The False Discovery Rate in ACS: Helping Users to Understand Estimates for Small Domains

Robert E. Fay[1]
[1]7252 Greentree Rd., Bethesda, MD 20817

Abstract

In 2010, the U.S. Census Bureau will publish the first set of 5-year period estimates from the American Community Survey (ACS), based on data for 2005-2009. Published for small places, tracts, and other small areas, the 5-year ACS estimates will attempt to replace the long-form data from recent decennial censuses. Because the effective sample size for the ACS will be somewhat less than half that of previous censuses, users will face an increased challenge to distinguish true variation from sampling error.

The concept of the false discovery rate has become increasingly useful in other disciplines confronted by large numbers of estimates, such as micro-array analysis in genetics and fMRI studies of the brain. The paper will review this concept and suggest its possible future application, based on a preliminary analysis of published data from the ACS Multiyear Estimates Study.

Key Words: American Community Survey, multiple testing, small area estimation, FDR

1. Introduction

The Census Bureau has undertaken the American Community Survey (ACS) as a replacement for the decennial census long form. The survey continuously collects data from the U.S. population in monthly samples, totalling approximately designated 3,000,000 housing units per year. As a consequence of subsampling and nonresponse, it completes interviews with approximately 2,000,000 households per year. In 2006, the Census Bureau published data from the first full year of data collection in 2005, beginning annual publication of 1-year estimates.

In spite of the large size of the ACS's annual sample, it is still small relative to the long-form sample of previous censuses. Instead, the strategy will be to accumulate ACS data over periods of 3 and 5 years to form period estimates. ACS publishes 1-year period estimates for states and areas with population of 65,000 or more. Late in 2008, the Census Bureau will publish its first set of 3-year period estimates for areas of 20,000 or more. The ACS will only attempt the geographic detail of the previous decennial census long form in 2010, when 5-year period estimates for 2005-2009 will be published.

As recently emphasized by a National Academy of Sciences (NAS) panel (NRC, 2007), however, the effective sample size supporting the ACS 5-year estimates will still be somewhat less than half the size of the Census 2000 long form. The title of the panel's report, *Using the American Community Survey: Benefits and Challenges,* suggests the scope of its findings. The panel concluded that the ACS held substantial promise as a replacement for the long form but also faced significant challenges. Among the challenges were to guide users (1) in interpreting the complex array of 1-, 3-, and 5-year estimates to be eventually produced; and (2) in appreciating the effect of sampling error on the variability in the estimates. The report offered 10 guidelines to users in interpreting ACS data, the first of which was to "Always examine margins of error before drawing conclusions from a set of estimates." To follow this guideline, users must be able to access or construct margins of errors for the range of analyses that they undertake.

As the NAS panel was completing its work, the Census Bureau was producing 1-year, 3-year, and 5-year estimates from the Multiyear Estimates Study (MES), a study of a test version of the ACS in 34 test counties. The results have now been published. The MES illustrates and tests the panel's general guidelines. Previous work (Fay 2007a, 2007b) illustrated some of the challenges that would be encountered in analyzing ACS data for relatively large subcounty

areas, such as areas with population around 100,000. Even at that relatively high level of aggregation, separating real change from the effect of sampling variability was found challenging.

This paper extends the analysis to smaller areas, specifically census tracts, which are areas with an average population of about 4,000 people. Publication of 5-year estimates at the tract level results in overwhelmingly large number of estimates. Even if statistical tests are applied at conventional levels of significance in an attempt to filter out real differences from random variation, many of the nominally significant results may be so at random. The concluding remarks of the previous study (Fay 2007a, 2007b) suggested applying the notion of the *false discovery rate* (FDR) (Benjamini and Hochberg 1995) and the associated statistical procedures designed to control it. This paper illustrates the possible effect of the approach on the MES data. By filtering out nominally statistically significant results that are likely to stem simply from random variation, users will improve their chances of discovering interpretable and understandable results.

1.1 The Multiyear Estimates Study

Although the ACS started field work at essentially the full level in 2005, more than a decade of research work included conducting ACS-size samples in test counties. Beginning in 1999, the ACS interviewed essentially full-size samples in a set of 34 test counties (out of more than 3,000 counties in the country). The counties spanned a range of environments, including Bronx County, New York and San Francisco, California, as well as small rural counties.

The Multiyear Estimates Study (MES) used the data from the period 1999-2005 for the 34 counties to assemble the possible 1-, 3-, and 5-year estimates. As has been practice for many ACS products, individual estimates were accompanied by estimated margins of error. Besides the operational experience, a primary goal of the study was to anticipate the eventual appearance of the ACS to most users. For example, the MES produced five-year estimates for 1999-2003, 2000-2004, and 2001-2005 for the 34 counties. The availability of three partially overlapping sets of five year estimates anticipates the future status of the ACS in late 2012, when 5-year estimates for 2005-2009, 2006-2010, and 2007-2011 will be available. At http://www.census.gov/acs/www/AdvMeth/Multi_Year_Estimates/overview.html, the Census Bureau's web site provides further information about the MES.

1.2 The Lake County Challenge

Lake County, Illinois is among the 34 test counties included in the MES. Beaghen and Weidman (2007) selected the trend in the ACS estimates of the percent speaking Spanish at home to illustrate a point in their suggested guidelines to ACS data users. As Figure 1 illustrates, there is a clear upward trend.

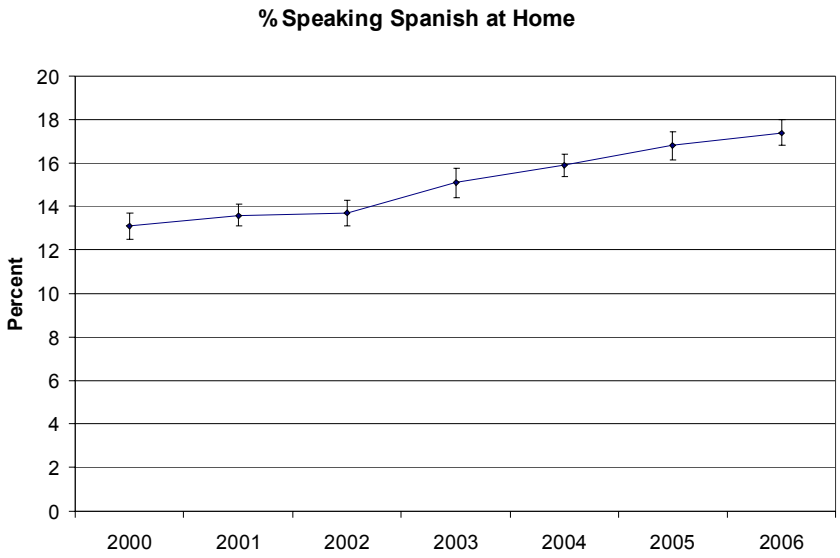

Figure 1: Percent speaking Spanish at Home in Lake County, Illinois, ACS estimates and 95% confidence intervals. All estimates are from the Census Bureau's Multiyear Estimates Study, except the 2006 estimate which is included as part of the 2006 ACS release.

The statistically significant trend in Figure 1 invites the question of the extent to which the ACS data permit disaggregation within the county. This notion is further elaborated in the following:

The Lake County Challenge

The government of Lake County, Illinois, is quite aware of the rapid population growth and changing demographics of the county. The county website, www.co.il.us, offers a summary of basic changes by comparing the 1990 and 2000 censuses, at www.co.lake.il.us/about/demographics/default.asp. A table shows a 25% change in total population from 1990 to 2000, from 516,418 to 644,599, and a 140% increase in the Hispanic population (any race) from 38,570 to 92,716. The page also points to the Census Bureau website for additional census data.

At the 2007 Joint Statistical Meetings, Michael Beaghen and Lynn Weidman presented data from the ACS Multi-Year Estimates Study (MYES) for the percent speaking Spanish at home for Lake County. Lake was among the counties for which 1-, 3-, and 5-year period estimates are available. Over the period from 2000 to 2005, the 1-year estimates for Lake County show a clear increase in the proportion speaking Spanish at home, which is mirrored in a smoother upward trend in both the 3- and 5-year estimates. The statistical significance of the overall trend isn't even in question.

Hypothetical Situation

Suppose a senior county official in Lake County becomes interested in this trend of Hispanic speaking households and interprets this measure as an indicator of where bilingual services may be most needed, now and potentially in the future. Suppose also that Lake County is fortunate to have your part-time consultation as a statistician, demographer, or sociologist. You have all of your current (real) skills, but you are only able to access ACS data as an outsider. The official asks you to find geographic patterns in the ACS for the increasing number of persons speaking Spanish at home as best as you can. Accordingly, you will read the general advice provided on the ACS website but, because you are serving the county's goals, you will use all of your professional skills, too. You take up the question:

> *To what degree can you pinpoint areas where Hispanic-speaking households are growing, and are there possible emerging trends?*

The Challenge

> *How would you use the ACS data and what would be your findings?*

The challenge led to an analysis (Fay, 2007a, 2007b) of subcounty change of this one variable. Following a recommendation of the NAS panel, one analysis was based on the 5 Public Use Microdata Areas (PUMAs) in the county. Estimates for the 18 townships in the county were first examined separately, but then analyzed further by grouping them into four groups based on past intercensal growth during 1990-2000 in the percent speaking Spanish at home.

One primary finding from this case study was that analysis of ACS data, formally taking into account margins of error, is time-consuming when the margins of error must be calculated for change and for groupings of areas that are not directly published. The finding amplified the NAS panel's warning that the Census Bureau needed to prepare for the challenges that users will face. Also consistent with the panel's analysis, the ACS is generally only able to provide meaningful measures of trends at relatively high levels of aggregation, such as areas with population of around 100,000. An exception was made to one panel recommendation, however, that users should not attempt to analyze multi-year estimates from overlapping periods, such as the periods 1999-2003 and 2001-2005. In fact, a possibly important exception was discovered under two conditions: (1) that the 5-year data is aggregated into large population units, such as areas of 100,000 or more; and (2) to ease interpretation, change over time is translated into measures of annualized change. Because many users form their own geographic units for analysis by using lower levels of geography as building blocks, this finding may have important consequences for users.

One of the requirements for the ACS to successfully replace the decennial census long form is that the ACS must provide credible statistics for small areas of geography, such as census tracts. In must allow relatively reliable

comparisons among areas for the same time period, as does the census, in spite of its reduced sample size compared to the census. The ACS improves upon the census if it can provide sufficient reliability to monitor change at the local level. The previous study (Fay 2007a, 2007b) noted the importance of tracts but deferred the problem, suggesting the possible application of the false discovery rate (FDR) and procedures that control it, specifically citing Benjamini and Hochberg (1995). This is the direction taken by the current paper.

2. Methods

2.1 Increasing Impact of the False Discovery Rate on Science

Although there is an expanding literature related to the false discovery rate, a few key papers illustrate the growing impact of the concept on science generally. The Benjamini and Hochberg (1995) paper presented the basic concept of the FDR, the motivation for controlling the FDR in some areas of application, and an FDR-controlling procedure for statistically independent tests. The Benjamini and Yekutieli (2001) paper significantly broadened the areas of potential application by showing that the Benjamini and Hochberg (1995) procedure could also control the FDR under a broader class of testing situations in which specific forms of correlation were present, and also offered an alternative, although more conservative, procedure for dependent tests generally. It appears that the expository paper of Genovese, Lazar, and Nichols (2002) substantially contributed to the uptake of these ideas in disciplines using neuroimaging, such as fMRI, in the scientific investigation of the brain. Figure 2 traces the approximate number of citations of these three papers in two journals using neuroimaging, and in the scientific literature generally, as reported at scholar.google.com in July, 2008.

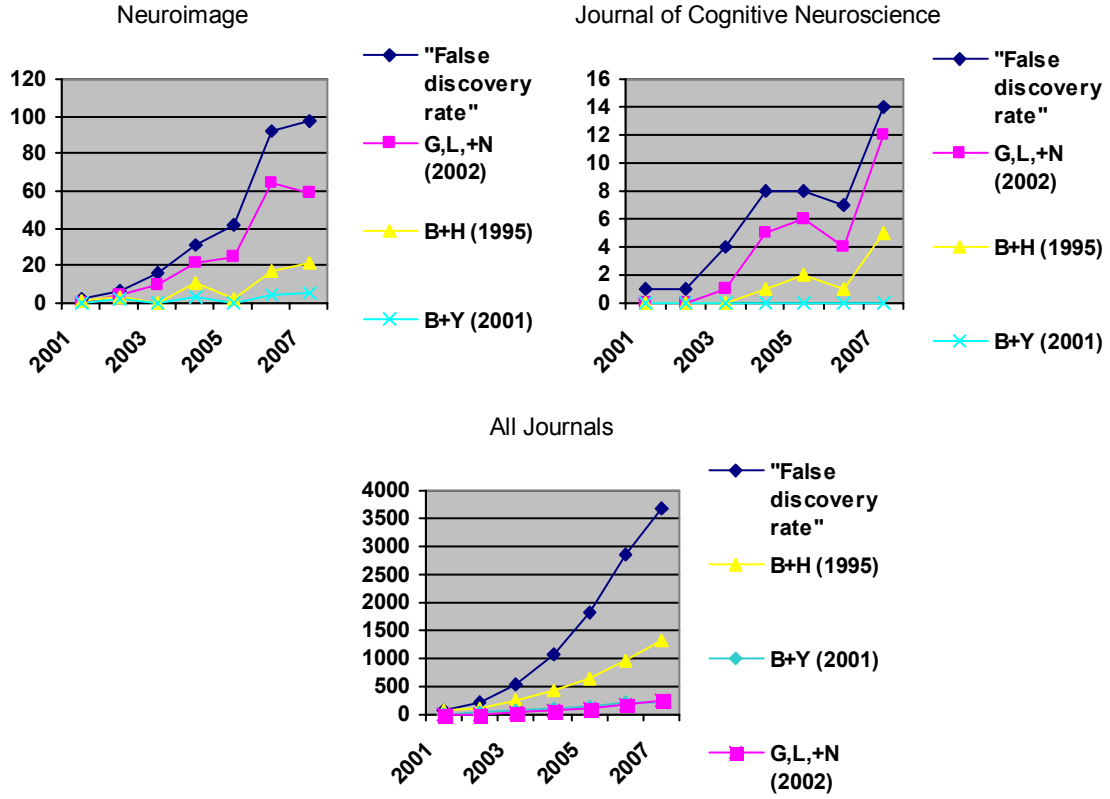

Figure 2: Numbers of papers referencing Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), or Genovese, Lazar, and Nichols (2002), and the count of papers using the specific phrase "false discovery rate" by year. All counts were obtained from scholar.google.com in July, 2008, and the counts for "all journals" refers to all sources scanned by this source.

Although in general arguing for a new approach because "everyone is doing it" is suspect, these data suggest that there has been a rapid uptake by segments of the scientific community. In turn, this observation suggests that some of the more technically skilled users of ACS data, such as those with quantitatively oriented social science backgrounds, may readily learn the approach and apply it appropriately if they are offered appropriate examples and summaries.

The number of citations to Genovese, Lazar, and Nichols (2002) suggests that this expository paper has encouraged the use of FDR-controlling procedures in neuroimaging applications. The paper would also serve as an introduction to the topic for statisticians who have had limited exposure to the approach and can serve as a first paper to read on the subject.

2.2 Controlling the False Discovery Rate

Although the paper by Genovese, Lazar, and Nichols (2002) may be the most accessible of the three, the path-breaking paper by Benjamini and Hochberg (1995) is also relative accessible. Statisticians wishing to guide ACS data users in application of the technique would do well to read both papers. Much of this subsection summarizes key components of the argument that Benjamini and Hochberg (1995) presented in more detail.

Their paper not only contributed to the statistical literature on multiple testing, but it redefined the goals for what a multiple-testing procedure should do in many contexts. In frequentist theory, practically all familiar statistical tests are defined in terms of their probability of falsely rejecting a true null hypothesis. At a fixed α-level, such as .05, a conventional test will falsely reject an expected proportion of true null hypotheses. When testing a moderately large number, such as 100, of true null hypotheses at .05, falsely rejecting at least one of them becomes highly likely. Multiple testing procedures, such as the Bonferroni procedure, instead strongly control the *family-wise error rate* (FWER) to some probability, such as .05. This probability is the upper bound on the change of rejecting *any* true hypothesis, even if some of the hypotheses are false. Suppose there are m hypotheses, H_1, H_2,…H_m, and conventional statistical tests yield for each of them corresponding p-values P_1, P_2,…P_m. Among the hypotheses m_0 are true. Regardless of the unknown value of m_0, the Bonferroni procedure controls the FWER to a desired significance level, q^*, with the rule

$$\text{Reject } H_i, \text{ if } \quad P_i \leqslant (1/m)\, q^* \qquad (2.1)$$

Depending on m_0, the Bonferroni test is conservative, assuring FWER $\leqslant q^*$.

As their paper points out, however, strong control of the FWER often seems too conservative for applications. For example, they point out that in clinical trials there are often multiple endpoints or outcomes being tested for a single drug. Provided that the drug appears to have solely beneficial effects, it may not be necessary to assure strong control of the FWER to identify the specific end-points where the trial points to benefits.

They use the term *discovery* to refer to rejection of null hypotheses generally and *false discovery* to refer to false rejection of true null hypotheses. They suggested that in many applications where a large number of hypotheses are being tested, it seemed more appropriate to control only the *false discovery rate*, which they informally defined as "the expected proportion of errors among the rejected hypotheses." More formally, for V = the number of rejected true hypotheses, S = the number of rejected false hypotheses, they defined Q = V/(V+S) except where, by definition, Q = 0, for V+S = 0. They defined FDR = Q_e = E(Q).

Their paper offered a specific, Bonferroni-like procedure to control the FDR. Under the same setup as the Bonferroni test, they proposed ordering the hypotheses $H_{(1)}$, $H_{(2)}$,…$H_{(m)}$ according to the order of the corresponding p-values, $P_{(1)} \leqslant P_{(2)} \leqslant \ldots \leqslant P_{(m)}$. They proposed rejecting $H_{(i)}$ sequentially, beginning with $H_{(1)}$, according to the rule

$$\text{Reject } H_{(i)}, \text{ if } \quad P_{(i)} \leqslant (i/m)\, q^* \qquad (2.2)$$

If the tests are statistically independent (a condition not required by the Bonferroni procedure), they showed that their procedure controls FDR $\leqslant (m_0/m)\, q^* \leqslant q^*$. Note that the rejection of $H_{(1)}$ is identical to the Bonferroni procedure, but (2.2) makes it easier to reject subsequent hypotheses $H_{(i)}$ with $i > 1$.

Their paper used the independence of the tests in the proof of their results, but many applications require dependent tests. This problem may have indeed been the case for their paper's example of multiple endpoints in clinical trials, where the multiple endpoints are often measured on the same participants. Benjamini and Yekutieli (2001) provided two extensions to dependent tests. They showed that under more general conditions, loosely a positive association among the tests in the context of one-sided testing, (2.2) could still be applied with the same result. They also presented an alternative procedure that was more conservative but requiring no special conditions.

Genovese, Lazar, and Nichols (2002) summarized both of the preceding papers and illustrated applications in the context of neuroimaging. Many neuroimaging experiments provide contexts for one-sided testing, such as fMRI experiments comparing activation of regions of the brain when a subject is doing a task compared to a resting condition.

Unlike many applications of FDR-controlling procedures, the application to ACS estimates of change over time at the tract level can for all practical purposes be regarded as independent tests. Thus, the original results of Benjamini and Hochberg (1995) are sufficient here. Other uses of the FDR-controlling procedures in the ACS context may require further consideration of the consequences of dependence, however.

In the context of a large number of tests involved, such as in the hundreds, the following formula provides a crude estimator, $\hat{Q}_b$, of the FDR resulting from applying a standard testing procedure. For example, in providing 90% confidence intervals directly to users, the Census Bureau facilitates testing at the q = .10 level in simple cases. Estimating the FDR according to the following formula would provide a useful perspective on the possible consequences of standard testing.

$$\hat{Q}_b = \frac{q\,(m-(V+S))}{(1-q)(V+S)} \qquad (2.3)$$

2.3 Data

Estimates for the 149 tracts in Lake County, Illinois, were downloaded from the Census Bureau's website for the two 5-year periods, 1999-2003 and 2001-2005. Three characteristics were considered. One was the characteristic of interest in the Lake County Challenge, the percent speaking Spanish at home. As Figure 1 suggested, the characteristic exhibited a relatively pronounced trend, increasing from 13.7% to 15.0% at the county level in the comparison of these two 5-year period estimates, in spite of their overlap of 3 years in common. The second characteristic was high educational attainment, specifically the percentage and the number of persons age 25+ with a graduate degree. Increasing from 15.3% to 16.0%, this variable exhibited a more moderate trend. The third characteristic, those living at another address in the U.S. 1 year ago, exhibited no appreciable trend at the county level, remaining at 13.2%.

3. Results

3.1 An Application under a Strong Trend

The growth in percent speaking Spanish at home provides an example of a strong county-level trend. Out of 149 tracts, conventional testing using the 90% confidence level points singles out 7 tracts with decreases and 24 tracts with increases. In the previous study, there was no evidence of a decrease anywhere in the county; the more relevant and challenging question was whether some areas were increasing faster than others. Application of (2.2) with q^* = .20 would reduce the count to 3 tracts decreasing, 7 increasing. In the context of a strong upward trend, however, consideration of 1-sided tests seems justified. One-sided application of (2.2) gives 0 decreasing, 22 increasing. Particularly because of this last result, the 7 tracts showing apparent decreases can be reasonably disregarded. Rather than representing evidence of specific areas running against the overall trend, the 7 apparent decreases can be ascribed to likely false discoveries.

Using (2.3) to estimate the FDR gives 42% for the original set of 7 tracts decreasing, 24 increasing. Applied to the one-sided tests with q = .05, however, (2.3) gives (effectively) 100% for the decreasing tracts and 27% for the 24

increasing tracts. Thus, $\hat{Q}_b$ from (2.3) gives the same impression about the relative strength of evidence for decreases and for increases as application of the FDR-controlling procedure, (2.2).

3.2 An Application under a Moderate Trend

The increase in the percent of graduate and professional degrees for persons age 25 and over is an example of a moderate trend. Out of 149 tracts, 2 tracts decrease and 16 tracts increase with conventional testing. Applying (2.2) in the same manner as the previous example results in 0 decreasing, 2 increasing for 2-sided testing, and the same outcome for 1-sided testing. Again, tracts apparently moving against the trend are eliminated by the FDR perspective, but the number of tracts individually established as increasing is negligibly small. Application of (2.3) gives 81% for 2-sided testing, and 100% and 44% for 1-sided testing. These largely inconclusive results discourage attempts to distinguish tracts on the basis of the comparison. Again, (2.3) provides the same overall impression as application of (2.2).

An analysis of the number of graduate or professional degrees would distinguish more tracts. This characteristic combines growth in the proportion with possible growth in the tract-level populations. Out of 149 tracts, only 1 is shown to decrease and 25 to increase under conventional testing. Applying (2.2) gives 0 decreasing, 8 increasing with 2-sided testing, and 0 decreasing, 20 increasing with 1-sided testing. Application of (2.3) gives 53% for 2-sided, 100% for 1-sided decreasing, and 26% for 1-sided increasing. Thus, either approach indicates that evidence for any decrease in number is questionable, but some tracts can be individually distinguished as increasing.

3.3 An Application without a Clear Trend

Finally, the percent living at another address shows no county trend. Out of 149 tracts, 10 tracts decrease and 9 increase. Application of (2.2) leaves 0 decreases and only 1 increase for the 2-sided test; 0 decreases and only 5 increases for the 1-sided tests. Application of (2.3) gives 73% for the 2-sided test, 82% for 1-sided decreases, and 76% for 1-sided increases. None of these results are within the target 20%. Thus, the FDR perspective largely discourages further interpretation of results that are dominated by noise.

4. Discussion

If properly introduced and encouraged, the concept of the false discovery rate may be accessible to a number of technically-oriented ACS users. Some of them may choose to apply the Benjamini and Hochberg (1995) procedure, (2.2). Others may find the estimated FDR bound, (2.3), useful as an initial screen when a large number of tests are in question, such as the previous example.

When a discernable trend at a higher level of geography provides a context, the 1-sided vs. 2-sided distinction may be quite useful. In the previous example, evidence for specific tracts changing in the direction opposite the trend was quite limited once the FDR was taken into account.

Finally, the preceding application employed a generous 20% FDR bound; very few tracts would have been significant at 5% or even 10%, more conventional levels of significance. Genovese, Lazar, and Nichols (2002) state that Benjamini was of the opinion that some analysis may be served by setting the FDR bound higher in some applications. In many cases, ACS users may agree. In the example, the 20% FDR bound limited the number of significant results far more than the Census Bureau's convention of 10%. Without some means to account for the false discovery rate, users may find that standard testing of individual hypotheses at the small area level, such as tracts, leads to many confusing and unstable results.

Acknowledgements

The author is currently associated with Westat, Inc., having retired from the U.S. Census Bureau. The responsibility for the paper rests solely with the author, since the work was done on his own time. It does not necessarily reflect the views of either organization.

References

Beaghen, M. and Weidman, L. (2007), "Statistical Issues and Interpretation of the American Community Surveys One-, Three-, and Five-Year Period Estimates," prepared for presentation at the Joint Statistical Meetings, Salt Lake City, UT, 29 Jul-2 Aug 2007.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B, 57,* 289-300.

Benjamini, Y. and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing under Dependency," *Annals of Statistics, 29,* 1165-1188.

Fay, R.E. (2007a), "The Lake County Challenge," unpublished manuscript dated 6 Dec 2007, U.S. Census Bureau.

_____ (2007b), "The Lake County Challenge," PowerPoint presentation, available from http://www.census.gov/acs/www/AdvMeth/Multi_Year_Estimates/presentations.html.

Genovese, C.E., Lazar, N.A., and Nichols, T. (2002), "Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate," *NeuroImage, 15,* 870-878.

National Research Council (2007), *Using the American Community Survey: Benefits and Challenges,* Panel on the Functionality and Usability of Data from the American Community Survey, C.F. Citro and G. Kalton (eds.), Committee on National Statistics, Division of Behavioral and Social Sciences and Education, The National Academies Press, Washington, DC.