Ipsos MediaCT
The Media, Content and Technology Research Specialists

Probability Modelling

Bite Sized White Paper | **2012**

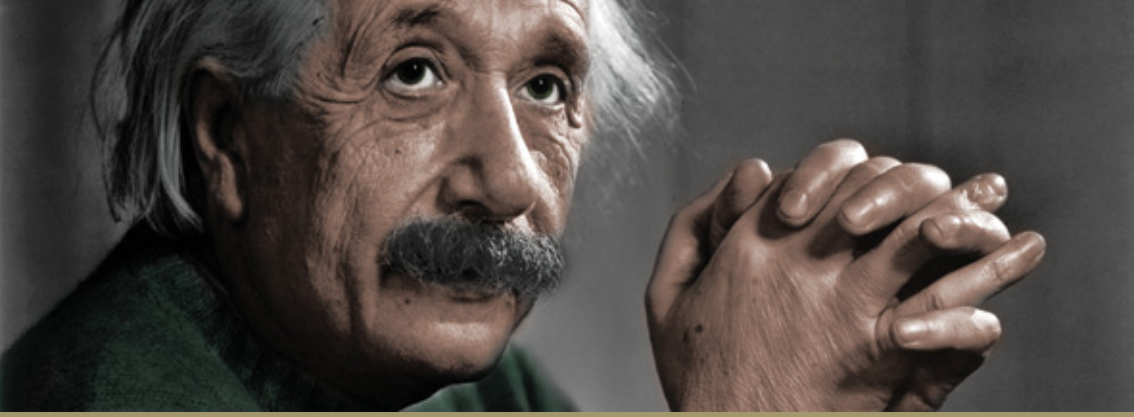
Introduction

The actor and director Sidney Poitier is quoted as saying "**So much of life, it seems to me, is determined by pure randomness.**" Many have struggled to come to terms with this situation, using strategies ranging from denial ("God does not play dice" – Albert Einstein), through the acceptance shown by Poitier, to analysis. Gamblers noted long ago that the random outcomes of dice and cards in fact exhibit patterns. An early example was the impecunious Renaissance mathematician Gerolamo Cardano, who generously published his findings in *Liber de Ludo Aleae* (Book on Games of Chance).

Probability models are formulaic representations of random phenomena and as such have wide applicability in media research. Our data acquire randomness from several causes:

- Variation in behaviour of a given population member over time (such as number of cinema visits, or number of billboards seen)
- Variation in behaviour across the population
- Fuzzy relationships between variables
- Divergence of the sample from the population.

For simple analysis such as estimating a proportion (e.g. media reach), we don't actually need to describe all these sources of variation via a model. But suppose we want to obtain greater insight, such as what reach would be over a longer time period, or what the inter-relationships



between different variables are. To do this, we need a model. A useful definition of a probability model is '**a description of the data-generation mechanism**'. This is akin to 'The Truth' that the now-notorious 'quants' on Wall Street sought to describe the behaviour of stock markets. But unlike them, we never forget that stationarity of behaviour over time is no more than a valuable simplifying assumption that is never quite true and sometimes violated. No model could predict the timing of Rupert Murdoch's launch of The Sun on Sunday or the impact it will have on the market.

Two Types of Model

Just like models of the natural world and in economics, we can model at either the macro or micro level. Newton's law of gravity is a macro model – it describes the how rather than the why. Einstein's successors work at the particle level in their continuing search for the why.

Alternative terms sometimes used in media research are functional model (if macro) and respondent-probability model (if micro). As in physics, early media models were macro in nature, partly due to the lesser computational load. Examples are the Sainsbury, Agostini and Metheringham formulae for print reach, and the Copland model for outdoor. But today the emphasis is firmly on micro-modelling. This includes for example regression and factor models that analyse relationships between variables using respondent-level data. Let's look at some other examples in the two media already mentioned.

Readership

Ipsos MediaCT measures the reading habits of the Great Britain population on behalf of NRS Ltd. for the National Readership Survey.

Topline reading data are reported from the binary data collected (One count for each positive reading claim). However there are other reporting requirements that require the binary data to be converted to probabilities. These include use of the data in reach and frequency tools and the ability to combine the NRS survey data with external data sources.

The NRS data collected includes socio-economic, demographic and lifestyle information about the readers of various publications, as well as what topics they are interested in reading about and how frequently they read a publication. This rich variable source enables us to use a data mining technique to calculate a probability of reading for all those respondents who claim to have read the publication in the past year (RPY).

Those variables which best discriminate between readers and non-readers of a publication are selected to build a classification tree for each publication. The result is a clustering of RPY respondents (ranging between 6 and 159 clusters, depending on the publication) each with its own probability of reading.

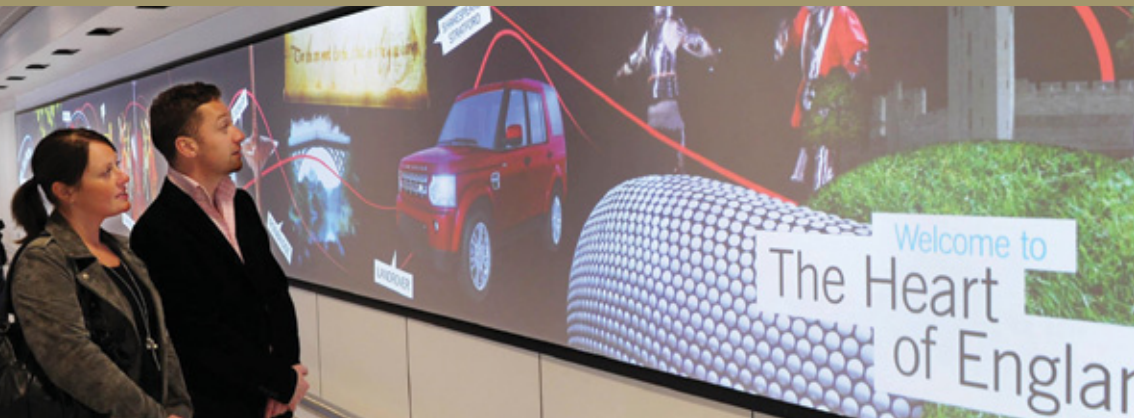
Finally, as recency of reading is also collected in the survey, a technique known as simulated annealing is applied to the probability so that, once expanded, the weekly, monthly and quarterly reach are as close as possible to the corresponding recent reading claims from the survey.

Out of Home

In 2008, POSTAR, the JIC for the Out of Home sector, commissioned Ipsos MediaCT to provide a new continuous audience measurement system. A specific requirement was to model the data to produce coverage and frequency results that span any combination of the measured media segments (roadside, bus-sides, tube and rail, malls etc.) and formats (6-sheet to 96-sheet billboards, dynamic and digital frames etc.), as well as for campaigns beyond the survey period.

Opportunities to see a poster are generated by tracking the location, direction and speed of travel of a sample of respondents using MobiTest GPS devices. For roadside, these data are related to the location of the frames, generating actual contacts over nine days for each respondent. For each indoor segment such as stations and malls, a probability model describes how people navigate their way and Monte-Carlo sampling is used to generate passages and hence contacts. For bus-sides, route and frequency data are used.

The contact database is then input to the main probability modelling process. The first stage is to acknowledge that journeys either start or end at home, so home location is a powerful predictor of which frames are contacted. This relationship is modelled to predict which journeys might be made after the ninth day. This part of the model is used as a data-generation mechanism, from which virtual contacts are generated. This expansion of the database (by a factor of about 10) helps improve the granularity in analysis. Then the probability for each actual and



virtual contact is estimated in a unified way, using the NBD (negative binomial distribution) within a Bayesian framework. The probabilities are appended to each contact record and from this point on standard tools can be used to provide reach and frequency in a flexible manner.

In conclusion

We think these examples demonstrate that the real value of probability models in survey research is **to extract the maximum information from the data**. We see three benefits in this. First, the mere process of fitting a model requires exploration of the data which is itself beneficial to understanding what the data are trying to tell us. Second, as our examples demonstrate, a good model then has many applications. Third, it has a life beyond the survey: it becomes acquired knowledge and intellectual capital which can be brought to bear on future research.

For more information, please contact:

Sandra Collins
t: +44(0)20 8861 5515
e: sandra.collins@ipsos.com
www.ipsos-mori.com

Trevor Sharot
t: +44(0)20 8861 8217
e: trevor.sharot@ipsos.com
www.ipsos-mori.com

Ipsos MediaCT - Winner - 2011 MRS Awards: "Best Innovation"