AUDIENCE ACCUMULATION MODELS: BINOMIAL, BETA-BINOMIAL, FULL BINOMIAL AND BEYOND

Gilles Santini, Informatique Medias Systemes

1) INTRODUCTION :

One of the oldest question in media research is :

"How does audience built up over time ?".

For the past two decades or so the most widely accepted mathematical model used to describe the accumulation of the audience over several issues of a magazine has been the Betabinomial Model.
Notwithstanding the practical value of such model one may gain significative benefits from re-examination of the matter.

The classical research cycle is :

- ❑ modelisation
- ❑ validation
- ❑ rejection of the premises.

Any theory even widely accepted bears in itself its own obsolescence. Only the best ones are rich enough to transmit their heritage.

The belief here is that the original Binomial accumulation model which gave rise in the past to the Betabinomial Model is still the proper seed for new classes of model of wider applicability.

Before any confrontation against experimental data new models should be built theoritically. Such is the view angle of the present paper. It should not be interpreted as a dismiss of empirical validation nor of lack of interest for practical implementation. It is merely the relation of the first phase of a new research cycle.

2) BASIC FRAMEWORK :

From now on the following framework will be considered :

- ❑ A representative sample of the total population containing I individuals named i and weighted by a sample weight π_i. Without lack of generality it is assumed that the sum of the sample weights equal 1 ($\sum_{i=1}^{I} \pi_i = 1$).
- ❑ Each individual i has a known probability p_i of reading an issue of a given magazine M.
- ❑ $e_n^k(i)$ is the probability for individual i to read k issues of magazine M among n.

- ❑ E_n^k is the average probability over the entire population to read k issues of magazine M among n.

All individuals being independent one can write that :

$$E_n^k = \sum_{i=1}^{I} \pi_i e_n^k(i)$$

This *"aggregation principle"* states that the average probability of reading k issues of magazine M among n is the sample average of the individual probabilities of reading k issues amoung n.

3) THE INDIVIDUAL BINOMIAL MODEL (IBM) :

Under the hypothesis that *at individual level* reading of various issues of a given magazine are *independent* events it is easy to derive the individual frequency distribution of reading :

$$e_n^k(i) = \binom{n}{k} p_i^k (1 - p_i)^{n-k}$$

This is a straightforward Binomial distribution.

Averaging over the population sample leads to the average frequency distribution of reading E_n^k. Such model is often called Full Binomial.

The Binomial distribution benefits from nice properties that allow fast computation of E_n^k. Namely, if R_n denotes the average coverage of n issues for the total population :

$$R_n = 1 - \sum_{i=1}^{I} \pi_i (1 - p_i)^n$$

Then if one compute the sequence of differences :

$$\Delta_n^1 = R_n - R_{n-1}$$

$$\Delta_n^2 = \Delta_n^1 - \Delta_{n-1}^1$$

...etc...

$$\Delta_n^k = \Delta_n^{k-1} - \Delta_{n-1}^{k-1}$$

the distribution of reading E_n^k follows readily since it can be proved that :

$$E_n^1 = \Delta_n^1$$

$$E_n^2 = -\binom{n}{2}\Delta_n^2$$

$$E_n^3 = +\binom{n}{3}\Delta_n^3$$

...etc.....

$$E_n^k = (-1)^{k-1}\binom{n}{k}\Delta_n^k$$

The previous calculation mecanism involves a limited number of multiplications. This is a good thing since multiplications are much slower operations to perform with an electronic processor than additions and multiplications.
More over since the probability p_i takes in practice a limited number of distinct values it is a good idea to store as precalculated quantities the terms $(1-p_i)^n$ which are involved in the calculation of R_n.
Assuming that the Binomial coefficient $\genfrac{}{}{0pt}{}{n}{k}$ are stored too, the whole process of calculation of the frequency distribution of exposures for the individual binomial model turns out to be extremely fast.

4) THE AGGREGATED BETA BINOMIAL MODEL (ABBM) :

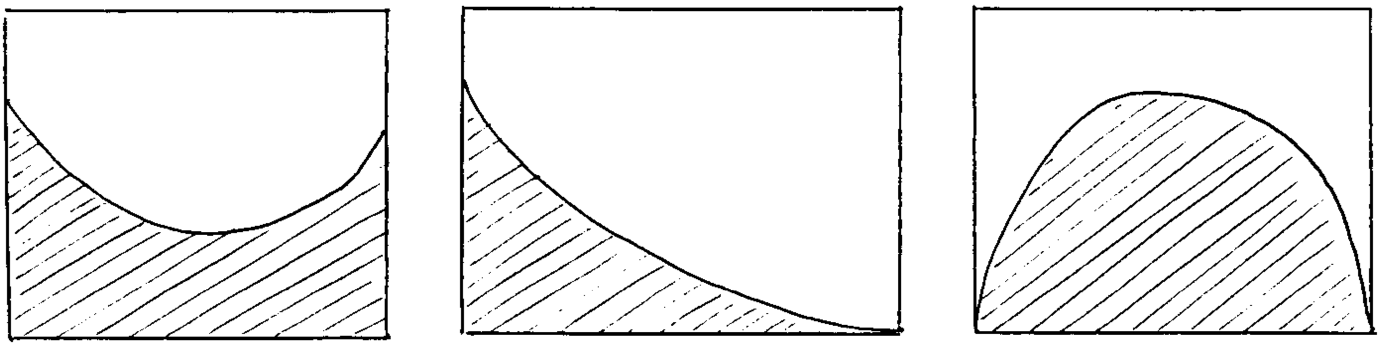
The individual Binomial model leads under further hypothesis to the well known agregated Betabinomial Model.

The additional hypothesis made is that the p_i probabilities are distributed among the population according to a Beta distribution. Although, such an hypothesis is made within the models inside widely used mediaplanning systems, one should be aware that it is a very questionnable one.

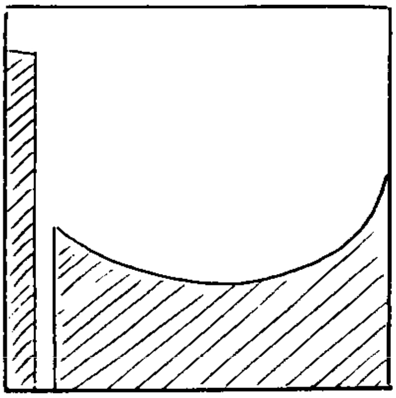
In fact the hypothesis readily fails to be true as soon as the distibution of the p_i is not a smooth distribution between 0 and 1 with no accumulation around 0 (never read) or 1 (always read).

When mediaplanning was concerned with the performence of general interest magazines over large unspecific target groups such an hypothesis might have been good enough. This is not the case any more with more focused vehicules and precise media-marketing strategies.

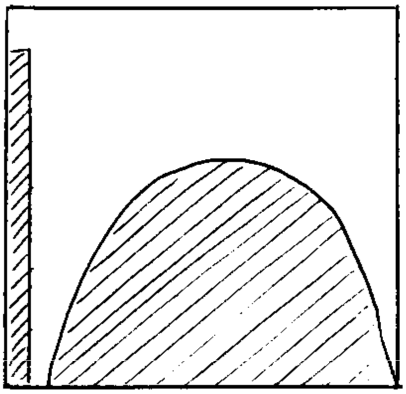
Typically if the distribution of probabilities in the target group looks like :



or similar shapes, the aggregated Betabinomial model can be considered safely.
On the contrary, if the distribution of probabilities in the target group is :



or



or similar shapes the Beta hypothesis is not valid. In such case the Betabinomial Model is misleading and the original individual Binomial Model should then be used.

5) THE INDEPENDENCE HYPOTHESIS :

The individual Binomial model and the aggregated Betabinomial models postulate independence at individual level of the reading of several issues of a magazine. Such *independence is not true at population level.* However the form of the individual Binomial model implies a rather strong property at population level : *the probability of reading k issues among n is independent of which specific k issues are considered.* This property is known in statistical theory as exchengeability of the events. It can be proved (De Finetti's theorem) that if such property is verified then the model comes necessarily from the individual Binomial class. This may sound good and give a good foundation to the accumulation models above. However, considering the fact that independence of reading successive issues of magazine is a basic assumption of these models, serious criticisms may be raised. Indeed, droping such independence assumption stimulates new findings.

6) THE INDIVIDUAL BETA BINOMIAL MODEL (IBBM) :

On way to relax the independence hypothesis at individual level is to reduce the speed with which the accumulation builts up over issues. For such purpose it is handy to use the casuality ratio :

$$\rho = \frac{R_2 - R_1}{R_1 (1 - R_1)}$$

that measures the growth of the coverage from one to two issues compared to the growth under the independence assumption.

Clearly under the independence assumption ρ equals 1 but if the growth is lower ρ is less than 1.

The Betabinomial law with parameters α et β has a constant casuality ratio :

$$\rho = \frac{\alpha + \beta}{\alpha + \beta + 1}$$

One may then think using the Betabinomial law to describe at individual level an accumulation[1] process that accumulate less quickly than under the independence assumption. The coefficient ρ being fixed from some other source, it is possible to write for each individual that :

$$\rho_i = \frac{\alpha_i}{\alpha_i + \beta_i}$$

which leads to :

$$\alpha_i = p_i \frac{\rho_i}{1 - \rho_i}$$

$$\beta_i = (1 - p_i) \frac{\rho_i}{1 - \rho i}$$

Then the frequency distribution of exposure of the individual i can be calculated by recursion from the following formules :

$$C^{k}_{n} (i) = A_k \frac{n - k + 1}{k} C^{k-1}_{n} (i)$$

[1] This Full Betabinomial model should not be confused with the usual Betabinomial one described in § 4.

where

$$A_k = \frac{\alpha_i + k - 1}{\beta_i + n - k}$$

The never read frequency being :

$$\overset{0}{C_n}(i) = 1 - \frac{\beta_i \ (\beta_i + 1) \ldots\ldots\ldots (\beta_i + n - 1)}{(\alpha_i + \beta_i) \ (\alpha_i + \beta_i + 1) \ldots\ldots\ldots\ldots (\alpha_i + \beta_i + n - 1)}$$

Although, such calculation by recursion is rather fast one still has to perform it for each individual in the sample population. The aggregation over all individual being done sequencially. The individual Betabinomial Model is identical to the individual Binomial model if $\rho = 1$. It clearly relaxes the independence assumption of the latter and bring improvment to it. However, this benefit is obtained through a much heavier computation task.

Last but not least, the individual Betabinomial model fails to differenciate the various issues. This is a major limitation which calls for a different, more radical, way of challenging the individual Binomial model.

7) THE INDIVIDUAL MARKOVIAN BINOMIAL MODEL (IMBM) :

In order to develop this new Model, it is useful to introduce additional notations.

From now on X with values 1 (YES) or 0 (NO) will denote the fact that a given individual reads or not read the k^th^ = issue of a sequence of n issues of a magazine. The vector

$$P_k = \begin{pmatrix} \text{Prob} & \{X_k = 1\} \\ \text{Prob} & \{X_k = 0\} \end{pmatrix}$$

gives the probability of reading and not reading the k^{th} issue. The basic idea is to write that what happens for the k^{th} issue only depends on what has happened for the previous one. There is dependency from one issue to the next one but such influence is not carried over onto further issues.

Using this property known has the Markovian property and assuming that the influence of reading the kth issue on the fact of reading the (k + 1)th issue is quantitatively the same as the one of reading any other issue on its next one, leads to a good looking result :

$$P_{k+1} = \begin{pmatrix} 1 - \rho q & \rho p \\ \rho q & 1 - \rho p \end{pmatrix} P_k$$

where ρ is the casuality ratio and p (resp.q) the probability of reading (resp. not reading) an issue.
The correlation between two consecutive issues can be calculated and is :

$$\mathrm{Corr}\,(X_k\,,\,X_{k+1}) = 1 - \rho$$

If $\rho = 1$ the correlation is null and the accumulation is Binomial
If $\rho = 0$ the correlation is total and there is no accumulation
If $\rho = 2$ ($p = q = 0{,}5$) the correlation is total with a fixed read/not read behaviour.

The coverage turns out to be :

$$R_n = 1 - q\,(1 - \rho p)^{n - 1}$$

which can be rewritten

$$R_n = 1 - (1 - R_1) \left(\frac{1 - R_2}{1 - R_1} \right)^{n - 1}$$

under such form it can be compared to the coverage of the Betabinomial which accept the following approximation :

$$R_n \simeq 1 - (1 - R_1) \left(\frac{1 - R_2}{1 - R_1} \right)^{\log_2 (n)}$$

This stresses the fact that accumulation is in general faster with the Markovian Binomial Model than with the Betabinomial. This is not a satisfactory fact at aggregated level where Betabinomial accumulates already two quickly in some cases (in particular if there are individuals with zero probabilities). However, if the Markovian Binomial model is applied at *individual level*, it provide a good intermediate between the IBM model and the IBBM model when the latter accumulates two slowly which is often the case.

In the past the Markovian Binomial Model has only been considered at the aggregated level and dismissed because of poor performence and intractability of the computation of the distribution of frequencies of reading.

Using it at individual level dramatically change the scope. It provides a sound alternative to the independence assumption and offer a significant flexibility to model the accumulation process.

This can only be considered because of new algorithmic advances that allow to compute quickly the distribution of the frequency of exposition for the total population. Appendix I gives some indications to how fast calculation can be performed.

Appendix 1: Calculation of the Distribution of the Frequency of Reading Under the IMBM Model

□ For each individu i with probability of reading p_i ($q_i = 1 - p_i$) compute by recursion the joint quantities (a_n^k , b_n^k) :

$$a_n^{k+1}(i) = (1 - \rho\, q_i)\, a_{n-1}^k + \rho p_i\, b_{n-1}^k$$

$$b_n^k(i) = \rho q_i\, a_{n-1}^k + (1 - \rho p_i)\, b_{n-1}^k$$

□ The distribution of the frequency of reading of individual i is :

$$C_n^k(i) = a_n^k(i) + b_n^k(i)$$

□ For the total population, the distribution of the frequency of reading is :

$$E_n^k = \Sigma_i\ \pi_i\ C_n^k(i)$$