IMPROVING STATISTICAL RELIABILITY OF MAGAZINE AUDIENCE ESTIMATES

Martin Frankel, Jim Collins, Julian Baim and Joe Agresti, GfK MRI

INTRODUCTION

One of the most vocal client complaints related to magazine audience ratings is the variability of overall ratings estimates and audience demographic profiles from one report to the next. These issues are not unique to any single country; it is one of the reasons why publishers and agencies continue to ask for even larger total samples to improve reliability in their respective countries' ratings service. Despite these well-intentioned requests, three fundamental factors suggest that larger samples, while certainly laudable, are not a viable answer. First, assuming the constancy of sample design, sampling variability (i.e., sampling error) is only halved when the sample is quadrupled. The likelihood of quadrupling current sample sizes using probability frames and the same sample design is minimal insofar as the cost of doing such is prohibitive. Second, an overwhelming number of measured magazines has audience ratings of 2% or less. Relative sampling error will in all likelihood remain quite high despite sample increases. Finally, most industry clients examine an individual magazine's demographic profile, which means the underlying basis for calculating profile sampling error is the number of claimed unweighted readers of a publication's average-issue audience. In most cases, these unweighted bases are fractions of the total sample and hence subject to volatile swings.

In the quest for larger samples at lower cost a number of organizations either completely or partially abandoned door-to-door interviewing based on full probability samples. Some organizations have adopted hybrid methods that use telephone, mail or online data collection/sampling based on probability sampling. Other organizations have abandoned probability methods altogether.

Although the difficulties and costs associated with probability sampling methods are well known and often highly touted, media measurement generally exists in a highly competitive and often zero sum economic environment. As a result, the prospect of "bias" related to certain sample design or data collection methods, presents a barrier to the adoption of other sampling methods.

Within the context of both probability and non-probability sampling, there are a number of methods that have been used by survey organizations to produce "stabilized" estimates of certain behaviors, attributes and measures. This paper briefly describes a number of different smoothing or stabilization approaches that have been used in various survey and data reporting contexts. It then focuses on a class of methods that are viewed as potentially suitable for the current MRI-GFK product "The Survey of the American Consumer."

STATISTICAL ADJUSTMENT IS A PART OF MOST SAMPLE SURVEYS.

Virtually all sample surveys use some degree of statistical adjustment. Until recently, most of the statistical theory and literature involving the topics of "statistical inference", "estimation" and "testing" did not mention non-response and other "bias" producing non-random errors at either the element (individual) or item (question) level. In practice however, most statistical summarization, estimation and testing use some degree of adjustment. By adjustment we mean that there are differences in simple summarization (either counting or averaging) of the actually collected data and the reported results.

In the US, there is evidence that beginning with the first constitutionally mandated Census of 1790, some degree of "adjustment" was discussed and undertaken.[iii] Adjustments of the US Census were and continue to be controversial, both statistically and politically. Over the years, various adjustments to the simple count of the US Census have resulted in the increased representation in the US House of Representatives at the expense (loss of seats) of other states. Since the Census is used in various revenue allocation formulas, much of the statistical adjustment undertaken by the US Bureau of the Census and the Bureau of Labor Statistics exist in a zero-sum revenue distribution environment, there are always a number of lawsuits and some introductions of congressional bills that accompany the release of the US decennial census.

All data adjustments (whether or not they are labeled as such) involve the use, either implicit or explicit, of an underlying statistical model. The justification of these models is typically based on the argument(s) of random error reduction/stabilization/removal or bias attenuation/reduction/elimination.

In most research studies based on probability samples (including media research), statistical adjustment is typically undertaken on the basis of the probability of selection of sample units and differential non-response among various groups of sampling units. This is generally described as weighting and will often involve the adjustment of respondent level weights so that the resulting "weighted results" conform to external and generally accepted "population characteristics." In the US, these characteristics include the distribution of persons on be basis of Gender, Age, Race/Ethnicity, Geography, Household Income, Educational Attainment and Language Spoken.

The statistical adjustments described above are generally well accepted by most of the scientific community and the informed public. There is less knowledge and possibly acceptance other forms of statistical adjustment designed to smooth or stabilize survey estimates.

STATISTICAL ADJUSTMENT BASED ON "STRONG MODEL" ASSUMPTIONS

Some surveys, make use of a class of adjustments that might be characterized as involving "strong model" assumptions. Perhaps the best known of these in the United States is the Monthly Report of Unemployment issued by the US Bureau of Labor Statistics and US Bureau of the Census. Since the early 1940s the US Bureau of the Census has carried out monthly data collection based on a probability sample of US households on behalf of the US Department of Labor. The monthly data is then weighted using the probability of selection and the type of non-response/conforming methods previously described. These weighted data are then used to compute various well known labor force statistics such as: In the labor force; Currently working; Looking for work and the Unemployment Rate. However, the labor force statistics developed from the weighted survey data are actually further adjusted before they are generally released. These further adjustments involve "smoothing" and adjustment for seasonality and other random variations.

Thus, if one were to compute the various labor force estimates using the publicly available data files, the results of this direct tabulation would not agree with the publicly released labor force statistics.

The use of seasonal and stabilizing adjustment can be traced to the 1950s and seems to be associated with the introduction of non-military use digital computers.[iii] At the present time US monthly labor statistics are modified by a method known as "X-13 Seasonal Adjustment,"

ECONOMIC TIME SERIES ADJUSTMENTS CANNOT BE SIMPLY REPRODUCED BY SIMPLE SURVEY TABULATION

One of the key features of many "adjusted" statistical estimates that are released for both public consumption and for paid subscription use is that they are not directly reproducible from associated databases. That is, if one were to obtain the required data base and carry out standard weighted tabulations, the final results will differ from the final released estimates. This simply means that one or more "statistical adjustment steps" must be carried out to produce the final estimate. This differences between the estimates produced by direct tabulation and those that are the output of adjustments represents an important constraint for survey organizations (such as GFKMRI) which also release their databases to clients.

For GFK-MRI this means that final published estimate must be reproducible by direct tabulation. This constraint holds true for overall statistical estimates (e.g. the Average Issue Audiences of Magazines), but corresponding estimates for various domains and sub-groups. For example, users of GFKMRI should be able to produce audience estimates for various groups of the population (e.g. Males 18-49 with children in their households). Users should also be able to produce estimates of Median Household Income among readers of a specific magazine for households in the South region of the United States.

Not all media research companies produce and distribute a weighted database as part of their research product. However, because GFKMRI not only produces "published" estimates as well as a database which is available to clients and third-party processors on behalf of clients, a number of critical constraints were imposed on the development of smoothing (variation reduction) adjustments. Specifically any system must satisfy the following requirements.

A. ESTIMATES MUST BE REPRODUCABLE BY DIRECT TABULATION
B. THE SYSTEM MUST BE CAPABLE OF PRODUCING CONSISTENT ESTIMATES FOR SUB-DOMAINS AND SUB-CLASSES
C. MUST BE UNDERSTANDABLE AND CONSISTENT ACROSS A WIDE RANGE OF ESTIMATES OF BOTH MEDIA AND PRODUCT UTILIZATION

SOME DEGREE OF SMOOTHING (VARIANCE REDUCTION) IS ALREADY PART OF THE GFK-MRI SYSTEM

The basic GFKMRI survey (The Survey of the American Consumer) is fielded and released using a 6-7 month survey period. This is generally referred to as a Survey Wave. Specifically, a Full Probability Area Based Sample of Households is selected and fielded (i.e. data is collected) during a 6-7 month time period. Each six months, a national probability area sample of households distributed among 168 primary sampling units and 1250 secondary sampling units is fielded. This process results in approximately 12,250 cooperating households. The data for these households is edited, weighted and projected to represent the full population of approximately 220 million adults living in 110 million US households. While this weighted sample is available to clients, the standard GFKMRI sample release consists of two successive 6-7 month waves of data which produce average magazine audiences and other media audiences that span a single year.

MRI'S CURRENT DOUBLE-BASE PRODUCT

In an effort to provide increased reliability for various magazine and media profiles, MRI has provided a product called DoubleBase. DoubleBase is released every year and consists of the most recent four waves of the Survey of the American Consumer. At the time of release, the survey midpoint is approximately one year prior to the release. Thus, at the time of release the Doublebase projects to a point in time approximately one year prior to its first release. As time progresses the survey midpoint grows from one year to two years. While the Doublebase product is widely used, it is often criticized as being a bit out of date.

INCREASING THE EFFECTIVE SAMPLE SIZE WHILE MAINTAINING TIMELINESS

The proposal put forth in this paper for increasing the stability of audience estimates and associated audience level while maintaining time currency is the result of two basic research findings.

The first finding was reported in 1999 by Stuart Grey and co-authors with a follow-up in 2001. Grey found that the prediction of the "next" wave of audience levels could be improved by applying a higher relative weight to the more recent wave of survey results.

The second finding, reported for the first time in this paper, shows that for the vast majority of magazines, the wave to wave variation is consistent with the stationarity hypothesis. This means that while it is well understood and accepted that magazine audience levels are not constant over time, the random variation in audience estimation generally exceeds the true wave to wave audience size variation. This means that the elimination of random variation by some form of averaging is justified. This justification arises because in the short term (over the period of 3 waves of data collection) the random error eliminated substantially outweighs the error attributable to actual audience change.

A SOLUTION THAT IS SIMPLE AND TRANSPARENT

In the process of developing a proposal for increasing the stability of audience estimates and associated within audience demographics we considered a number of options, some of which were quite complex. For example, we made use of the iterative linear programming function bundled with the Microsoft Excel product to find the optimal weighting the three most recent wave in the prediction of the most current future audience on a magazine by magazine basis. The overall optimization function involved the minimization of both mean absolute error and mean absolute relative error.

In the end we chose a solution that was not only simple, understandable and transparent, but possessed a certain intuitive logic. Specifically our strategy for estimate stabilization is based on the following formula.

ESTIMATED AUDIENCE AND DEMOGRAPHICS = .4 WAVE(t), + .4 WAVE(t-1)+.2WAVE(t-2)

This formula allows GFKMRI to release a data file that is easily integrated into all currently used tabulation systems since simply involves to use of a modified weight for the most recently sampled three waves of respondents. (approximately n=37,000). As we show below, this results in decrease of random error of approximately 13% for average issue audience (AIR) estimates and 21% for title specific median household income levels.

EXAMINING THE TRADEOFF BETWEEN RANDOM ERROR REMOVAL AND BIAS INTRODUCTION

Our first analysis examined the proportion of times and number of titles for which wave to wave variation was within one and two sigma limits. On a title by title basis we computed 10 Z-scores by subtracting the wave specific audience estimate from the 10 wave average. This difference was then divided by the audience standard error. The audience standard error for each title, was based on the average of 10 jack-knife standard errors.

On the basis of standard normal theory, coupled with the assumption of stationarity, we expect that 66.7% percent of all z scores should fall within plus and minus one. Further, we expect that 95 percent of all z-scores should fall within plus and minus two. Table I shows the distribution of the number of titles (out of 163) that the z scores exceeded one (i.e. one sigma).

TABLE I: AUDIENCE ESTIMATES (Z-SCORES) OUTSIDE ONE SIGMA

	NUMBER OF TIMES Z SCORE EXCEEDS ONE (OUT OF 10)		NUMBER OF TITLES
	0		5
	1		14
	2		23
	3		31
	4		37
	5		26
	6		14
	7		7
	8		4
	9		2
AVERAGE	3.787	TOTAL	163

Table II shows the distribution of the number of titles (out of 163) that the z scores exceeded two (i.e. two sigma).

TABLE II: AUDIENCE ESTIMATES OUTSIDE TWO SIGMA

	NUMBER OF WAVES OUTSIDE 2 SIGMA		NUMBER OF TITLES
	0		96
	1		34
	2		18
	3		5
	4		6
	5		1
	6		2
	7		1
	8		0
	9		0
AVERAGE	0.816	TOTAL	163

From Table I we find the average number of times (relative to 10) that the z-score exceeded one is 3.78 This average is obtained by multiplying the two columns, adding the total and dividing by 163 (614/163). In similar fashion, from Table II we find that the average number of times that the z-score exceeds 2 is 0.816.

Under the assumption of full stationarity for all magazines the expected average number (out of 10) of z scores that exceed one is 3.34. The expected number of z scores that exceed two is 0.05. Comparison of 3.787 with 3.34 and 0.816 with 0.050 provides strong evidence that while there is some departure from full stationarity for some titles, overall, the elimination of random error is certainly justified.

HOW MUCH VARIATION IS REDUCED BY THE PROPOSED METHOD OF WEIGHTING?

In order to examine the degree to which our proposed weighting solution actually reduces variation we examined the impact of our procedure over 10 waves of data collection spanning a five year field period.

We computed average issue audience as well as within title median household income using successive sample waves assigning a (50%, 50%) weighting, as is the current practice. We also computed estimates of average issue audience and within title median household income using three most recent waves and applying the proposed (40%, 40%, 20%) weighting.

In order to equalize the comparison of methods we computed 7 successive audience estimates using the current (50-50) method and 7 successive estimates using the proposed (40-40-20) method. We then computed the relative standard deviations of these 7 successive audience estimates (AIR) on a title by title basis. We applied the same method of computing two sets of estimates for title specific median household income. Finally we computed the title specific relative standard deviations of these median household income estimates. The results of this analysis is shown in Table III.

TABLE III: PERIOD TO PERIOD VARIATION AVERAGE RELATIVE STANDARD DEVIATIONS

ESTIMATE	CURRENT	PROPOSED	RATIO TO CURRENT
AIR AUDIENCE	0.0777083	0.0676415	0.870454
TITLE SPECIFIC MEDIAN HOUSEHOLD INCOME	0.055954	0.044068	0.787574

Over the full set of magazine titles the average relative standard deviation for AIR estimates computed using the current method was 0.0777083 and was 0.0676415 for the proposed method. The average standard deviation across titles for median household income was 0.055954 using the current method and 0.044068 using the proposed method. Thus for overall audience estimates the proposed method produced average standard deviation that were 13% lower than the current method (0.0676415/0.0777083 = 87.04%). For estimates of title specific median household income the proposed method produces estimates with average standard deviation 21% lower than the current method (0.044068/0.055954 = 78.76%).

CONCLUSIONS

On the basis of our research to date we feel that the use of the proposed three wave weighting provides a viable and easily adoptable method for reducing wave to wave bounce in overall audience estimates as well as key title specific title profiles. More research will be conducted to further examine the behavior of this method and a changing magazine audience environment.

[i] Constitution of the United States of America, Article I, Section 2, Clause 3: Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers…The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.

[ii] Cantwell, P.J; Hogan, H.; and Styles, K.M. "Imputation, Apportionment, and Statistical Methods in the U.S. Census: Issues Surrounding Utah v. Evans, ; Statistical Research Division, U.S. Bureau of the Census, Washington, DC, (2005).

[iii]Shiskin, J., and Eisenpress, H., (1957), "Seasonal Adjustment by Electronic Computer Methods," JASA, Vol. 52, No. 280.
Marris, S.N., (1960), "The Treatment of Moving Seasonality in Census Method II," *Seasonal Adjustments on Electronic Computers*, Organization for Economic Cooperation and Development.