SAMPLING ERRORS VERSUS REAL CHANGES IN MAGAZINE AUDIENCE RESEARCH

Roland Soong and Lindsey Draves, KMR Group

1. Background

A print audience research study is a survey that measures the readership of magazines and/or newspapers. Such a survey may be repeated periodically, such as on an annual basis. The audience numbers for a print title may change from one survey to the next due to a number of reasons.

First of all, the changes in the audience numbers may represent real changes in the world. Reading habits may have changed, the print title may have undergone significant editorial changes and circulation strategies may have been altered.

Secondly, there may have been changes in the survey methodology. For example, the length of the questionnaire may have been increased in order to measure more print titles, and this may adversely impact the results.

Thirdly, surveys have sampling errors in the sense that different survey samples will produce somewhat different results.

The usual strategy in understanding the audience changes is to test for statistical significance and place some bounds on the magnitude and direction of the methodological changes based upon prior research. Afterwards, we do some detective work to consider the nature of the 'real' changes (that is, those that have been found to be statistically significant beyond what methodological changes would allow).

2. Statistical Significance Testing

For this paper we analyze an empirical example of a readership study in order to understand changes in audience numbers.

The MARS OTC/DTC Pharmaceutical study is an annual print readership survey that has been conducted every year from 2001 through 2007. The survey methodology has remained the same: the field period is from January to March, the sample frame is ordered from the same list of names and addressed compiled and maintained by Acxoim Corporation, the questionnaire is 20 to 24 pages long, and the intab sample size is between 21,000 to 23,000. Over these seven years, 87 print titles (83 magazines and 4 newspapers) were measured every year.

The statistical significance test requires us to calculate a standard error around each audience rating.

Let us define:

p_1 = the audience rating (as a percentage) for year 1
s_1 = the standard error of the audience rating for year 1
p_2 = the audience rating (as a percentage) for year 2
s_2 = the standard error of the audience rating for year 2

The test statistic

(2.1) $z = (p_1 - p_2) / \sqrt{(s_1^2 + s_2^2)}$

is known to follow the Gaussian (or normal) distribution. The value of the actual z-statistic may be compared against standard tables to ascertain statistical significance. For example, if the z-statistic is larger than 1.96 in absolute value, then the difference is declared to be statistically significant at the 5% level.

The audience rating (p) is a percentage. The textbook formula for the standard error of a percentage from a simple random sample of size n is given by:

(2.2) $s = \sqrt{[p (100 - p) / n]}$

In practice, real-life survey samples are never simple random samples. In the case of the MARS OTC/DTC Pharmaceutical Study, the following factors make the sample not so simple.

First, the readership question is not a simple binary "Yes/no" question. Instead, it is a frequency question that results in a probability of reading. For magazines, the probabilities are 0.00, 0.25, 0.50, 0.75 or 1.00 based upon the number of issues

read out of the last four issues. For weekday newspapers, the probabilities are 0.0, 0.2, 0.4, 0.6, 0.8 or 1.0 based upon the number of issues read out of the last five issues. The average audience is calculated as the arithmetic mean of the probabilities. This is therefore not a pure binomial variable; instead it is an overdispersed binomial variable which has a different standard error depending on how dispersed the the distribution of the probabilities is.

Secondly, certain missing variables may have to be imputed. For example, if a respondent does not state the marital status, then the variable outcome has to be imputed based upon correlated variables such as age and sex. As another example, if a respondent does not state the household income, then it has to be imputed based upon correlated variables such as the education and occupation of the head of household. Generally speaking, imputation will increase standard errors.

Thirdly, the MARS sample is weighted in two ways. The first part is a design weight to account for the fact that the sampling frame is household-based and one (and only one) respondent is selected per household. All households have identical probabilities of being selected. But a person in a single-person household is selected within that household with certainty while a person in a two-person is selected with 0.5 probability within that household. The design weighting accounts for these different selection probabilities due to the different number of adults within households.

The second part of the weighting is the traditional sample balancing (also known as marginal weighting, raking, or iterative proportionate fitting). The intab sample experiences differential response rate from different segments. It is a common phenomenon in survey research for intab samples to be under-represented in young people, men, blacks, Hispanics, and so on. The MARS sample is weighted on age, sex, geography, employment, occupation, education, personal income, household income, race, household size, and so on.

A common belief is that weighting will automatically cause the standard error to become bigger. In practice, the issue is more subtle. Other things being equal, when the weights are positively correlated with readership, the standard error becomes bigger. But when the weights are negatively correlated with readership, the standard error becomes smaller.

These considerations mean that formula (2.1) is unlikely to be applicable. While formula (2.1) may be simple and easy to use, there are consequences if it is misused. If the actual standard error is higher than formula (2.1) implies, then the user of (2.1) will be led to believe that there are many significant differences when they are not. If the actual standard error is lower, then the user will get the wrong impression that nothing much changes over time.

In the past, MARS has used a standard error model of the form:

(2.3) $(\text{DEFT}) \sqrt{[p\,(100 - p) / n]}$

where DEFT is an adjustment factor known as the Design Factor.

For audience ratings, the DEFT for the MARS study is stated to be 2.5. This design factor was based upon an earlier empirical replication study that determined an average adjustment factor. As such, it may overstate or understate the adjustment factor for individual titles because the complex sample design features (such as weighting) can have different impact.

Over the years, MARS (and its users) have observed that certain print titles are more likely to be found to be statistically significant under this standard error model. Sometimes, a print title may be showing an increasing (or decreasing) trend over the years and we may think that we understand this reflects real change. For example, the Internet has been eroding readership of certain titles and this is an ongoing process. In other cases, a print title may bounce up one year, tumble down the next year, bounce up again the year after, and so on. Absent any plausible explanation, we suspect that such a print title may have a larger standard error than indicated and it was frequently declared to be statistically significant when it was not.

The purpose of this paper is to use an extensive and accurate empirical study to determine precise title-specific standard errors for the 87 titles across the seven years of the MARS studies. Hence we can determine the true pattern of changes across the years. As a side benefit, we want to be able to understand how the DEFT's vary across titles and their determinants.

3. Jackknife Replication

The methodology used here is known as jackknife replication. The technical details of the method can be found in Wolter (1985). Large-scale applications to media audience research include Occhiogrosso and Frankel (1982) and Soong (1988). A brief description of the process will be given here.

The MARS study begins with a mailing sample of about 25,000 names selected as a systematic sample from Acxiom Corporation's compiled household database of about 110 million compiled household names and addresses. For this study, we divide this mailing sample into 50 mutually exclusive systematic samples. For example, the first sub-sample might consist of the 1st, 51st, 101st, … cases; the second sub-sample might consist of the 2nd, 52nd, 102nd … cases. These sub-samples are known as 'simple replicates.'

These simple replicates are just 1/50th the size of the full sample. As such, they cannot be treated as replicates of the full sample because their behavior is different under the survey process. For example, while you can reasonably weight a full sample on a long list of variables (e.g. age, sex, geography, education, employment, occupation, personal income, household income, race, etc), a 1/50th sub-sample cannot be put under so much weighting. If this were attempted, there will be non-convergence in the sample balancing as well as extreme values for the weights.

In the jackknife replication method, each jackknife replicate is defined to be the complement of the simple replicate. If a simple replicate is 1/50th the size of the full sample, then the jackknife replicate is 49/50th the size of the full sample. There are 50 jackknife replicates, one for each simple replicate. The intab sample of each jackknife replicate is processed using the same data editing, imputation and weighting. Audience ratings are calculated for all the print titles in the jackknife replicates.

Let

p = the audience rating for the total sample
p_j = the audience rating for the j-th jackknife replicate
k = 50 (which is the number of jackknife replicates)

Then the jackknife standard error of the audience rating (p) is given by the formula:

$$(3.1) \qquad \sqrt{[(k-1)\sum_{j=1}^{k}(p - p_j)^2 / k]}$$

It should be noted that this formula is not the same as the usual one for a standard error because there is an additional factor of (k-1), which is therefore to account for the fact that the jackknife replicates overlap each other.

4. Results of Statistical Testing

When the processing was completed there were 87 x 7 = 609 audience ratings, each with its own individually fitted standard error. Statistical testing was conducted on the audience ratings between adjacent years (that is, year 1 against year 2, then year 2 against year 3 and so on). Out of 87 x 6 = 522 statistical tests, 109 (or 21%) were statistically significant at the 5% level. If there were no real changes either in readership or methodology, we would expect (by definition) that 5% of these statistical tests would be significant at the 5% level. The fact that 21% were statistically significant showed that there were real changes in the audience data.

In Table 4.1, we show the percentages separately by time period.

Table 4.1. Percent of significant statistical tests by time periods of comparison

Time Periods	% statistically significant at 5% level
2001 vs. 2002	29.9%
2002 vs. 2003	17.2%
2003 vs. 2004	21.8%
2004 vs. 2005	21.8%
2005 vs. 2006	19.5%
2006 vs. 2007	10.3%

Why are there so many more statistically significant differences for 2001 vs. 2002? Methodological changes can easily account for some of them. 2001 was the first year of the MARS study. Even though many of the features were tested in pilot studies, we were able to uncover some of the problems only after we completed the study and saw the full results. For example, in 2001, we listed the four newspapers at the top of the first page. We ended with numbers that everybody believed were far too high. In 2002, we moved the four newspapers to the bottom of page six and added write-in answers for other newspaper titles. The audience ratings for the four newspapers dropped an average of 25%, which were obviously statistically significant. From 2002 onwards, the newspapers were treated virtually the same way and there has been only one statistically significant difference since (note: one out of 4 x 5 = 20 comparisons is exactly 5%!).

As for why the percentage of significant results should decrease over time, one explanation is the "MRC effect." MARS is an accredited service by the Media Ratings Council. It took several years to achieve that status, and the process involves specifying and justifying the procedures and adhering to them very closely thereafter. Thus, there is a lot less change in methodology from 2006 to 2007 than from 2001 to 2002. If there is a major change (such as the increase of the questionnaire from 20 pages to 24 pages in 2007), then the expected effect is well understood via pilot tests.

The next issue is about whether there are any print titles that tended to swing significantly from one year to the next. For this analysis, we can construct a baseline based upon a binomial model with 0.05 probability of success and 6 trials. For example, the probability of having zero significant results requires six successes in a row (=0.95^6 or 73.5%). This is shown as the second column of Table 4.2. In the third column, we show the actual data that were observed.

Table 4.2. Empirical and theoretical percentages of significant results for print titles

Number of Significant Results	Theoretical Percentage	Empirical Percentage
0	73.5	33.3
1	23.2	31.0
2	3.1	20.7
3	0.2	12.6
4	0.0	1.2
5	0.0	1.2
6	0.0	0.0

We inspected the titles that had three or more significant changes over the seven year period. We can understand some of them. For example, *US Weekly* had this series of audience ratings over the seven years: 2.4, 3.3, 3.7, 6.7, 7.2, 9.7, 10.6, with three of these being statistically significant. We think that we understand this because this is a rapidly rising title and this hypothesis is supported by the results from many other independent studies. But there are at least half a dozen titles for which we cannot understand the oscillatory nature of their audience ratings over time. There may be some logical explanation, but we do not have all the facts (such as editorial changes and circulation strategies) in our hands.

5. Design Factors

The design factor DEFT is defined as the ratio of the standard error from the jackknife replication divided by the standard error from the formula of the simple random sample (see formula 2.2). There are 87 x 7 = 609 DEFTs from the 87 print titles over 7 years. Here are the summary statistics for those DEFTs.

Maximum = 6.13
90-th percentile = 2.75
75-th percentile = 2.50
Mean = 2.27
Median = 2.23
25-th percentile = 1.99
10-th percentile = 1.76
Minimum = 1.29

Previously, MARS had conducted a one-time-only study in 2003 which yielded an estimate of 2.5 for the design factor. As it turns out, this estimate is conservative in the sense that it overstates the impact of the sample design on the reliability of the audience estimates. Being conservative means using larger standard errors in statistical tests and therefore declaring fewer 'statistically significant' differences over time. Being conservative is usually considered better than being more aggressive, because it is preferable in the business sense. To understate the standard error and detect more 'statistically significant' results could result in turmoil. Senior managers have been dismissed because their media titles performed 'statistically significantly' worse than before so that a completely different direction is followed.

Why should the DEFTs vary across titles? In Section 2, we hinted that this depends on the correlation between the weights and readership. This is easily illustrated by a couple of examples.

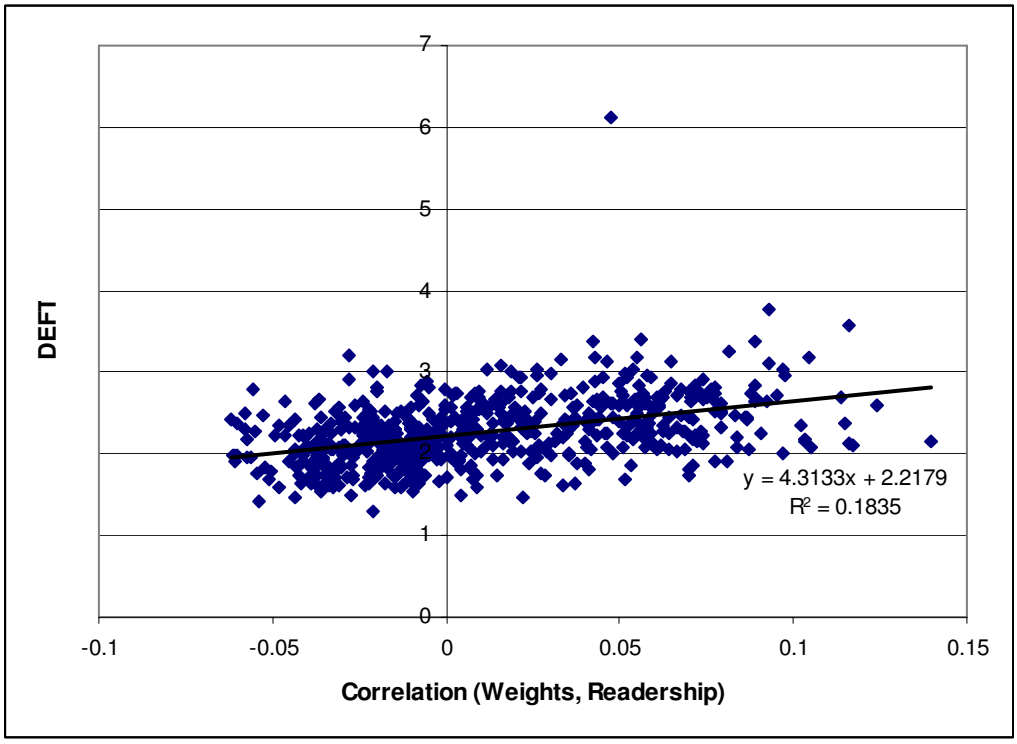
In the first instance, a certain group is underrepresented in the intab sample. An example might be blacks in the United States, who are well-known to be less cooperative to survey research for a number of reasons (such as coverage in the sampling frame, interest in the subject matter, etc). Another example might be young males. Survey samples are usually weighted to the universe estimates for those groups. Meanwhile, there are certain print titles that have extensive appeal to that group (e.g. magazines that appeal to blacks or young males). These print titles will benefit as a result of the weighting.

This means that a relatively small group of persons has a disproportionately large impact on the audience rating of the relevant print titles. Generally, this means greater volatility and therefore larger standard errors. The characteristic of the situation is that the weights are positively correlated with the readership behavior. As a result, one might be reasonably led to think that a positive correlation between the case weights and readership might lead to a larger DEFT.

Conversely, it is possible that for some print titles, the weights may be negatively correlated with readership. For example, some groups in the population are over-represented (for example, elderly women) in the MARS study. Such groups have to be weighted down. For those print titles which appeal to these groups, the correlation between the weights and readership will be negative. As a result, one might be reasonably led to think that a negative correlation between the case weights and readership might lead to small smaller DEFT.

In our case, we calculated the correlation coefficients between the weights and readership for each of the 87 print titles over the 7 years. Figure 5.1 shows the scatterplot between the DEFT and the correlation coefficient.

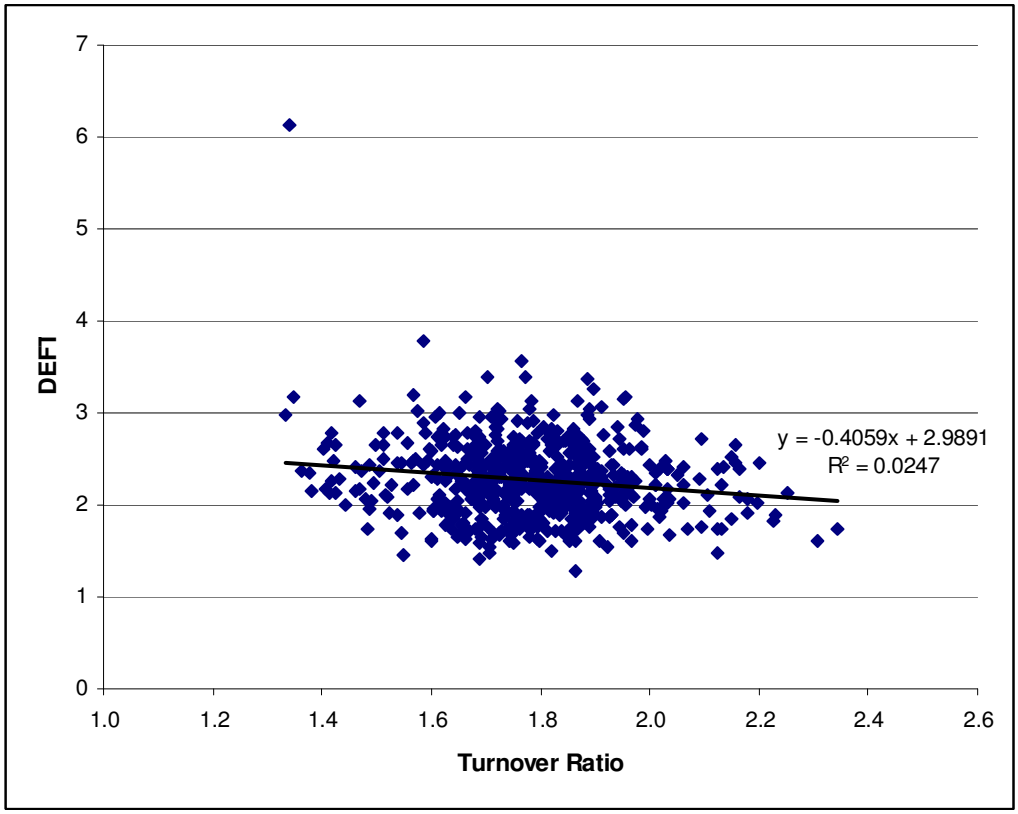
Figure 5.1. DEFT versus Correlation Coefficient (Weight, Readership) across all print titles 2001-2007



As hypothesized, there is a positive relation between DEFT and the correlations. The regression of DEFT on the correlations was statistically significant at the 1% level with 18.35% of the variance being explained.

We also stated in section 2 that the frequency of reading question was an overdispersed binomial variable, and that may affect the DEFT. The impact depends on the degree of dispersion. When most of the readers read every issue of a print title the results would be close to a standard binomial variable. When the answers are distributed more evenly among the various frequency classes, the overdispersion would be larger. This can be captured in the turnover ratio, which is the ratio between the cumulative audience (which is the number of people who read any of the issues) divided by the average audience. We calculated the turnover ratio for each of the 87 print titles over the 7 years. Figure 5.2 shows the scatterplot between the DEFT and the turnover ratio.

Figure 5.2. DEFT versus Turnover Ratio across all print titles, 2001-2007



There is a small negative correlation between the DEFT and the turnover ratio, which does not achieve statistical significance at only 2.5% of the variance accounted for.

6. Standard Error Models

In this section we will deal with the consequences of various standard error models for statistical testing. The formula for comparison two independent audience estimates (such as the audience ratings for a print title between two years) as given in formula (2.1). The audience ratings are known, but what should we assume about the standard errors in that formula?

The simplest model is to assume that formula (2.2) for simple random samples is valid. We simulated this model on the MARS data. Out of the 87 x 6 = 529 between-year pairwise comparisons, we found that 52% of the tests were statistically significant at the 5% level. Thus, on the average, half of the print titles will have statistically significantly different audience ratings compared to the previous year.

The next simplest model is the so-called first-order Taylor series approximation. Here, we recognize that weighting destabilizes audience estimates. The formula for DEFT is given by:

(6.1) $\quad DEFT = \sqrt{[\ 1 + RV(w)]}$

where RV(w) is the relative variance of the weights (i.e. the variance of the weights divided by the square of the mean weight). If all weights are equal, then DEFT is 1.0 as expected. If there is a lot of variation in the weights, the DEFT will be big.

In the MARS studies, the average DEFT for the first-order Taylor series formula is 1.55. When applied to the 87 x 6 = 529 between-year pairwise comparisons, we found that 33% of the tests were statistically significant at the 5% level. Thus, on the average, one-third of the print titles have statistically significantly different audience ratings compared to the previous year.

A more complicated model is the so-called second-order Taylor series approximation. The formula is based upon considering that the variance of a weighted average is impacted by the relative variances as well as the covariance between the weights and readership (see Goodman (1960)). This relationship is demonstrated in Section 4 above.

In the MARS studies, the average DEFT for the second order Taylor series formula is 1.86. When applied to the 87 x 6 = 529 between-year pairwise comparisons, we found that 37% of the tests were statistically significant at the 5% level. Thus, on the average, more than one-third of the print titles have statistically significantly different audience ratings to the previous year.

In Section 4 we had used the jackknife replication method to calculate standard errors and then found that 21% of the between-year comparisons were statistically significant at the 5% level. This is considered an accurate method. All three models above are based upon simplifying assumptions and understate the standard errors. As a result, we found many more 'statistically significant' results.

7. Conclusions

In this study we made an elaborate effort to calculate standard errors properly in order to identify the real changes in the MARS survey over the past seven years. For the 87 print titles, the percentage of changes that were statistically significant was 21% at the 5% level. In some cases, it is clear that the changes occur as a result of changes in survey methodology. Only 2 print titles had statistically significant year-to-year changes four or more times out of the six year-to-year comparisons.

By making the effort to calculate the standard errors accurately, we were able to assess the use of standard error models. First of all, if we used a naïve Simple Random Sample formula, then this will be a serious underestimation as the correct standard error is 2.25 times larger on the average. This led to 52% of the tests being 'statistically significant.' Similar problems exist with a couple of other standard error models.

We also studied the use of other variables to custom-fit DEFTs. While the correlation between the weights and readership is a statistically significant predictor, it is hard to calculate in practice. Therefore, it is difficult to implement it. The turnover ratio is easy to calculate, but it turned out not to be statistically significant.

We conclude with the reminder that since data-driven decision-making by statistical tests are based upon standard errors, the use of incorrect standard errors leads to serious mistakes of judgment. Caution should be exercised about the results from statistical testing if there is doubt about the standard error model.

REFERENCES

Goodman, L.A. (1960) On the exact variance of products. *Journal of the American Statistical Association*, 55, 708-715.

Occhiogrosso, M.G. and M.R. Frankel (1982) *Arbitron Replication II: A Study of the Reliability of Radio Ratings*. Arbitron Ratings Company. New York, NY.

Soong, R. (1988) The Statistical Reliability of People Meter Ratings. *Journal of Advertising Research*. February/March, 1988.

Wolter, K. (1985) *Introduction to Variance Estimation*. Springer-Verlag New York, Inc. New York, NY.