Gilles SANTINI
Readership Symposium

AUDIENCE ACCUMULATION MODELS :

A framework for a theoretical approach

0. INTRODUCTION

Since the first international Media Research Symposium in New Orleans research has mainly concentrated on audience measurement and survey techniques (questionnaires design, telephone interviews, fusions ... etc).

Indeed very limited efforts have been devoted to mediaplanning models. Still most of the everyday usage of audience data involve audience accumulation models.

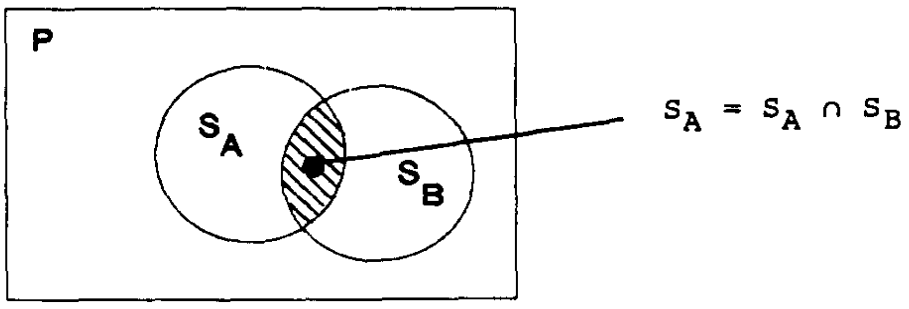
The intent of this paper is neither to present a new model nor to evaluate the relative goodness of fit of the commonly available ones. It attempts to put in a coherent framework the various types of models and to outline their interrelations, inheritage and underlying hypothesis.

1. SET THEORY BASED MODEL

The basic assumptions of the set theory based model are simple :

1. the base population is a set made of a countable number of people.
2. Audience can be described through a static notion allowing to decide by yes or no for each individual in the base population whether he reads or does not read a specific magazine.

Given these two assumptions it is most elementary to express the matter in a set theoretic manner :



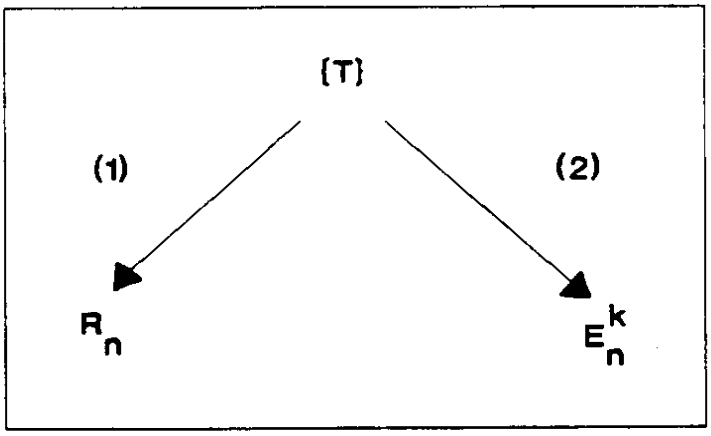
P represents the base population
The size of the set s_A respective to P is equal to the audience o magazine A : all readers of A are within the set s_A and all points within s_A are readers of A.

The duplication between two magazines A and B is represented by the intersection of the two associated sets s_A and s_B.

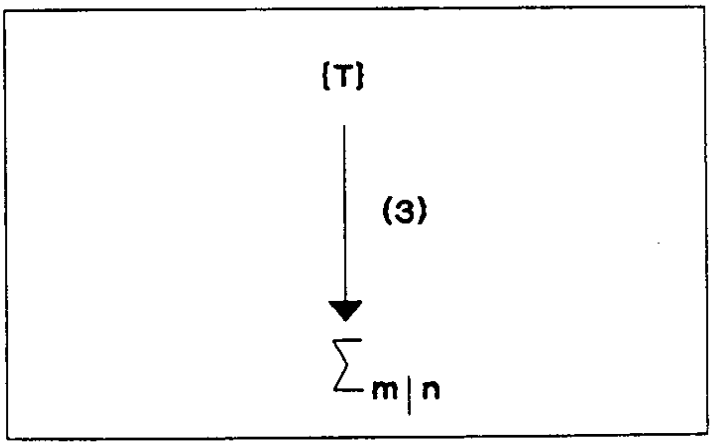
In order to derive some fondamental results we need to introduce the following quantities :

T_0 = 1
T_1 = sum of all audiences
T_2 = sum of all duplications
T_3 = sum of all triplications
.
.
.
T_n = sum of all intersections of n distincts vehicules.

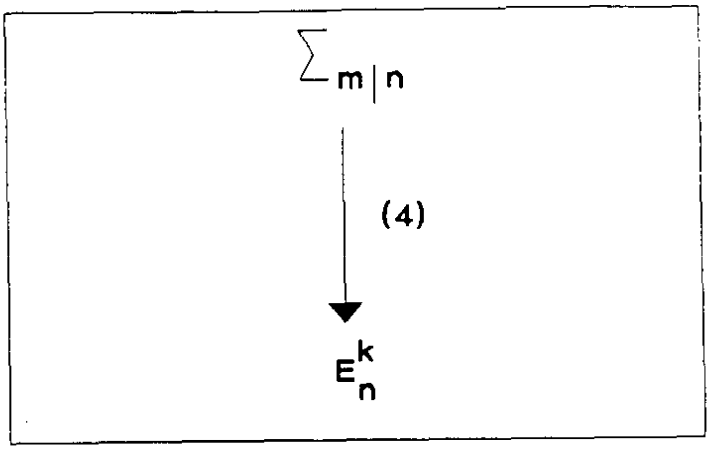
The first result comes from straightforward set theory algebra and states that knowing the above T quantities it is possible to compute exactly the audience of a combination of n magazines and the number of readers of exactly k magazines among the n. If we denote R_n and E_n^k respectively these two notions we have



Using a combinatonial argument one also can prove that the sum of the reach of each combination of m magazines amoung n ($\Sigma m/n$) is computable from the knowledge of the same T quantities



Inverting this last relationship we have established a second result that tells us how to compute E_n knowing the $\Sigma m/n$:



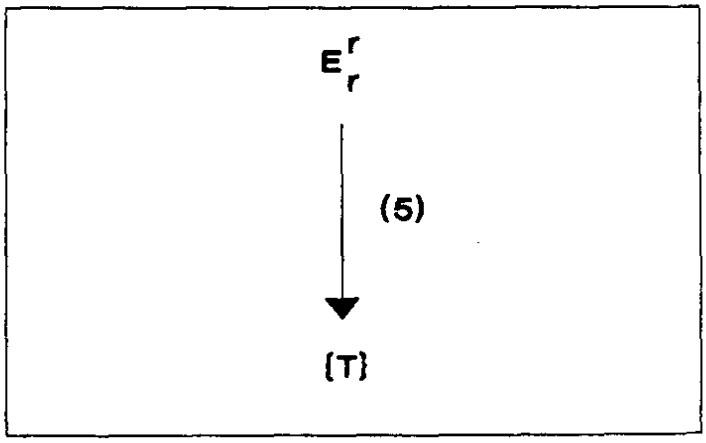
One should note that this result has been derived without any form of approximation. It provides a way to exactly compute the frequency of reading whenever one has the benefit of a formula that gives the reach of a combination of m magazines.

2. EXCHANGEABILITY MODELS

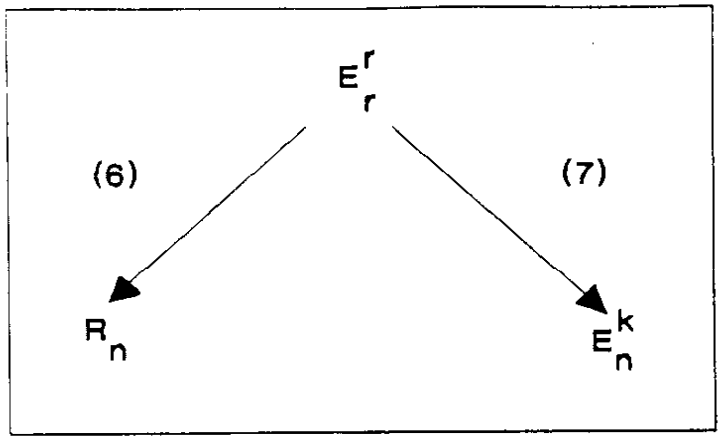
For a moment we will only consider the combination of n issues of a given magazine. We say that these n issues are exchangeable if and only if the reach of m issues choosen among the n ones is independant of which specific m issues are considered. This hypothesis clearly implies that all n issues have the same audience.

The exchangeable hypothesis has received a rather wide interest among statisticians for it is a weaker hypothesis that the independance one. In the case of the accumulation of the audience of n issues of a magazine it allows to push forward the theory.

Under the exchangeable hypothesis it is very easy to compute the T quantities as a fonction of the regular readers E_r (ie those who have read r issues out of r)



Hence one can derive a set of two relations that let compute reach and frequency of reading from the number of regular readers.



The concept of exchangeability not only allows us to derive the above interesting relationship but leads in addition to the following result :

If a magazine behaves accordingly to the exchangeability assumption it accumulates as the average of a binomial over the base population(1).

In mathematical terms this can be writen with on integral sign

$$E_n^k = \int_0^1 \underbrace{\binom{n}{k} \rho^k (1-\rho)^{n-k}}_{A} \underbrace{f(\rho)}_{B} d\rho$$

Part A is the binomial law
Part B is a function that accounts for the form of the distribution of the binomial parameter f(p) over the population.

If f(p) is choosen to be a Beta density functions, which is wise for a lot of practical reasons and for ease of adjustement, one ends up for E_n^k with a BetaBinomial distribution.

Hence the BetaBinomial accumulation law can be considered as an outcome of the exchangeability assumption.

Exchangeable accumulation has a lot of other nice properties such as usefull recurrence relationship between the reach and the various levels of frequency (8,9). To state only one result interesting enough to remember one should note the fact that magazines under the exchangeability assumption always accumulate with a slower rate than under the independance hypothesis.

$$R_{n-1} \leq R_n \leq 1-(1-R_1)^n$$

(1) This result comes from a theorem by DE FINETTI (1930)

3. FUNCTIONAL MODELS

Functional Models have been introduced in media research on a Ad-Hoc basis. Most of them are related to Set theory and Exchangeable Models by way of approximation. For exemple the well know and often used Agostini-Hoffmans formula (10) that estimates the reach of a combination of magazines knowing their audiences and their pairwise duplications can be considered as an approximation of formula (1) that link the T quantities to R_n. Likewise Politz has used a functional Model (11) where the accumulation of a magazine grows as the logarithm of the number of issues. Such model can be proved to be a crude approximation of the BetaBinomial one. Except for the work by Armand MORGENSTERN that follows a logic of its own, it is possible to draw similar conclusions for most of the Ad-Hoc models that the media researcher may comes across. This is to say that audience behaviour do not obey functional laws but can better be modeled by models based on set or probability theory.

The practical value of simple functional formulas comes from the goodness of fit that they allow but such quality should be considered as the result of their proximity to more theoretical models.

4. CASUALITY MODELS

This paper has sketched so far how exchangeable and functional models are related to the set theory framework. Such framework allow to develop another class of models less known of the media research community. Such approach originates from work done by Christopher FRY but the following construction is new.

Let us first generalize an old notion

$$\rho = \frac{R_2 - R_1}{R_1 - (R_1)^2}$$

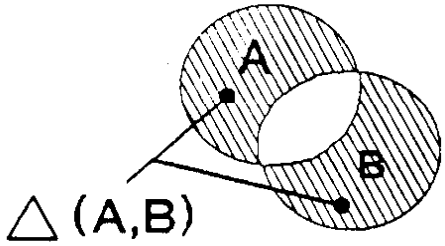
known as the casuality coefficient that measure the ratio of the growth shown by the reach when one go from one issue to two issues respective to what would have been such growth under the hypothesis of independance of the two issues.

For two magazine A and B one can define

$$\delta_{AB} = \frac{(A+B-D) - \frac{A+B}{2}}{(A+B-AB) - \frac{A+B}{2}}$$

δ_{AB} measures the ratio of the growth of the reach of the combination of the two magazines over the average of their own audiences, respective to what would have been such growth under the hypothesis of independance of the two magazine. Clearly δ_{AB} reduces to $\mathcal{D}$ when one consider two issues of the same magazine instead of two distinct magazines.

δ_{AB} can be rewritten using the set theory concept of symmetric difference



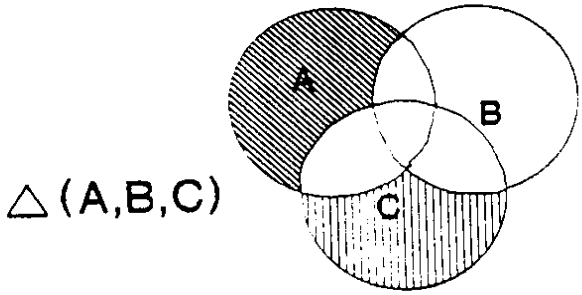
This can be seen by simplily rewritting

$$\delta_{AB} = \frac{A+B-2D}{A+B-2AB} = \frac{\triangle(A,B)}{\dot{\triangle}(A,B)}$$

The little dot over the $\triangle$ symbol means that the independance assumption is supposed to be true.

A consequence of this rewritting is that one can generalize the concept to more than two magazines.

The symmetric difference of several sets is defined as follows :



In media research terms it is the same notion as the exclusive readers of the combination of several magazines. Hence the casuality coefficient can be generalised as follows

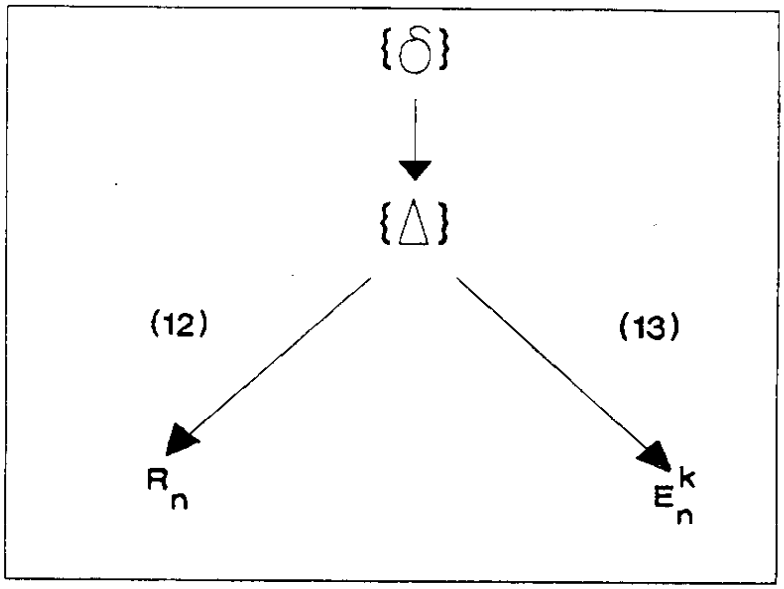
$$\delta_{A,B,C,} = \frac{\triangle(A,B,C)}{\dot{\triangle}(A,B,C)}$$

If one can make an Ad-Hoc assumption on the value of $\delta_{ABC,}$ since $\triangle(A,B,C)$ is easily computable, one can estimate the $\dot{\triangle}(A,B,C)$ quantities.

The missing link to understand kow all this lead to a new class of models comes from the following fundamental result.

The reach and frequency distribution of a combination of n magazines can exactly be computed from quantities that only involve symmetric differences of all possible combinations of them.

The proof of such theorem is not straightforward but the result is very neat and provides us with a powerful exact relationship



5. INDIVIDUAL MODELS

The previous quick overview of how audience accumulation models are derived from basic set theory find further developpments by introducing individual behaviour over time. Such models which draw heavily form stochastic models theory are beyond the scope of this paper. However it is interesting to know that they lead to non contradictiory results specially with the exchangeable models. Over all one can be striken by the global harmony of all theses constructions as if reality was operating under the ruling of mathematics. This is true in other fields too.

As a motivation for future work it is interesting to read what Albert EINSTEIN was writting in 1933 :

«I am convinced that it is possible to discover by mean of pure mathematical constructions the concepts and the laws linking them that provide the key to understand the natural phenomenons».

This do apply too to sociocultural phenomenons such as audience behaviour.

MATHEMATICAL ANNEX

1- NOTATIONS

S_i = Readers of magazine S_i

S_{ij} = Readers of magazine S_i and magazine S_j

$S_{ili2...ik}$ = Readers of magazines S_{i1} and S_{i2} ... and S_{ik}

R_n = Reach of the combination of S_i... S_n

E_n^k = Frequency of exactly k magazines read among n

$$T_0 = 1$$

$$T_2 = \Sigma S_i$$

$$T_3 = \sum_{i<j} S_{ij}$$

.
.
.

$$T_r = \sum_{i1<i2<...<ir} S_{ili2...ir}$$

$$Z_1 = \Sigma S_i$$

$$Z_2 = \sum_{i<j} \Delta(S_i, S_j)$$

.
.
.

$$Z_r = \sum_{i1<i2<...<ir} \Delta(S_{i1}, S_{i2}, \ldots S_{ir})$$

2- FORMULAS

2.1. SET THEORY BASED MODELS

$$(1)\quad R_n = T_1 - T_2 + T_3 - \ldots + (-1)^{n-1} T_n$$

$$(2)\quad E_n^k = T_k - \binom{k+1}{k} T_{k+1} + \binom{k+2}{k} T_{k+2} - \ldots + (-1)^{n-k} \binom{n}{k} T_n$$

$$(3)\quad \Sigma_{m/n} = \binom{n-1}{m-1} T_1 - \binom{n-2}{m-2} T_2 + \ldots + (-1)^{m-1} T_m$$

$$(4)\quad E_n^k = \sum_{m=n-k}^{n} (-1)^{n-k+m+1} \binom{m}{n-k} \Sigma_{m/n}$$

2.2 EXCHANGEABILITY MODELS

$$(5)\quad T_r = \binom{n}{r} E_r^r$$

$$(6)\quad R_m = \sum_{r=1}^{m} (-1)^{r+1} \binom{m}{r} E_r^r$$

$$(7)\quad E_n^k = \binom{n}{k} \sum_{r=k}^{n} (-1)^{r-k} \binom{n-k}{n-r} E_r^r$$

$$(8)\quad R_m = \sum_{k=1}^{n} \left[1 - \frac{\binom{n-m}{k}}{\binom{n}{k}} \right] E_n^k$$

$$(9)\quad E_{n-t}^k = \sum_{i=0}^{t} \frac{\binom{t}{i}\binom{n-t}{k}}{\binom{n}{k+i}} E_n^{k+i}$$

2.3 FUNCTIONAL

$$(10)\quad R_n = \frac{\Sigma S_i}{1+\dfrac{\Sigma k_{ij}S_{ij}}{\Sigma S_i}} \qquad \text{with} \qquad k_{ij} = \frac{S_i+S_j}{S_iS_j-S_{ij}}$$

$$(11)\quad R_n = R_1+(R_2-R_1)\log_2(n)$$

2.4 CASUALITY MODELS

$$(12)\quad nR_n = \sum_{k=1}^{n} \begin{bmatrix} n-1 \\ k-r \end{bmatrix}^{-1} Z_k$$

$$(13)\quad E_n^1 = Z_n$$

$$\vdots$$

$$E_n^k = \frac{1}{k}\left[Z_{n-k+1} - \sum_{r=1}^{k-1} r \begin{bmatrix} n-r \\ k-r \end{bmatrix} E_n^r \right]$$

$$\vdots$$

$$E_n^n = \frac{1}{n}\left[Z_1 - \sum_{r=1}^{n-1} r\, E_n^r \right]$$