MODELLING READERSHIP CORRELATIONS

Gilles Santini, G.S. IT Services

1. Introduction

One of the conclusions of the paper entitled "*Revisiting mediaplanning models assumptions*" that we gave at the Cambridge Readership symposium in 2003, was the necessity to develop ways to provide relevant models accounting for the correlations which exist between different issues readership.

The motivation behind this demand was to fill a major gap of mediaplanning models that ignore such correlations at personal probabilities level and weaken reach and frequency calculations.

Also if one could solve the problem in the simple case of print media it would open a way for a better handling of cross-mediaplanning.

The difficulty was three folds:

- A lack of adequate bivariate distributions to describe those correlations
- The complexity of their integration within existing mediaplanning models
- Critical data processing speed issues

Since 2003 major statistical theory advances have happened allowing for a better understanding and theorisation of bivariate beta distributions and several research works are available now.

2. New advances

In a recent paper published in 2005 [1], J. Danaher and B. Hardie from the Marketing Department of the University of Auckland, have studied how those new developments could be used for modelling exposure to correlated magazines. After some analytic developments they conclude that the observed overestimation that results from neglecting such a correlation "*would result in campaign effectiveness expectations (such as sales) being much higher than would actually be observed"*.

Using a statistical framework introduced by Sarmanov in 1966 they introduce a bivariate beta distribution to model the joint probability of exposition to two magazines.

This distribution has the following form:

$$g(p_1, p_2) = f(p_1 \mid \alpha_1, \beta_1) f(p_2 \mid \alpha_2, \beta_2)\left[1 + \omega(p_1 - \mu_1)(p_2 - \mu_2)\right] \quad (1)$$

Where $f(p_1 \mid \alpha_1, \beta_1)$ and $f(p_2 \mid \alpha_2, \beta_2)$ are the usual marginal Beta distributions of each magazine with mean equal to the audience level : $\mu_1 = \frac{\alpha_1}{\alpha_1 + \beta_1}$ and $\mu_2 = \frac{\alpha_2}{\alpha_2 + \beta_2}$.

The last term, within square brackets, is a mixing function that carries the dependence of the two magazines exposition.

The ω parameter which appears in this term is directly linked to the correlation of the two probabilities of exposure. It is easy to show that $corr(p_1, p_2) = \omega\sigma_1\sigma_2$ where σ_1 and σ_2 are the known variances of the two marginal distributions. If $\omega = 0$ we get the independence case.

The parameters (α, β) can be estimated the usual way for each magazine.

For each magazine this lead separately to a Beta-Binomial distribution of the frequency of exposures :

$$P_{BB}\{X = k \mid \alpha, \beta\} = \binom{n}{k} \frac{\alpha^{\overline{k}} \beta^{\overline{n-k}}}{(\alpha+\beta)^{\overline{n}}} \tag{2}$$

In the case of two correlated magazines one gets :

$$\Pr\{X_1 = k_1, X_2 = k_2\} = P_{BB}\{X_1 = k_1\} P_{BB}\{X_2 = k_2\} \left[1 + corr(X_1, X_2) \frac{[k_1 - E(X_1)][k_2 - E(X_2)]}{\sqrt{\mathrm{var}(X_1)\,\mathrm{var}(X_2)}} \right] \tag{3}$$

This is a clearly generalizes the notion of Beta binomial to the two variables cases.

Danaher and Hardie' paper opens the way to a new advance in media planning but unfortunately their work is restricted to the case of two magazines.

Also being based on a global modelling, by opposition to a respondent level modelling, their model does not guarantee consistency of the results across several sub-targets groups.

What one really needs is a generalization of the Full Binomial model that most media planning models implement nowadays.

We have personally been working on this for several years without much success but the above work has finally pointed us in the right the direction.

3. Extending the Full-Binomial

Let i be any respondent in the sample with a probability to read a magazine equal to $p(i)$.

The Full-Binomial model makes the assumption that several issues of a magazine are independent.

So the number of exposures of a given respondent i after n issues can be considered as the sum of n independent Bernoulli random variables $x^{(1)}, x^{(2)}, \cdots, x^{(n)}$ taking the value 1 or 0 (read/don't read) with probability $p(i)$ and $1 - p(i)$ respectively.

It follows that X is distributed as a Binomial law:

$$\Pr\{X = k \mid n\} = \Pr\{x^{(1)} + x^{(2)} + \cdots + x^{(n)} = k\} = \binom{n}{k} p(i)^k \left(1 - p(i)\right)^{n-k} \tag{4}$$

For n_1 issues of magazine 1, n_2 issues of magazine 2, … , n_M issues of magazine M one gets the probability of k exposures as :

$$\Pr\{Z = X_1 + X_2 + \cdots + X_M = k \mid n_1, n_2, \cdots n_M\} \tag{5}$$

The Full-Binomial postulates that two issues of two different magazines are independent and writes:

$$\Pr\{x_1 = 1, x_2 = 1\} = p_1(i)\, p_2(i) \tag{6}$$

This additional assumption enables us to easily compute the quantity (5) by convolution and produce the requested result in other words the frequency distribution of a media schedule.

To extend the Full-Binomial one need to relax the independence assumption (6).

We propose to add a mixing factor and write:

$$\Pr\{x_1 = 1, x_2 = 1\} = p_1(i)\, p_2(i) \left[1 + \rho \varphi_1(i) \psi_1(i)\right] \tag{7}$$

This is similar to what is done in correspondence analysis as explained in [2].

In that context one also writes :

$$\Pr\{x_1 = 0, x_2 = 0\} = (1 - p_1(i))(1 - p_2(i))[1 + \rho\varphi_0(i)\psi_0(i)] \tag{7b}$$

Where ρ is an unknown parameter and (φ_1, φ_0) and (ψ_1, ψ_0) four quantities that should obey the following normalizing conditions:

$$\begin{cases} p_1\varphi_1 + (1 - p_1)\varphi_0 = 0 \\ p_1\varphi_1^2 + (1 - p_1)\varphi_0^2 = 1 \end{cases} \qquad \begin{cases} p_2\psi_1 + (1 - p_2)\psi_0 = 0 \\ p_2\psi_1^2 + (1 - p_2)\psi_0^2 = 1 \end{cases} \tag{8}$$

Solving (8) leads to :

$$\Pr\{x_1 = 1, x_2 = 1\} = p_1(i)\, p_2(i)\left[1 + \rho\sqrt{\frac{(1 - p_1(i))}{p_1(i)}}\sqrt{\frac{(1 - p_2(i))}{p_2(i)}}\right] \tag{9}$$

This expression turns out to be similar to (3) when the marginal Beta distributions degenerate to point distributions which is the case at individual level.

It is not hard to check that ρ is the correlation between the two magazines and rewriting (7) as :

$$\Pr\{x_1 = 1, x_2 = 1\} = p_1(i)\, p_2(i) + \rho\sqrt{p_1(i)(1 - p_1(i))}\sqrt{p_2(i)(1 - p_2(i))}$$

one can check that it always produce a legitimate value when ρ is positive and in most cases practical cases otherwise.

Similarly to what one gets in the context of Analysis of Correspondences a direct consequence of relations (7) and (8) is that ρ^2 is equal to the χ^2 distance which measure the amount of departure of the joint distribution from independence.

Finally the sample correlation coefficient will provide an estimate of the value of this unknown parameter.

4. Computing the Corr-Binomial

As we have already mentioned the most widely used media planning models are based on the Full-Binomial calculation. In order to enhance our current practice by the introduction of correlated audiences we need to develop a new method to actually carry the necessary calculations. Let us call such method, candidate successor of the Full-Binomial, the Corr-Binomial.

The difficulty is that we cannot compute (5) by convolution as in the Full-Binomial case since the random variables $X_1, X_2, \cdots, X_M$ are not independent anymore.

However a simple algebra trick will help us out.

Let us consider the following quantity:

$$Q_l = \Pr\left\{\frac{Z(Z-1)}{2} = l\right\} \quad \text{where} \quad Z = X_1 + X_2 + \ldots + X_M$$

Since the equation $k(k-1) = 2l$ has 0 or a unique solution for integers larger than 1 we have :

$$\Pr\{Z = X_1 + X_2 + \cdots + X_M = k \mid n_1, n_2, \cdots n_M\} = Q_{\frac{k(k-1)}{2}} \quad k = 2, 3, \ldots, M$$

If we are able to calculate the Q_l this will get us the values of the frequency distribution for 2 exposures or more.

After some algebra it can be shown that the quantity $\frac{Z(Z-1)}{2}$ reduces to a sum of independent random variables and can be computed using a convolution method.

More precisely :

$$\frac{Z(Z-1)}{2} = \sum_{m=1}^{M} \frac{n_m(n_m-1)}{2} x_m^2 + \sum_{1<m'<m''<M} X_{m'} X_{m''} \tag{10}$$

The leading terms are easy to get since it is a sum of independent Binomials x_m^2.
The trailing ones involve more work but they can be seen as the sum of independent two way Binomials.

To proceed we will use the following canonical type expansion similar to (3) which extends the Bernoulli case (9):

$$\Pr\{X_1 = k_1, X_2 = k_2 \mid n\} = P_{Bin}\{X_1 = k_1\} P_{Bin}\{X_2 = k_2\} \left[1 + \rho_B \frac{(k_1 - n_1 p_1)(k_2 - n_2 p_2)}{\sqrt{n_1 n_2 p_1 p_2 (1-p_1)(1-p_2)}} \right] \tag{11}$$

where we will set $\rho_B = corr(X_1, X_2) = corr(x_1, x_2) = \rho$.

The calculation of the above terms provides a probability estimate for all levels of exposures except 0 and 1. The 1 exposure will be deduced from the mean of the distribution which is a known quantity equal to $\mu = n_1 p_1 + n_2 p_2 + \cdots + n_M p_M$ and the 0 exposure from the sum which is equal to 1 by definition. Alternatively one can compute directly the 0 exposure.

So finally if we call $\{E_n^k(i) \mid k = 0,1,...,n\}$ the schedule frequency distribution for respondent i with sample weight π_i and $\{E_n^k \mid k = 0,1,...,n\}$ the total sample frequency distribution we will have :

$$\begin{cases} E_n^k = \sum_i \pi_i E_n^k(i) & k = 0,1,...,N \\ E_n^k(i) = Q_{\frac{k(k-1)}{2}}(i) & k = 2,...,N \\ E_n^1(i) = \left(\sum_{m=1}^{M} n_m p_m(i) \right) - \left(\sum_{k=2}^{M} k E_n^k(i) \right) & \\ E_n^0(i) = 1 - \sum_{k=1}^{M} E_n^k(i) & \end{cases} \tag{12}$$

So expressions (11) and (12) enable us to carry out the computations using the marginal and the joint distributions only.

We still need however to know, as for the Full-Binomial, the individual probabilities plus the specific ρ parameter.

This parameter will be calculated for each pair of magazines using the observed sample two ways crosstab:

		MAG 2	
MAG 1	f_1	f_{12}	$f_1 - f_{12}$
	$1 - f_1$	$f_2 - f_{12}$	$1 - f_1 - f_2 + f_{12}$
	1	f_2	$1 - f_2$

From which one can get easily the sample estimate of the correlation between two magazines :

$$\tilde{\rho} = \frac{f_{ij} - f_i f_j}{\sqrt{f_i (1 - f_i) f_j (1 - f_j)}} \tag{13}$$

In practice it may be advisable to estimate different values for this correlation coefficient separately for disjoint subpopulation such as sex and age segments.

5. Full-Binomial and Corr-Binomial comparison

We have applied the Corr-Binomial model in comparison with the Full-Binomial model on a real life dataset provided to us by courtesy of the AEPM.

We have been working with 10 magazines (5 weeklies and 5 monthlies) on a men target.

Here are the observed correlations:

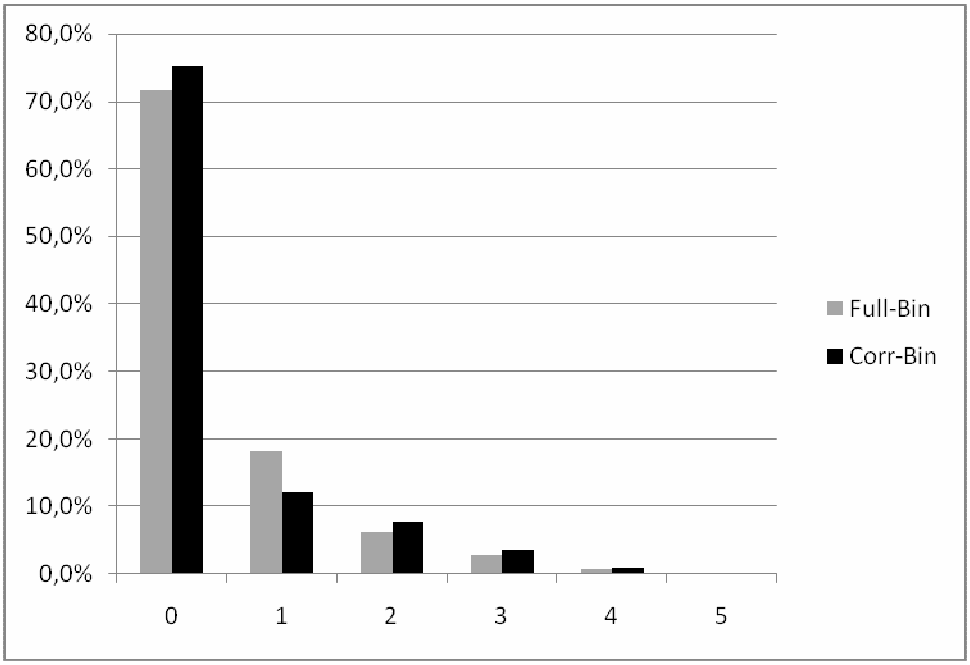
Rho	B	C	D	E	F	G	H	I	J
A	1	2	2	5	0	2	1	2	3
	B	35	1	5	12	2	1	23	27
		C	3	4	12	1	0	28	32
			D	0	0	7	2	2	2
				E	9	1	0	4	4
					F	0	0	16	15
						G	0	1	2
							H	2	3
								I	41

Some are rather high which makes sense since several magazines focus on cars.

In order to make the comparison easier we have considered no multiple issues and have built and evaluated a schedule with 1 insert in each magazine.

The resulting frequency distribution are the following ones :

	Full-Bin	**Corr-Bin**
Frequency		
0	71.7%	75.3%
1	18.2%	12.0%
2	6.2%	7.8%
3	2.9%	3.6%
4	0.8%	1.0%
5	0.2%	0.3%
Reach	28.3%	24.7%



Clearly the Full-Binomial over estimate the reach while the Corr-Binomial exhibits a higher proportion of reached 2 or 3 times.

Here is another example on women for a schedule including 4 uncorrelated magazines:

	Full-Bin	**Corr-Bin**
Frequency		
0	84.9%	84.9%
1	14.1%	14.0%
2	1.0%	1.0%
3	0.1%	0.1%
4	0.0%	0.0%
Reach	15.1%	15.1%

As expected the Full-Binomial and the Corr-Binomial match in that case.

6. Conclusion

The Corr-Binomial model introduces a mechanism to handle a long standing issue: the need for an adequate treatment in media planning of the correlations between magazines readership.

Although this model improvement implies a complexification of the calculations, the work can be done with the existing data sets.

The switch from the Full-Binomial model to the Corr-Binomial should not affect results in the general case but would produce a better reach and frequency estimate in the case of media schedules focusing on a center of interest. It may also have an impact on reach optimization since decorrelated audiences will then be used more efficiently (a small negatively correlated magazine can impact significantly the efficiency of a combination of other ones).

The use of the Corr-Binomial model can also be considered in other contexts such as exposure to magazine web sites.

A final word of caution though: the Corr-Binomial model is quite difficult to implement efficiently not only because of the underlying statistical complexity but also because of the numeric precision requested by the computations.

We strongly advise whoever will consider using this new model to take great care of those aspects and carry multiple control tests against the Full-Binomial model. We however believe that it is worth the effort.

References :

[1] Danaher P.J. & Hardie G.S. (2005), "Bacon With Your Eggs ? Application of a New Bivariate Beta-Binomial Distribution", *The American Statistician* November 2005, Vol. 59, N0. 4

[2] Santini G. (2003), "Mathematical Models & Methods for Media Research", *G.S. IT Services,* Chap. 11, 264-269