dhaene, J. and Denuit, M. (2001). *Modern Actuarial Risk Theory*. Kluwer Academic Publishers, Dordrecht.

[7] Klugman, S.A., Panjer, H.H. and Willmot, G.E. (2008). *Loss Models: From Data to Decisions, Third Edition*. John Wiley & Sons, New York.