ISSUE SPECIFIC ESTIMATION – MATHEMATICAL AND STATISTICAL ISSUES, PROCEDURES AND MODELS

Martin R. Frankel, Julian Baim, Michal Galin, Joseph Agresti, Mediamark Research Inc

I. Introduction and Background

In recent years the need for timely issue specific audience estimates has become more pressing. The fact that print has relied on and provided only average issue audience is often viewed as detrimental to the full integration of print in the media planning and evaluation process.

In early 2006 MRI began a continuous online survey of approximately 2,500 web-based interviews per week in order to gather information to be used in producing estimates of individual issue audience approximately 6-8 weeks (in the case of weeklies) and 12-16 weeks (in the case of monthlies) after the "on-sale" date. A full description of the survey system, the sampling system, the questionnaire and the survey results are covered in a companion paper. This paper's focus is the basic mathematical, statistical and inferential issues, procedures and models that are being used to transform survey results into actual issue specific audience estimates.

The basic technical issues to be discussed in this paper include:

- Integration of audience data based on a non-probability Internet panel sample with the MRI currency AIR measure to produce issue specific audience estimates that are consistent with the currency measure.

- The impact and removal of memory decay (forgetting) on issue specific audience estimates and choice of an estimation function.

- The use of recently developed statistical techniques and methods (James-Stein, Empirical Bayes, borrowing strength) to increase reliability by removing a portion of sampling error from the issue specific estimates.

- The procedure used to compensate for missing date specific data for issues that are partially missing.

- The interaction of issue specific estimates and existing models of issue velocity (accumulation).

II. Integration of audience data based on a non-probability Internet panel sample with the MRI currency AIR measure to produce issue specific audience estimates that are consistent with the currency measure.

In developing an estimation strategy for producing issue specific audience estimates, we considered the relevant features and costs associated with four basic data collection options. In addition to sampling and response rate issues, we considered total costs and the ability to control the timing of data collection. The following data collection options were considered:

A. Face-to-Face interviewing conducted in households based on a probability sample with a high response rate, high cost and partial control over timing.
B. Telephone Interviews conducted with a full probability sample with a medium to low response rate, medium cost, and control over timing.
C. Mail Interviews conducted with a full or partial probability sample with a medium to low response rate, medium to low cost and partial control over timing.
D. Internet interviews conducted with a non-probability sample with a low response rate, low cost, and full control over timing.

It was also recognized that given the basic audience sizes for which estimates were required, (i.e. population proportions between 0.005 and 0.250 per issue) sample sizes of between 10,000 and 30,000 respondents would be required. For example, in the US, if a sample size of 20,000 is used, a publication with an audience of one million (1,000,000) will have a relative standard error of 10%.[i]

If we assume, a per issue sample size of 20,000, then the estimation of issue specific audiences for a weekly requires a sample size of n= 52 x 20,000 = 1,040,000 for a single year. However, we decided that it would be possible to measure eight (8) different issues of a weekly with the same sample of persons, which would span a period of about two months from the title's initial publication data. By measuring eight different weekly issues at the same time, we could reduce our required sample size to approximately n= 1,040,000/8 = 130,000.

The sample size requirement of approximately 130,000 interviews per year and our assessment of the financial realities of the US print measurement market restricted our data collection choices to mail and Internet. The use of a mail questionnaire would allow us to come closer to a probability sample, but we had significant concerns about our ability to implement a "screen-in, then read" approach for 50 weekly and 150 monthly titles; eight (8) issues for weeklies and four (4) issues for monthlies. Further, based on prior research with subscriber and audience accumulation studies, we recognized that controlling of "timing" would present serious difficulties.

Internet data collection seemed an ideal way to apply a "screen-in / then ever read the issue" approach to the 50 weekly and 150 monthly titles, since it was possible to control both timing (to the level of day) and stimuli by showing relatively large pictures of issue covers to the restricted sub-sample of individuals who screened-in for the title. Rather than being faced with a 200 page mail questionnaire with skip instructions, the respondent would only see the 15 or 20 pages that were relevant.[ii]

The serious drawback associated with the use of the Internet was the non-probability nature of sample selection process. We knew from our previous research, based on nearly 100,000 interviews[iii], that the absolute audience levels produced by both weighted and unweighted Internet samples were not consistent with those produced by the MRI annual "probability based" average issue audience (AIR) readership survey. However, analyses from the first 37 weeks of Internet data collection provided very strong evidence that it was possible to derive title specific, statistically consistent and reliable, measures of relative issue-to-issue variation. By applying these Internet derived measures of relative issue-to-issue variation to probability sample based estimates of average issue audience, it was possible to produce useful estimates of issue-to-issue audience.[iv]. As is more fully described below, the evidence consisted of the behavior of issue specific recognition measures over weekly replications of the non-probability sampling process. It was found that measures of "higher than average" and "lower than average" readership for specific issues of a title, taken over eight weekly samples, showed consistency and reliability far in excess of what would occur if we were observing random, haphazard or quasi-random noise or variation. That is, if one were simply looking at haphazard responses or noise in the week-to-week sampling process, we would expect the above and below average behavior of specific issues to be haphazard as well. This is particularly true given the issue overlap from week-to-week.[v] We did not observe this type of haphazard variation on an issue-by-issue basis. We found that issue specific claimed reads showed substantial consistency in above and below the mean behavior. It was these observed results that provided evidence that Internet based sampling was reflecting issue-to-issue variation rather than random or haphazard noise.

We applied a number of statistical tests to examine this non-random behavior. The most convincing, in terms of intuitive and statistical significance, was based on the non-parametric sign test.[vi] This was accomplished as follows:

The null hypothesis H_0, states that there is no issue-to-issue variation. That is, we assume, for the purposes of the statistical test, a constant issue-to-issue readership level over time. Thus any variation observed is simply due to sampling variation and noise. The alternative hypothesis H_1, states that at least one, but possibly more, of the issues do not have the same true readership levels as the others.

Suppose we consider 30 successive issues of a weekly magazine for which measurements of cumulative issue readership are taken in each of the eight (8) weeks after publication. Under our null hypothesis of no issue-to-issue variation, we would expect that the measured audience (cumulative) for each of the eight (8) weeks to be the same, except for sampling error, over the 30 different issues.

If we let P_{ij} denote the true cumulative readership of issue i of a particular title, j weeks after publication. The null hypothesis may be stated as

$$H_0: \ P_i = P_{1j} = P_{2j} = P_{3j} \ \dots P_{30j}, \text{ for each j=1,..8}$$

That is if we focus on a particular week after publication, we hypothesize that all issues (i=1,..,30) have the same cumulative readership. This hypothesis is assumed to hold for each of the eight (8) week of observation (j=1,…8)

This null hypothesis implies that within a given week after publication the observed issue-to-issue variation should be random and simply a result of sample, but not true issue-to-issue variation. Thus, if we examine the cumulative readership at a given week, across the 30 issues, we should see variation that is simply the result of sampling error, but not true issue variation.

In order to carry out the test we first define CA_{ij} as the sample reported "read this issue audience" for the ith issue, j weeks after publication.

Table I shows the values CA$_{ij}$ for 30 issues of a weekly magazine measured over 8 weeks after publication. We note that these are actual observed values for female readers of one of the weeklies in the study. For example, in this table, find the entry in the second data row (Issue #2, Date 6/26/2006) and third column (Weeks after publication=3) is 0.123890. This means that 12.4% of female respondents claim to have read the 6/26/2006 issue, three weeks after its publication.

TABLE I - READ THIS ISSUE AUDIENCE									
		Week After Publication							
Issue #	Issue Date	1	2	3	4	5	6	7	8
1	6/19/2006	0.15428	0.18592	0.19688	0.19294	0.23159	0.25537	0.26026	0.26110
2	6/26/2006	0.10465	0.12540	0.12389	0.13414	0.15390	0.15390	0.17797	0.18453
3	7/3/2006	0.08444	0.10027	0.14453	0.15920	0.13506	0.15485	0.14289	0.14447
4	7/10/2006	0.11056	0.17617	0.22689	0.21606	0.23264	0.24415	0.23813	0.23336
5	7/17/2006	0.10214	0.15353	0.15119	0.16496	0.17434	0.16765	0.17861	0.16270
6	7/24/2006	0.15072	0.17862	0.19592	0.20623	0.20761	0.22737	0.23722	0.23804
7	7/31/2006	0.10865	0.14045	0.16111	0.13881	0.16133	0.15589	0.15561	0.17908
8	8/7/2006	0.11444	0.14356	0.13756	0.14547	0.17325	0.16638	0.16819	0.17223
9	8/14/2006	0.10368	0.14498	0.17039	0.16472	0.18687	0.20778	0.21019	0.19336
10	8/21/2006	0.10131	0.12755	0.14017	0.14621	0.15097	0.15546	0.14535	0.14860
11	8/28/2006	0.08736	0.10776	0.12484	0.13660	0.14114	0.11077	0.13883	0.11392
12	9/4/2006	0.10567	0.17784	0.18721	0.20526	0.21247	0.23824	0.19376	0.20589
13	9/11/2006	0.07862	0.11004	0.12361	0.09788	0.11549	0.11290	0.12090	0.12463
14	9/18/2006	0.12116	0.17568	0.16770	0.18997	0.21394	0.21738	0.25661	0.24229
15	9/25/2006	0.12770	0.15733	0.17650	0.16944	0.16877	0.20750	0.19502	0.22884
16	10/2/2006	0.10455	0.14662	0.17028	0.16946	0.20897	0.19706	0.21136	0.21059
17	10/9/2006	0.13861	0.18368	0.18398	0.21135	0.21146	0.21063	0.20494	0.23858
18	10/16/2006	0.09087	0.11208	0.15325	0.16668	0.17528	0.15725	0.17913	0.16683
19	10/23/2006	0.08098	0.12396	0.12882	0.12044	0.12496	0.15255	0.13122	0.13065
20	10/30/2006	0.13185	0.17801	0.19294	0.18819	0.20089	0.20669	0.18741	0.20308
21	11/6/2006	0.11842	0.15524	0.17016	0.18831	0.17797	0.20910	0.20076	0.17796
22	11/13/2006	0.10915	0.16268	0.20467	0.20920	0.23813	0.23694	0.20661	0.22438
23	11/20/2006	0.08468	0.12899	0.12399	0.11757	0.13230	0.11429	0.13016	0.14728
24	11/27/2006	0.11899	0.15574	0.15646	0.17996	0.15536	0.18411	0.20012	0.20646
25	12/4/2006	0.11236	0.18403	0.19674	0.19903	0.23364	0.24194	0.25716	0.26412
26	12/11/2006	0.10428	0.11839	0.12305	0.14475	0.15169	0.14873	0.15656	0.17341
27	12/18/2006	0.07345	0.08585	0.09933	0.10315	0.11397	0.12468	0.12128	0.11014
28	12/25/2006	0.08112	0.12332	0.15050	0.16282	0.16661	0.16854	0.18527	0.21247
29	1/8/2007	0.10619	0.14706	0.15684	0.15562	0.15071	0.17177	0.15152	0.16148
30	1/15/2007	0.09410	0.11720	0.13401	0.14209	0.14149	0.13639	0.13552	0.14314
	Average	0.10683	0.14426	0.15911	0.16422	0.17476	0.18121	0.18262	0.18679
	Median	0.10516	0.14580	0.15665	0.16484	0.17101	0.17015	0.18220	0.18180

Reading across the row we find the reported "read this issue audience" for the 6/26/2006 issue from weeks 1-8 after publication. If we read up and down the column for this entry we find the third week after publication "read this issue audience" for the thirty (30) issues from June 19, 2006 to January 15, 2007.

At the bottom of each column we show the mean and median values. Examination of this summary shows a general increase in both median and mean "read this issue" levels over successive weeks with asymptotic behavior in weeks 7 and 8. When compared with expected audience accumulation, we see that reporting error associated with forgetting seems to be present. This observation will be discussed in the next section.

Under the null hypothesis H$_0$ of no issue-to-issue variation, we expect that for a particular column (week after publication) as we look up and down the column we should see haphazard or random variation. That is, when we compare the particular issue level to other issues measured at the same point in time after publication we should see random variation due to sampling error. If we focus on a particular column, we do see variation from issue-to-issue.

However, if we focus on a particular issue and the mean or median for all issues we find that issues that start with higher than average (or median) readership in the first week tend to retain this behavior over successive weeks. Conversely, issues having lower than average (or median) readership in the first week tend to retain this behavior over successive weeks. If this is the case, then our null hypothesis of no issue-to-issue variation is not supported.

In order to conduct the formal test of persistence (non-randomness) in relative levels across the eight measured weeks we construct a "test statistic," and then examine the behavior of this test statistic under the null hypothesis H_0.

The test statistic is actually a vector of 9 values and is constructed as follows. Within each of the 8 columns (weeks) of the table, we examine each of the 30 sample values and assign a code of 1 if the value "read this issue" level is above the median over all issues and a value of 0 if the read this value is below the median. We use the median rather than the mean to assure that there will be 15 above the median values (i.e. 1's) in each column and 15 below the median values (0's) in each column. For example, the second row and third column (3rd week after publication) has value of 0.123890. The median overall all issues of week 3 is 0.156652, thus the value of 0 is assigned in Table II (since 0.123890 is below the median). Table II shows the corresponding zero or one values for all 8 x 30 = 180 cells in Table I. It should be noted that each column of Table II has 15 ones and 15 zeros. For each week after publication, half of the issues show "sample read this issue" values above the median and half are below the median. The last column in Table II shows the sum of zeros and ones (i.e. the number of ones) for each of the 30 issues. Thus, for issue #2 the number of ones is equal to one (1) which means that the reading level is above the median for only one of the eight weeks.

TABLE II - TEST STATISTIC CONSTRUCTION (ABOVE AND BELOW COLUMN MEDIAN)

Issue #	Week After Publication								SUM
	1	2	3	4	5	6	7	8	SUM
1	1	1	1	1	1	1	1	1	8
2	0	0	0	0	0	0	0	1	1
3	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	8
5	0	1	0	1	1	0	0	0	3
6	1	1	1	1	1	1	1	1	8
7	1	0	1	0	0	0	0	0	2
8	1	0	0	0	1	0	0	0	2
9	0	0	1	0	1	1	1	1	5
10	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0
12	1	1	1	1	1	1	1	1	8
13	0	0	0	0	0	0	0	0	0
14	1	1	1	1	1	1	1	1	8
15	1	1	1	1	0	1	1	1	7
16	0	1	1	1	1	1	1	1	7
17	1	1	1	1	1	1	1	1	8
18	0	0	0	1	1	0	0	0	2
19	0	0	0	0	0	0	0	0	0
20	1	1	1	1	1	1	1	1	8
21	1	1	1	1	1	1	1	0	7
22	1	1	1	1	1	1	1	1	8
23	0	0	0	0	0	0	0	0	0
24	1	1	0	1	0	1	1	1	6
25	1	1	1	1	1	1	1	1	8
26	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	1	1	2
29	1	1	1	0	0	1	0	0	4
30	0	0	0	0	0	0	0	0	0
SUM	15	15	15	15	15	15	15	15	

Under the null hypothesis the, zero-one values will be distributed at random within each week. That is, if there is no issue-to-issue variation, then within a week we expect the pattern of above and below the median measurements should be distributed across issues at random.

If the distribution of zeros and ones is random or haphazard across the various issues, within each week, we expect that if we add the zeros and ones across the weeks, within a row, the results taken over the various issues should behave like a random coin toss. In effect, within each issue, the chance of getting a zero or one each week should be 0.5, and the results of one week should be independent of the next week. Under the null hypothesis we expect that the sum of zeros and ones across the eight weeks will be close to four. This is the same as saying that if we flipped a fair coin eight times, we expect the number of head (or tails) to be 4. If the number of heads were 3 or 5, this would not be a surprise, but if we obtained no heads or all eight heads, we would be very surprised, and indeed convinced that the coin was not fair.[vii]

In symbolic terms, we define a variable (shown in Table II)

D_{ij} = 1 if $CA_{ij} > MEDIANj(CA_{ij})$ and
    = 0 if $CA_{ij} < MEDIANj(CA_{ij})$, where

$MEDIANj(CA_{ij})$, is the median across the index i=1,…30

Furthermore let

X_i = $\sum_j D_{ij}$, the sum taken over the subscript j, of D_{ij}

Under the null hypothesis, the distribution of the vector of values

$X = \{ X_1, X_2, X_3, …. X_{30}\}$ should follow the binomial distribution with p=0.5 and N=8.

The right most column of Table II shows the sums X_i for the 30 issues.

Table III shows the expected (under the null hypothesis) and the actual frequency distribution of the values of the vector $X = \{ X_1, X_2, X_3, …. X_{30}\}$

TABLE III-EXPECTED ACTUAL DISTRIBUTION p<0.00000001			
X	B(X|8,0.5)	Expected	Actual
0	0.003906	0.12	9
1	0.031250	0.94	1
2	0.109375	3.28	4
3	0.218750	6.56	1
4	0.273438	8.20	1
5	0.218750	6.56	1
6	0.109375	3.28	1
7	0.031250	0.94	3
8	0.003906	0.12	9
SUM	1.000000	30.00	30

The first column of the table shows the possible values for $X_i = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$, the second column shows the binomial probability distribution for X, which describes the expected probability distribution under the null hypothesis. These probabilities are computed by the standard binomial formula

Prob $\{X| N=8, p=1/2\} = B(X|8,0.5) = [N! / (X!*(N-X)!)] * p^X * (1-p)^{N-X}$

$$= [8! / (X! * (8-X)!)] * 0.5^X * 0.5^{8-X}$$

Column three shows the expected frequency distribution of the 30 X_i values under the null hypothesis and is obtained by multiplying the column of probabilities by 30. The final column shows the actual empirical distribution of 30 observed X_i values.

Looking at the row of Table III corresponding to X=4, (fifth row of data) we see the expected number of times that X=4 is 8.2(column 3). This means that under the null hypothesis we should observe a sum in the last column of Table II) equal to 4 approximately 8 of 30 times. We actually observe this sum equal to 4 only once. This corresponds to the actual value of 1 in column 4 of Table III. Furthermore, under the null hypothesis of no issue-to-issue variation, we would not expect to see a sum equal to 8 at all (expected value 0.12). We would also not expect to see the sum of zero. The sum of 8 occurs 9 times and the sum of zero also occurs 9 times. When we compute the probability of observing the actual frequency distribution, given the null hypothesis we find that the chance of observing this is substantially less than 0.00000001[viii]. This means that under any type of decision rule, we reject the null hypothesis of no real issue-to-issue variation and accept the alternative that there is issue-to-issue variation.

This testing was carried out for various magazines and showed similar results time after time. It was our conclusion that our survey results were not simple sampling error or noise, but rather actual issue-to-issue variation.[ix] As a result we concluded that the "relative" levels of audience size provide useful issue specific levels when used in conjunction with appropriate average issue audience levels.

III. The impact and removal of memory decay (forgetting) on issue specific audience estimates and choice of an estimation function.

At the outset of our design process we recognized that one of the sources of survey error in our measurement process would be related to the dimension of time. Since we were asking for any reading or looking into each of the issues that were being shown, we expected that the major source of this type of error would be memory decay or forgetting, rather than telescoping since we were asking for an "ever read" this issue response. For weeklies, the issues would be from one to eight weeks old, while for monthlies, the issues might be between 1 and 16 weeks old. Given that we were not attempting to measure total audience for each issue, but rather the relative size of the total audience (in comparison to other issues of the same age at the time of measurement), we felt that there were three basic approaches the might be used in the initial development of relative audience. These three basic approaches described as simple average, time-compensate average, and indexed.

Following the same notation used above, for a particular title, let CA_{ij} denote the reported "read this issue audience" for the ith issue, j weeks after publication

The general function for producing an estimate of the relative audience of the ith issue of a publication may be expressed as

$$RA_i = \sum_j K_j * CA_{ij}$$, where K_j's are constants, but may vary by j

The three basic approaches to estimating the relative audience of the ith issue of a publication involve three different choices of the values K_j. The simplest possibility (which we describe above as simple average) results if all values of K_j are the same or constant. In the case of weeklies, with 8 observations associated with weeks after publications, we have $K_j = 1/8$. For monthlies, if we use the full 16 weeks of observation we have $K_j = 1/16$. In this case, the relative audience for issue i is computed by summing the reported "read this issue" audience levels across all of the weeks that the issue is included in the survey and dividing by J, the number of weeks included.

The second possible set of values that was considered involved the use of the values proportional to the inverse of the expected percentage of cumulative audience for the publication j weeks after publication for K_j. In this case the values associated with early weeks after the publication data would receive increased weight, based on the inverse of expected audience accumulation. If 50% of the reading of an issue was expected to take place in the first week and 75% in the first two weeks, the relative weights for the first two weeks would be 2.0=1/0.50 and 1.33=1/0.75 respectively.

The third possibility involves the use of the inverse of the average CA_{ij} over all of the I issues measured. This is equivalent for creating an index for each issue (relative to the overall average) within each week after publication and then taking the average of these indices.

Define $ARA_j = \sum_i CA_{ij} / I$, the average over the I different issues within the jth week.

Then we have $K_j = [1 / (I * ARA_j)]$

We chose the third approach over the first two because we felt that the averaging of "relative variation" as expressed by the index levels would be more robust than the other two options. With the first option the later weeks of measurement play a more dominant role because of their increasing magnitude and, since we were not estimating total audience but rather relative variation, we wanted to give each week equal impact. We also rejected the second estimator for the same reason coupled with the fact that we could not be sure of the mixture of in-home and out-of-home readers that is so critical in accumulation levels. Recognizing that there are errors of various kinds the may change over time; we felt that allowing the variation in each week to have equal impact was the most robust choice.

**IV.	The use of recently developed statistical techniques and methods (James-Stein, Empirical Bayes, borrowing strength) to increase reliability by removing a portion of sampling error from the issue specific estimates.**

In the early 1970's Bradley Efron and Carl Morris published a series of papers that had a profound impact on the world of statistics[x]. In these papers they examined the implications and practical applications of a basic result on the "statistical admissibility" of estimates that had been first published by Professor Charles Stein in 1956 and then expanded with Willard James in the early 1960's. James and Stein (1961) proved that when making inferences about a series of three or more parameters, the simple means, which are both the BLUE, Best Linear Unbiased Estimator, as well as the maximum likelihood estimators, are "inadmissible." That is, it is possible to find another estimators with smaller expected squared error throughout the full range of possible parameters. Efron and Morris made this finding more understandable by using an example from the game of baseball. The data they used appear in Table IV. This table contains the batting averages for the 18 Major League Baseball players who had exactly 45 "at bats" as of April 26 or May 3, 1970 issues of the *New York Times*. In the first data column the batting averages after 45 games are shown. The second data column shows the batting average for the players at the end of the season. The third column shows an estimate that Efron and Morris termed the "James-Stein" estimator. The James-Stein estimate essentially shrinks all of values closer to the overall average. The amount of shrinkage depends upon how different the values are from the overall average and in this case involves the arc sin transformation. When the James-Stein estimator is compared to the MLE (the maximum likelihood estimator) consisting of the first n=45 "at-bats" the James-Stein superiority is quite evident. If we take the mean of the absolute errors it is 0.055 for the MLE and 0.026 for the James-Stein. If we take the square root of the mean of the squared errors is 0.064 for the MLE and 0.035 for the James-Stein.

TABLE IV - BATTING AVERAGES 18 PLAYERS

PLAYER	First 45 (MLE)	Full Season	James-Stein	Error MLE	Error J-S
Celmente	0.400	0.346	0.290	0.054	-0.056
F. Robinson	0.378	0.298	0.286	0.080	-0.012
F. Howard	0.356	0.276	0.282	0.080	0.006
Johnstone	0.330	0.222	0.277	0.108	0.055
Berry	0.311	0.273	0.273	0.038	0.000
Spencer	0.311	0.270	0.273	0.041	0.003
Kessinger	0.289	0.263	0.268	0.026	0.005
Alvarado	0.267	0.210	0.264	0.057	0.054
Santo	0.244	0.269	0.259	-0.025	-0.010
Swoboda	0.244	0.230	0.259	0.014	0.029
Unser	0.222	0.264	0.254	-0.042	-0.010
Williams	0.222	0.256	0.254	-0.034	-0.002
Scott	0.222	0.303	0.254	-0.081	-0.049
Petrocelli	0.222	0.264	0.254	-0.042	-0.010
Rodriguez	0.222	0.226	0.254	-0.004	0.028
Campaneris	0.200	0.285	0.249	-0.085	-0.036
Munson	0.178	0.316	0.244	-0.138	-0.072
Alvis	0.156	0.200	0.239	-0.044	0.039

The James-Stein estimator is defined as follows:

Let Y_i be the batting average of player i, i=1,18 (k=18) after n=45 at bats. Further define $X_i = f_{45}(Y_i)$ where $f_n(y) = (n)^{1/2}$ arc sin (2y-1). The James-Stein estimator is defined (Efron and Morris, 1975) as:

$$\partial_i^1(X) = \bar{X} + (1 - \frac{k-3}{V})(X_i - \bar{X})$$

$$V = \sum(X_i - \bar{X})^2,$$

As Efron and Morris noted at their presentation before the Royal Statistical Society (1973) "The reaction of the statistical community to this *tour de force* has been generally hostile, the usual suggestion being that this is some sort of mathematical trick devoid of genuine statistical merit." However, over the past 25 or so years, this estimation strategy has now been recognized as a great advance and is often described as "borrowing strength" to non-statistical audiences and "empirical Bayes" or "shrinkage estimation" to statistical audiences. While, this procedure had not found general acceptance in the field of media audience measurement, we feel that the measurement of issue specific audiences provides a situation where this type of estimation procedure may be used in a way that borrows strength from within a publication's issues but not across publications. As a result, we are not using the audience of U.S News and Newsweek, to modify the audience of Time. Rather we are removing expected random sampling error when we evaluate the performance of one issue of Time against other issues of Time, and separately removing the expected random sampling error when we evaluate the performance of Newsweek against other issues of Newsweek. And so on.

To make our procedures transparent and to avoid complex transformations, our estimation process proceeds as follows. Within each title we first develop the standard estimates of issue-to-issue variation by taking the average index of the issues relative performance (against other issues) across the weeks of measurement. Next, we estimate the amount of random sampling error or variation that we would observe in these issue-to-issue measures, under the assumption that there was no true issue-to-issue variation. In other words, we estimate the amount of sampling error we would find if we were measuring the same quantity from week-to-week. We compute the "most probable" value for this sampling error and subtract it from the actual observed sampling error to obtain an estimate of "expected variation with random sampling error removed." Finally we apply a James-Stein type of shrinkage function to the issue-to-issue values so that the variation conforms to the "expected variation with random sampling error removed."

For a particular publication and a particular demographic group we begin with the matrix of values CA_{ij}, which represent the "read this issue audience sample proportions" for the ith issue, j weeks after publication.

Within a given week j, we compute the empirical variance of the CA_{ij} values taken over the I different issues being considered. So, we have

$$VAR(C_j) = \frac{\sum_{i=1}^{I} CA_{ij}^2 - [(\sum_{i=1}^{I} CA_{ij})/I]}{I-1}.$$

We also compute the variance that would be expected if there were not true issue-to-issue variation but only random sampling variation. This value is computed as

$$RSVAR(C_j) = \frac{\bar{C}_j(1-\bar{C}_j)}{n}$$

$$\bar{C}_j = \sum_{i=1}^{I} CA_{ij})/I$$

where n=effective sample size for the measured CA_{ij}

Because we will be averaging across survey weeks, we transform these two measures into coefficients of variation as follows

$$CV(C_j) = \frac{\sqrt{VAR(C_j)}}{\bar{C}_j}$$

$$CVRS(C_j) = \frac{\sqrt{RSVAR(C_j)}}{\bar{C}_j}$$

Finally, within each week we compute the value of the expected true coefficient of variation as

$$ETCV(C_j) = CV(C_j) - PZ \bullet CVRS(C_j)$$

PZ denotes the probable error point under a standard normal distribution and is equal to 0.6745.[xi]

The values of ETCV(C_j) are averaged over the 8 or 16 weeks of observation to produce the overall ETCV(C).

The actual "shrinkage" or borrowing strength estimates are formed as

$$ISI_i = IS_i^r$$, where r is found, by iterative methods, so that

$$CV(ISI) = ETCV(C)$$, where CV(ISI) is the coefficient of variation of the ISI_i values.

Table V shows the computation of ETCV(C) and the components CV(C_j), CVRS(C_j) and ETCV(C_j) for the data shown in Table I. Table VI shows the values of $ISI_i = IS_i^r$. It should be noted that in this example r was found to be r=0.85885.

TABLE V: CV, DVRS and ETCV

Week	CV(Cj)	CVRS(Cj)	ETCV(Cj)
1	0.19006	0.07728	0.13794
2	0.19609	0.06509	0.15218
3	0.19078	0.06144	0.14934
4	0.19822	0.06029	0.15755
5	0.21127	0.05808	0.17210
6	0.23381	0.05681	0.19549
7	0.22607	0.05654	0.18793
8	0.23230	0.05577	0.19468
Average			0.16840

TABLE VI: RAW AND FINAL INDICIES

Issue #	Issue Date	IS	ISI
1	6/19/2006	134	128
2	6/26/2006	89	91
3	7/3/2006	82	84
4	7/10/2006	128	124
5	7/17/2006	97	97
6	7/24/2006	127	123
7	7/31/2006	93	94
8	8/7/2006	95	95
9	8/14/2006	106	105
10	8/21/2006	86	88
11	8/28/2006	75	78
12	9/4/2006	117	114
13	9/11/2006	69	72
14	9/18/2006	121	118
15	9/25/2006	110	109
16	10/2/2006	108	107
17	10/9/2006	122	119
18	10/16/2006	92	93
19	10/23/2006	77	80
20	10/30/2006	115	113
21	11/6/2006	108	107
22	11/13/2006	121	118
23	11/20/2006	76	79
24	11/27/2006	105	104
25	12/4/2006	128	124
26	12/11/2006	87	88
27	12/18/2006	64	68
28	12/25/2006	95	96
29	1/8/2007	93	94
30	1/15/2007	81	83

In general, the amount of shrinkage depends upon a number of factors, including the sample size associated with the demographic group and the magnitude of the values of C_j.

VI. The procedure used to compensate for missing date specific data for issues that are partially missing.

The logistics associated with implementation of a study that updates the covers of more than 200 magazines on a weekly basis is both complex and daunting. It depends upon the cooperation and interaction with a number of publishing organizations. As a result of this it is possible that data may be missing for a single week of magazine issue. Rather than eliminate the issue because of a missing week, we have implemented a procedure for imputing this missing data. In order to do this we have relied on some fundamental results and methods that were first developed and reported in the early days of digital computers at Rothamsted Experimental Station in the UK[xii]. The process is iterative and begins by computing an overall mean as well as row and column means from the data matrix using the non-missing weeks and issues. Using these means, a value is estimated for each missing entry in the data matrix. The process of determining means is repeated and revised estimates are placed in the missing data positions. This process continues until convergence is obtained.

VII. The interaction of issue specific estimates and existing models of issue velocity (accumulation).

Our initial development of issue specific audience estimates has followed the implicit assumption that the rate of audience accumulation is approximately the same from issue-to-issue. We believe that this basic assumption is probably somewhat of an oversimplification, but will provide useful results in both planning and evaluation. We believe that a refinement of this assumption will require several years of data. This will allow us to examine seasonality and departures from seasonality. At the present time, our plans are to examine issues that have different levels of overall accumulation and determine the degree to which the steps in the accumulation process are statistically the same or different from the others. For example, if we select issues that appear to produce larger than average audiences, we will then examine whether or not the audiences in the first weeks are statistically different from issue-to-issue. If these differences are present, then we feel that this supports the assumption of differential rates of accumulation. If all of the issues that produce larger than average audiences seem to accumulate at the same rate in our observed data, this tends to support the assumption of similar rates of accumulation. At this point we feel that we do not have sufficient data to carry out this evaluation. We will have the data after several yearly cycles.

VIII. Summary and Conclusions

In the US, the founding fathers of magazine audience research tried, unsuccessfully, to develop measurements of the issue specific audiences of magazines. The fact that this development was not possible, was not a reflection of human brain power, but rather of the limitations of technology and statistical methods. We believe that technological advances associated with the Internet, the World Wide Web, and the personal computer as well as advances in statistical theory and practice has made the development of issue specific audience estimates a practical reality.

References:

i The relative standard error of a proportion p from a simple random sample of size n is relse(p) = se(p)/p = [p(1-p)/n]1/2 / p = [(1-p)/pn]1/2. A title with an AIR of 1,000,000 means that that population proportion of readers is approximately p=0.005 (i.e. 0.5%). Substituting p=0.005 and n=20,000 we find the relative standard error is [(1-0.005)/0.005x20000]1/2 = 0.09975 (10%).

ii Because of the automated logic that is built in to an Internet administration, a respondent who screened into 15 titles would only be shown the 15 or so pages of relevant issue covers and would not have to work through a 200 page mail questionnaire containing all of the issues (200 titles x 8 or 4 issues) covered in that week.

iii Martin Frankel, Julian Baim, Michal Galin and Michelle Leonard (2003), "Measurement of Magazine Readership Via the Internet," Worldwide Readership Symposium-2003, Cambridge, Massachusetts, Session Papers, p.131-148. Julian Baim, Michal Galin and Martin Frankel (2005), "Title Confusion: The Impact of Response Error on Competitive Pairs," Worldwide Readership Symposium-2005, Prague, Session Papers, p.251-274.

iv We initially conduced the analyses described below after approximately 24 weeks and shared preliminary results with a number of interested people and groups. Our more formal evaluation and more definitive planning for data release occurred after a period of 37 weeks from initial data collection.

v In addition to the test described below we also simulated the impact of sample related differences over the 8 issues measured by each sample. We found out tests to be conservative against this alternative. In other words, increase sample level effects would tend to produce more random variation.

vi George W. Snedecor and William G. Cochran (1971), Statistical Methods, Sixth Edition, Ames Iowa, The Iowa State University Press, p 125-129.

vii It should be noted that in the case of correlation across the issues within a survey week, the behavior of sums within an issue, across weeks will become random. Thus even if we start with issue specific variation but add a random term to the observed readership of all magazines measured in a survey week, the behavior of these sums becomes more binomial. This has been confirmed by a number of simulations.

viii The evaluation of this probability requires the use of a procedure known as the Fisher-Freeman-Halton test. It is actually an extension of "Fisher's exact test" for 2 by 2 tables. The usual chi-square test in excel produces a p value of 3.3552E-290. However, several test chi-square test assumptions are not met. We used several simulation procedures and several cell-collapse procedures to obtain "robust" significance levels. All were in excess of 0.00000001, and most were much smaller.

ix It should be noted that even in the case of non-random behavior in the samples from week-to-week, the expected variation in results and thus the distribution of the number of times an issue was above or below the median should follow the binomial distribution, with large frequencies observed at 3, 4 and 5, rather than the U shaped results that were actually observed.

x Efron, B and Morris, C., (1973) "Combining Possibly Related Estimation Problems, (with Discussion)" Journal of the Royal Statistical Society. Series B, Vol. 35. No. 3, pp. 379-421.

Efron Bradley. and Morris, Carl.,(1975) "Data Analysis Using Stein's Estimator and its Generalizations," Journal of the American Statistical Association, Vol. 70.,No. 350, pp. 311-319.

Stein, C.(1956) , "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," Proceedings of the Third Berkeley Symposium Mathematics, Statistics, Probability., Vol 1. University of California Press, Berkeley, California, pp. 197-206.

Stein, Charles and James, Willard. (1961), "Estimation with Quadratic Loss," Proceedings of the Fourth Berkeley Symposium Mathematics, Statistics, Probability. Vol. 1. University of California Press, Berkeley, California, pp. 361-399.

xi Under a standard normal distribution one half of the total probability falls between Z=-0.6745 and Z=+0.6745. Multiplying CVRS by 0.6745 gives us the most likely value of the observed coefficient of variation under the assumption of no issue-to-issue variation.

xii Michael Healy and Michael Westmacott, (1956) "Missing Values in Experiments Analysed on Automatic Computers," Applied Statistics, Vol 5. No. 3, pp. 203-206.