REF:140/354

An Overview of some of the Theory behind the On-line Casualness System

prepared by

George Rennie

O.R. Systems Pty Ltd

50 Nicholson St

FITZROY 3065

(PH 417-5822)

for

The Roy Morgan Research Centre

499 BOURKE ST,

MELBOURNE 3000

(PH 602-5222)

October 14, 1986

CONTENTS

1. INTRODUCTION

This document describes most of the theory underlying the new casualness system. The document generally covers the same ground as the presentation given on Wednesday 24th September 1986, although some additional comments and suggestions have been incorporated. A definition of casualness is outlined followed by a description of its variation in a population. The next section is on the estimation of casualness from the reinterview sample and contains an account of the formulae and the smoothing techniques used by the system in order to overcome some of the inherent problems.

2. WHAT IS CASUALNESS?

The casualness (γ) of a publication is defined as the ratio of the additional reach* of a second issue (over the first) to the additional reach which would be expected if the persons reading the first and second issues were chosen independently of one another.

For example, given a 50% readership in a population of 100 people, if the 50 readers of the first issue were chosen independently of the 50 readers of the second issue we would expect a 75% reach (i.e. 75 readers) after two issues. This represents an additional reach of 25%. If in fact we found that the reach was only 60% (i.e. 40% read neither issue, 20% read one issue and 40% read both issues), the additional reach would be 10% and the casusalness would be:

$$\gamma = \frac{10\%}{25\%} = 40\%$$

---

* Throughout these notes the formula for additional reach (AR)

$$AR = \gamma rs$$

is frequently assumed. In this formula, r denotes the readership, s=1-r the proportion of non readers of the average issue and γ is the casualness.

3. THE VARIATION OF THE CASUALNESS OF A PUBLICATION ACROSS THE READERSHIP OF THE PUBLICATION.

(3.1) Introduction

By far the most important point to recognise is that if two subgroups have the same casualness (γ say) it does not follow that the combined group will also have the same casualness (γ). Indeed the casualness of the combined group will nearly always be less than γ. It can never be greater and it can only ever equal γ if the readerships of the two subgroups are the same.

The best illustration of this is probably by means of an example. Let us suppose that a population comprises two equal groups A and B. Further let us suppose that A has a 98% readership and a 100% casualness and that B has a 2% readership and a 100% casualness. The two issue reach of the total population will be

$$50\% \ (0.98 + 0.02 \times 0.98) + 50\% \ (0.02 + 0.98 \times 0.02)$$

$$= 52\%$$

The casualness of the total population will be

$$\frac{2\%}{25\%} = 8\%$$

Thus we have two subgroups each with 100% casualness which when combined yield a total population with only an 8% casualness.

This also works in reverse. As we subdivide a population we should expect average casualness not to stay steady but to rise. Casualness is a measure of homogeneity. A perfectly homogeneous population must by definition have a casualness of 100%. It follows therefore that the more tightly defined the demographics, the more homogeneous the population and therefore the higher the casualness.

Thus as a general rule, if we define small homogeneous subgroups of a population we should expect to get high casualness figures.

If we now consider the casualness of publication as a function of its readership we must take into account two components. These are the distributional component and the commitment component.

(3.2) The Distributional Component of Casualness

One possible model is to consider a population as comprising varying proportions of two subgroups

(A) A group of people who never read,

(B) A group of people who have a readership r^* and a casualness γ^*.

Under this model, if the proportions of (A) and (B) are α and $1-\alpha$, then the overall readership

$$r = \alpha r^*$$

and the overall casualness

$$\gamma = \frac{\alpha\gamma^* r^* s^*}{rs} \quad \text{where } s = 1-r$$

$$= \frac{\alpha\gamma^* r^* s^*}{\alpha r^* s}$$

$$= \frac{\gamma^* s^*}{s}$$

$$= \frac{\gamma^T s^T}{s}$$

where the superscript T denotes the corresponding value for the total population.

Thus under this model the casualness of the group selected is inversely proportional to s the fraction of non-readers of the average issue. The model is equivalent to the assumption of constant turnover.

The component of casualness

$$\gamma_D = \frac{\gamma^T s^T}{s}$$

which corresponds to this model is known as the distributional component. The reason for calling it the distributional component is that it gives a realistic representation of the effect of mixing areas where a publication is distributed with areas where it is not distributed (eg city v country for some publications). The distributional component rises as the readership rises.

(3.3) The Commitment Component of Casualness

The contrasting model is to consider the population as comprising varying proportions of

(B) A group of people who have a readership r^* and a casualness γ^*,

(C) A group of people who always read.

Under this model, if the proportions of (B) and (C) are α and $1-\alpha$, then the overall readership

$$r = 1-\alpha + \alpha r^*$$

and the overall casualness

$$\gamma = \frac{\alpha\gamma^* r^* s^*}{rs}$$

$$= \frac{\alpha\gamma^* r^* s^*}{r(1-(1-\alpha+\alpha r^*))}$$

$$= \frac{\alpha\gamma^* r^* s^*}{r\alpha\,(1-r^*)}$$

$$= \frac{\gamma^* r^*}{r}$$

$$= \frac{\gamma^T r^T}{r}$$

The corresponding component of casualness

$$\gamma_c = \frac{\gamma^T r^T}{r}$$

is known as the commitment component. It simulates the effect of adding or subtracting committed readers. In direct contrast to the distributional component, the commitment component falls as the readership rises.

(3.4) The General Case

In the more general case, casualness will vary according to the direction of travel in demographic space. In some directions it will resemble the distributional component and in other directions it will resemble the commitment component. Thus while both the distributional model and the commitment model conserve two issue reach+, as does any constantly weighted combination of them, none of these models can explain all the variations in casualness.

One model which can reproduce all the variations in casualness, however, is that which results when the population is considered as entirely comprising the three subgroups (A), (B) and (C), vis

(A) The group who never read,
(B) The people who have a readership r^* and a casualness γ^*,
(C) The people who always read.

Under this model it can be shown that the casualness of a group i which comprises fractions α_1, α_2 and α_3 of groups (A), (B) and (C) respectively and which constitutes a fraction f_i of the total population is

$$\gamma_i = \frac{\alpha_2}{f_i} \frac{r^T s^T}{r_i s_i} \gamma^T \qquad (1)$$

+ The proofs of these statements are as follows:
Given sets i with frequences f_i such that $\Sigma f_i = 1$ and $\Sigma f_i r_i = r^T$, then for the distributional model the additional reach

$$= \Sigma_i f_i \gamma_i r_i s_i$$

$$= \Sigma f_i \frac{\gamma^T s^T}{s_i} r_i s_i$$

$$= \gamma^T s^T \Sigma f_i r_i$$

$$= \gamma^T s^T r^T.$$

For the commitment model, the additional reach

$$= \Sigma_i f_i \gamma_i r_i s_i$$

$$= \Sigma_i f_i \frac{\gamma^T r^T}{r_i} r_i s_i$$

$$= \gamma^T r^T \Sigma f_i s_i$$

$$= \gamma^T r^T s^T.$$

To show that this formula for casualness can simulate the distributional component, we consider the ratio $\alpha_2 : \alpha_3$ constant. Then

$$\frac{\alpha_2}{f_i} = \frac{r_i}{r^T}$$

and so

$$\gamma_i = \frac{r_i}{r^T} \frac{r^T s^T}{r_i s_i} \gamma^T$$

$$= \frac{s^T}{s_i} \gamma^T$$

To show that the formula can simulate the commitment component, we consider the ratio $\alpha_1 : \alpha_2$ constant. Then

$$\frac{\alpha_2}{f_i} = \frac{s_i}{s^T}$$

and so

$$\gamma_i = \frac{s_i}{s^T} \frac{r^T s^T}{r_i s_i} \gamma^T$$

$$= \frac{r^T}{r_i} \gamma^T$$

Finally to show that the formula conserves two issue reach we evaluate the sum

$$\Sigma f_i r_i s_i = \Sigma \alpha_2 r^T s^T \gamma^T$$

$$= r^T s^T \gamma^T$$

since $\Sigma \alpha_2 = 1$ when the groups i add to the total population.

4. ESTIMATING CASUALNESS FROM A REINTERVIEW SAMPLE

(4.1) Raw Casualness

If the estimated single issue reach from the re-interview sample is R1 and the estimated two issue reach is R2 then the raw causalness is

$$\gamma_R = \frac{R2 - R1}{R1\ (1 - R1)}$$

(4.2) Drawbacks of raw casualness figures

While the raw casualness figure is probably the most suitable statistic to use for the estimation of the overall casualness of a single publication, it is certainly not the best to use in the estimation of casualness for demographic breakdowns of that population. There are two main reasons for this:-

(a) Raw casualness figures can be unduly influenced by random fluctuations arising out of small sample sizes, and

(b) Raw casualness figures will not give consistent answers due to differences in population and readership frequences between the re-interview survey and the population to which it is being matched (in our case the readership survey).

Examples of (a) are easy to visualize. For example, the sample might yield N0=10, N1=5, N2=0; this will give a raw casualness figure of 120%. Alternatively the sample might yield N0=15, N1=0, N2=2; this will give a raw casualness figure of 0%. This problem could even be exacerbated in the case of first time read publications where negative weights have to be used in the estimation of two issue reach*, although the program does overcome this problem in the estimation of α_2 (See section 4.3 below).

---

* This occurs for example where respondents read 2 say on the first round and 3 say on the second.

To see how situation (b) can arise, imagine a sample which gives a 60% raw casualness for males and a 40% raw casualness for females where each sex accounts for 50% of the total sample. In this case the raw casualness figure for the total sample will be 50%. To apply these raw casualness figures (males 60%, females 40%, total 50%) to the actual population (i.e. readership sample) would be fine so long as the numbers and the readers of each sex are equal. If they are not then discrepancies will occur. To see this let us suppose that male readership in the target population is in fact 30% and female readership is 70%. In this case assuming 60% casualness for males and 40% for females our total two issue reach will be

$$50\% \ (0.3 + 0.6 \times 0.7 \times 0.3) = 21.3\% \text{ for males}$$

$$+ \ 50\% \ (0.7 + 0.4 \times 0.3 \times 0.7) = 39.2\% \text{ for females}$$

$$= 60.5\% \text{ total.}$$

This is equivalent to a total casualness of 42% rather than 50%. In order to achieve an overall casualness of 50% in the target population (in this case) we would need higher casualness figures than 60% for men and 40% for women (e.g. 71.4% for men and 47.6% for women).

Similar examples can also arise when the relative proportions of demographic groups differ between the sample and the target population. However we now believe that this problem can (and should) be overcome by alternative means namely multidimensional weighting of the sample to the target population. By these means it should even be possible to automatically adjust overall casualness figures for changing socio-economic patterns (e.g. more/less unemployment, ageing population, etc).

(4.3) Overcoming these problems

The method the system uses to overcome problem (b) of Section (4.2) is that instead of using raw casualness figures it makes direct use of equation (1), vis

$$\gamma_i = \frac{\alpha_2}{f_i} \frac{r^T s^T}{r_i s_i} \gamma^T$$

with γ_i = the casualness of group i
f_i = its proportion in the total population

r^T = the readership of the total population
r_i = the readership of group i

$s^T = 1 - r^T$,
$s_i = 1 - r_i$, and

γ^T = the total casualness.

In applying this formula, the system first determines the overall (raw) casualness γ^T from the re-interview survey data. It then obtains all the values f_i, r_i, s_i, r^T, s^T from the readership survey data. α_2 is the only statistic specific to group i which is determined from the re-interview survey.

Currently for newspapers and specific issue publications the system estimates α_2 as d*/D* where d* is the number of once only readers in group i and D* is the total number of once only readers in the reinterview survey. This is equivalent to first using the raw casualness

$$\gamma_r^* = \frac{d^*/N_i^*}{r_i^* s_i^*}$$

where N_i^* is the number of group i respondents in the reinterview survey, and then using an adjustment factor of

$$\frac{f_i^*}{f_i} \frac{r_i^* s_i^*}{r_i s_i} \frac{r^T s^T}{r^{*T} s^{*T}}$$

to correct for differences in frequency and readership between the reinterview survey data and the readership survey data. (The asterisks in these equations denote figures derived from the reinterview survey data, figures without asterisks being taken from the readership survey data). To show this we substitute the equations

$$d^* = 2 r_i^* s_i^* \gamma_r^* N_i^*$$

$$D^* = 2 r^{*T} s^{*T} \gamma^T N^{*T}$$

$$\alpha_2 = \frac{d^*}{D^*}$$

in equation (1). Doing this we obtain

$$\gamma_i = \frac{1}{f_i} \frac{r_i^* s_i^* \gamma_r^* N_i^*}{r^{*T} s^{*T} \gamma^T N^{*T}} \frac{r^T s^T}{r_i s_i} \gamma^T$$

$$= \frac{1}{f_i} \frac{N_i^*}{N^{*T}} \frac{r_i^* s_i^*}{r_i s_i} \frac{r^T s^T}{r^{*T} s^{*T}} \gamma_r^*$$

$$= \frac{f_i^*}{f_i} \frac{r_i^* s_i^*}{r_i s_i} \frac{r^T s^T}{r^{*T} s^{*T}} \gamma_r^* \qquad \text{since } f_i^* = \frac{N_i^*}{N^{*T}}.$$

For first time read publications, the system first calculates D* using the correct weight for all possible response combinations, vis

Response combination	Weight
(0,0)	0
(0,1) (1,0)	1
(1,1)	0
(0,2) (2,0)	2
(1,2) (2,1)	-1
(2,2)	-4
(0,3) (3,0)	3
(1,3) (3,1)	-2
(2,3) (3,2)	-7
(3,3)	-12

The system then re-determines the weights so as to eliminate all the negative weights while keeping the total estimate of D* constant.

Values of d* and hence of α_2 are then determined using these revised weights.

Formula (1) is a very general formula and many different estimates of γ can be cast in its mould. In the opinion of the author the real focal point in the estimation of adjusted casualness lies not so much in the use of formula (1) but in the subsequent estimation of the paramenter α_2. However, while the current estimates may not be optimal and it may well be possible to improve on them in the future using techniques such as logistic regression, the degree of complexity of these techniques is almost certainly not warranted at this stage.

5. SMOOTHING

(5.1) Introduction

It is possible to use formula (1) to solve problem (a) of section (4.2) as well as problem (b), and the method used by the system is in fact mathematically equivalent. However it is conceptually simpler to approach the problem as an exercise in smoothing.

If γ_A is the casualness estimate obtained by taking $\alpha_2 = d^*/D^*$ in equation (1) we seek a smoothed casualness γ_{AS} of the form

$$\gamma_{AS} = \frac{\gamma_A + \Phi\gamma_S}{1 + \Phi} \qquad (2)$$

In this equation there are two variables which must be determined. These are the form of the smoothing casualness function γ_S and the smoothing factor Φ.

(5.2) The form of the smoothing casualness function.

Smoothing functions investigated have been of the form:-

λ Distributional Component + $(1-\lambda)$ Commitment Component

ie $$\frac{\lambda\,\gamma^T s^T}{s} + \frac{(1-\lambda)\gamma^T r^T}{r}$$

The first value of λ tried was $\lambda = s^T$, since this gives a local minimum when $r = r^T$, $s = s^T$. However this attempt was unsuccessful as were all attempts which involved the commitment component. With readerships most often in the 0-10% range, the 1/r term in this component proved virtually prohibitive.

Subsequent work have centred on the use of the distributional component alone. This has proved particularly good where readerships have been small (when the smoothing γ is very nearly constant). It has not been nearly as good when the readership is large as then it can introduce distortions.

(5.3) The smoothing factor Clearly the smoothing factor Φ needs to be small where the sample is large and large where the sample is small. Initially smoothing factors of the form

$$\Phi = k/\sqrt{D^*}$$

were tried. These however tended to introduce too great a distortion against prima facie commitment factors (eg women for women's magazines) in the well read magazines while apparently leaving too great a degree of random fluctuation in the less well read magazines. Subsequently factors of the form

$$\Phi = k/D^* \quad \text{and} \quad \Phi = k/D^{*2}$$

have been tried. The current version of the program has

$$\Phi = 500000/D^{*2}$$

This appears to give reasonably good results on most national publications and on small regional publications. It may not, however, be as good for State based publicatons when the readership is high and the sample base is relatively low. Examples of approximate degrees of smoothing obtained using this smoothing factor are:-

Women's Weekly	6%
New Idea	18%
Family Circle	27%
Cosmopolitan	48%
Mode	92%

6. WHERE TO NOW?

The current system represents a considerable improvement over what has previously been available. All figures are now determined on an equal objective pre-defined basis, whereas before they were subjectively smoothed, winsorized and otherwise processed in ways not usually known to the user. All breakdowns are fully consistent with one another and two issue reach is always conserved whereas before this was not so. The figures can be easily updated to account for changing patterns of readership and for shifts in population structure. This was previously impracticable. Similarly the data from new re-interview survey rounds can now be added to the system progressively in the same way that new readership data is progressively added to the survey data base. This was also previously impracticable. The new system is on-line; changes can be quickly implemented and there is scope for further automation (e.g. direct linkage to optimization routines).

It is acknowledged that the system as it now stands is not perfect. Changes, however, must not be made in isolation: by theoreticians divorced from practitioners. The development of a good system with an established, well documented technology, is in everyone's interest. Such development will only truly take place if the system is used; the results monitored and regular feedback established.

The system represents a considerable improvement over the past. It should therefore be used. Its use may result in suggested improvements which can be considered and if appropriate implemented. In this way real progress can continue.