Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Investigating alternative estimators for the prevalence of serious mental illness based on a two-phase sample

by Phillip S. Kott, Dan Liao, Jeremy Aldworth, Sarra L. Hedden, Joseph C. Gfroerer, Jonaki Bose and Lisa Colpe

Release date: June 21, 2018

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

© Her Majesty the Queen in Right of Canada as represented by the Minister of Industry, 2018

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2018
Vol. 44, No. 1, pp. 61-73
Statistics Canada, Catalogue No. 12-001-X

Investigating alternative estimators for the prevalence of serious mental illness based on a two-phase sample

Phillip S. Kott, Dan Liao, Jeremy Aldworth, Sarra L. Hedden, Joseph C. Gfroerer, Jonaki Bose and Lisa Colpe[1]

Abstract

A two-phase process was used by the Substance Abuse and Mental Health Services Administration to estimate the proportion of US adults with serious mental illness (SMI). The first phase was the annual National Survey on Drug Use and Health (NSDUH), while the second phase was a random subsample of adult respondents to the NSDUH. Respondents to the second phase of sampling were clinically evaluated for serious mental illness. A logistic prediction model was fit to this subsample with the SMI status (yes or no) determined by the second-phase instrument treated as the dependent variable and related variables collected on the NSDUH from all adults as the model's explanatory variables. Estimates were then computed for SMI prevalence among all adults and within adult subpopulations by assigning an SMI status to each NSDUH respondent based on comparing his (her) estimated probability of having SMI to a chosen cut point on the distribution of the predicted probabilities. We investigate alternatives to this standard cut point estimator such as the probability estimator. The latter assigns an estimated probability of having SMI to each NSDUH respondent. The estimated prevalence of SMI is the weighted mean of those estimated probabilities. Using data from NSDUH and its subsample, we show that, although the probability estimator has a smaller mean squared error when estimating SMI prevalence among all adults, it has a greater tendency to be biased at the subpopulation level than the standard cut point estimator.

Key Words: Bias; Bias-corrected estimator; Domain; Survey-sampling theory; Asymptotic.

1 Introduction

Serious mental illness is defined as currently or in the past year having a diagnosable mental, behavioral, or emotional disorder (excluding developmental and substance-use disorders) of sufficient duration to meet diagnostic criteria specified in the Diagnostic and Statistical Manual of Mental Disorders, 4^{th} edition (American Psychiatric Association, 1994). The National Survey on Drug Use and Health (NSDUH), sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA), provides national and state-level estimates on the use of tobacco products, alcohol, and illicit drugs in the civilian, noninstitutionalized population of the United States aged 12 years or older. The Mental Health Surveillance Study (MHSS) was a follow-up study to the NSDUH main interview, designed to estimate the prevalence of serious mental illness (SMI) among adults 18 years of age or older at the national level and within particular subpopulations of interest. It was not practical to administer clinical interviews to the entire NSDUH sample of approximately 46,000 adults per year in order to obtain estimates of SMI due to financial and time constraints. Thus, a second phase of sampling was employed as in the National Comorbidity Survey Replication (See Kessler, Chiu, Demler and Walters, 2005). From 2008 through 2012, a clinical interview was administered to a randomly-selected (before nonresponse) subsample of NSDUH adult respondents within four weeks of completing the NSDUH main interview. For more information about the

1. Phillip S. Kott, Dan Liao and Jeremy Aldworth, RTI International. E-mail: pkott@rti.org; Sarra L. Hedden, Joseph C. Gfroerer and Jonaki Bose, Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality (CBHSQ), 5600 Fishers Lane, Rockville, MD 20857, U.S.A.; Lisa Colpe, National Institute of Mental Health, 6001 Executive Blvd., Bethesda, MD 20892.

NSDUH and its clinical subsample, the reader is referred to Center for Behavioral Health Statistics and Quality (CBHSQ), 2014.

The clinical evaluations were treated as the "gold standard", despite the possibility of human error (this topic is discussed further in Section 4). A logistic prediction model was fit to the respondent subsample with the clinical SMI evaluations (yes or no) treated as the dependent variable and variables contained on the NSDUH as the model's explanatory variables. The NSDUH variables included in the prediction model, obtained from all 46,000 adult respondents in the main survey, were measures of psychological distress and functional impairment derived from responses to NSDUH survey items, an age variable, the existence of a self-reported past-year major depressive episode, and the existence of past-year serious suicidal thoughts. The fitted prediction model was then applied to the NSDUH full sample to predict every adult respondent's probability of having SMI based on their responses in the NSDUH main interview.

An adult NSDUH respondent with an estimated probability of having SMI that was greater than or equal to a cut point was assigned a predicted SMI value of 1 (i.e., has SMI); otherwise, he or she was assigned a predicted SMI value of 0 (i.e., does not have SMI). SMI prevalence rates for all adults and within various subpopulations were then estimated using these predicted values. The cut point was determined so that within the MHSS subsample the weighted fraction of false positives (subsample respondents predicted to have SMI based on the model but clinically diagnosed as *not* having SMI) and false negatives (subsample respondents predicted not to have SMI but clinically diagnosed as having SMI) came as close to being equal as possible. Strict equality was usually impossible because predicted probabilities of having SMI only take on a limited number of values depending on the covariates in the model.

The standard cut point estimator is derived from Receiver Operator Characteristics (ROC) theory. See, for example, Fawcett (2006). In Section 2, we use probability-sampling-based (also called "design-based") survey-sampling theory to propose alternatives to this estimator. One such is the probability estimator, which simply assigns each NSDUH respondent his (her) estimated probability of having SMI, with no dichotomous designation. Also introduced are bias-corrected versions of both the standard cut point and probability estimators. These are similar to standard GREG estimators (see, for example, Särndal, Swensson and Wretman, 1989) and are nearly unbiased under survey-sampling theory whether or not the logistic model holds. The bias correction term in each serves as the basis of a test statistic for determining whether the associated model-based estimator – whether cut point or probability – is significantly biased.

Section 3 uses NSDUH/MHSS subsample data collected between 2008 and 2012 to evaluate the alternative estimators using the logistic model. We show that although the probability estimator has a smaller mean squared error when estimating SMI prevalence among all adults, it has a greater tendency to be biased at the subpopulation level than the standard cut point estimator. This leads us to propose a hybrid cut point estimator that is (at least) more efficient than the standard cut point estimator for all adults while not having the tendency to be biased at the subpopulation level.

Section 4 provides some concluding remarks. It is important to remember that SAMHSA planned to use the logistic model fit in the 2008 to 2012 clinical subsamples to help compute annual estimates of adult SMI prevalences based on NSDUH responses beyond 2012 without any new clinical subsamples.

2 Some estimators

2.1 Across all adults

Let S denotes the relevant NSDUH respondent sample (adults 18 years or older) from 2008 through 2012, and w_k the NSDUH analysis (first-phase) weight for an individual $k \in S$. Let S' denotes the subsample of S responding to a clinical evaluation of their SMI status. Let $y_k = 1$ when k is diagnosed to have serious mental illness, and $y_k = 0$ when k is diagnosed not to have serious mental illness. Let ω_k be the two-phase weight for an individual k in S'. For convenience, we set ω_k to 0 for individuals in S but not S'.

In actual practice, both sets of weights have been adjusted to account for nonresponse and undercoverage and to increase their efficiency, but we will ignore that fact here for simplicity. Instead, we will assume $1/w_k$ is the probability of selection for a NSDUH respondent, $1/\omega_k$ the probability of selection for a MHSS subsample respondent, and thus w_k/ω_k the conditional selection probability of a subsample respondent given (s)he was a NSDUH respondent. A nearly unbiased estimator for the prevalence of SMI among adults between 2008 and 2012 based on the two-phase sample is $\bar{y}_U = \sum_{S'} \omega_k y_k \big/ \sum_{S'} \omega_k$, "nearly" because the denominator may contain some sampling error.

Suppose a $\omega_k -$weighted logistic regression is run on the all-adult MHSS subsample respondents in S' with y_k as the dependent variable using a reasonable vector of explanatory covariates, $\mathbf{x}_k$, available for every respondent in the adult NSDUH sample. Exactly how the covariates have been chosen is beyond the scope of this investigation (for that, the reader is directed to Center for Behavioral Health Statistics and Quality, 2015; Chapter 4). Let the predictor for y_k from this weighted-logistic regression be $p_k = p(\mathbf{x}'_k\mathbf{b}) = [1 + \exp(-\mathbf{x}'_k\mathbf{b})]^{-1}$.

The use of weights in fitting the logistic-regression model protects against the possibility that the model residuals are correlated with the probabilities of selection. It is also consistent with how SMI prevalence was estimated; that estimate resulted from the weighted regression of y_k on the constant 1 and no covariates.

Sorting the subsample by the p_k values, one can find the cut point value p_C such that

$$\sum_{\substack{k \in S' \\ p_k \geq p_C}} \omega_k = \sum_{k \in S'} \omega_k y_k \tag{2.1}$$

holds exactly or as nearly so as possible. That is to say, the estimated number of adults in the population having $p_k -$values at or above the cut point approximately equals the estimated number of adults with SMI. Define an indicator random variable c_k to be 1 when $p_k \geq p_C$ and 0 otherwise. A cut point determined using equation (2.1) also comes as close as possible to equalizing the weighted false positives $\left(\sum_{S':c_k=1} \omega_k (1 - y_k)\right)$ and false negatives $\left(\sum_{S':c_k=0} \omega_k y_k\right)$ in S'.

Two alternative estimators for SMI prevalence among adults are the model-driven *cut point* and *probability* estimators:

$$\bar{y}_C = \frac{\sum_S w_k c_k}{\sum_S w_k}, \tag{2.2}$$

and

$$\bar{y}_P = \frac{\sum_S w_k p_k}{\sum_S w_k}, \tag{2.3}$$

these estimators are computed using the entire NSDUH sample rather than the smaller MHSS subsample as is $\bar{y}_U$.

We assume now that one of the covariates in the logistic model is 1 or the equivalent $(\mathbf{x}_k'\boldsymbol{\gamma} = 1$ for some $\boldsymbol{\gamma})$. Under this assumption, the probability estimator for SMI prevalence is exactly equal to a *bias-corrected* probability estimator given below:

$$\begin{aligned} \bar{y}_{P\text{-BC}} &= \frac{\sum_{S'} \omega_k y_k}{\sum_{S'} \omega_k} + \left(\frac{\sum_S w_k p_k}{\sum_S w_k} - \frac{\sum_{S'} \omega_k p_k}{\sum_{S'} \omega_k} \right) \\ &= \frac{\sum_S w_k p_k}{\sum_S w_k} + \frac{\sum_{S'} \omega_k (y_k - p_k)}{\sum_{S'} \omega_k}. \end{aligned} \tag{2.4}$$

The equality between $\bar{y}_P$ and $\bar{y}_{P\text{-BC}}$ results from the numerator of the *bias-correction term* in the second line of equation (2.4), $\sum_{S'} \omega_k (y_k - p_k) \big/ \sum_{S'} \omega_k$, equaling zero. Fitting a logistic regression forces $\sum_{S'} \omega_k (y_k - p_k)\, \mathbf{x}_k = \mathbf{0}$, and we have assumed $\mathbf{x}_k$ contains 1 or the equivalent.

Since the expectation of the term in parentheses in the *first* line of equation (2.4) is nearly zero under mild conditions, $\bar{y}_P = \bar{y}_{P\text{-BC}}$, like $\bar{y}_U$, is nearly unbiased under survey-sampling theory. *This is true whether or not the model used to determine the* p_k *is correct* so long as $\mathbf{b}$ in $p_k = p(\mathbf{x}_k'\mathbf{b}) = [1 + \exp(-\mathbf{x}_k'\mathbf{b})]^{-1}$ converges to *something* as the MHSS subsample and NSDUH sample sizes grow arbitrarily large.

The estimator $\bar{y}_{P\text{-BC}}$ is analogous to the popular GREG estimator. It follows Lehtonen and Veijanen (1998), and computes the p_k with a logistic rather than the linear model of the GREG.

A bias-corrected cut point estimator is

$$\begin{aligned} \bar{y}_{C\text{-BC}} &= \frac{\sum_{S'} \omega_k y_k}{\sum_{S'} \omega_k} + \left(\frac{\sum_S w_k c_k}{\sum_S w_k} - \frac{\sum_{S'} \omega_k c_k}{\sum_{S'} \omega_k} \right) \\ &= \frac{\sum_S w_k c_k}{\sum_S w_k} + \frac{\sum_{S'} \omega_k (y_k - c_k)}{\sum_{S'} \omega_k}. \end{aligned} \tag{2.5}$$

Using the same logic as above, this estimator is also nearly unbiased under mild conditions. It is close to the model-driven cut point estimator since the bias-correction term, $\sum_{S'} \omega_k (y_k - c_k) \big/ \sum_{S'} \omega_k$, is close to zero. The bias-correction term would be exactly zero if there were a cut point p_C that satisfied equation (2.1) exactly.

2.2 Domain estimation

Let us now turn our attention to a subpopulation of all adults, such as males or all adults who have received treatment for mental illness (or all adults who live in a particular state). We call such a subpopulation a "domain" of interest. To estimate SMI prevalence in a domain, we can simply insert an indicator for domain membership, d_k, equaling 1 when k is in the domain, 0 otherwise, into all our estimates:

$$\bar{y}_{U(d)} = \frac{\sum_{S'} \omega_k y_k d_k}{\sum_{S'} \omega_k d_k} \tag{2.6}$$

$$\bar{y}_{P(d)} = \frac{\sum_{S} w_k p_k d_k}{\sum_{S} w_k d_k} \tag{2.7}$$

$$\bar{y}_{C(d)} = \frac{\sum_{S} w_k c_k d_k}{\sum_{S} w_k d_k} \tag{2.8}$$

$$\begin{aligned} \bar{y}_{P-\mathrm{BC}(d)} &= \bar{y}_{U(d)} + \left(\frac{\sum_{S} w_k p_k d_k}{\sum_{S} w_k d_k} - \frac{\sum_{S'} \omega_k p_k d_k}{\sum_{S'} \omega_k d_k} \right) \\ &= \frac{\sum_{S} w_k p_k d_k}{\sum_{S} w_k d_k} + \frac{\sum_{S'} \omega_k \left(y_k - p_k \right) d_k}{\sum_{S'} \omega_k d_k} \end{aligned} \tag{2.9}$$

$$\begin{aligned} \bar{y}_{C-\mathrm{BC}(d)} &= \bar{y}_{U(d)} + \left(\frac{\sum_{S} w_k c_k d_k}{\sum_{S} w_k d_k} - \frac{\sum_{S'} \omega_k c_k d_k}{\sum_{S'} \omega_k d_k} \right) \\ &= \frac{\sum_{S} w_k c_k d_k}{\sum_{S} w_k d_k} + \frac{\sum_{S'} \omega_k \left(y_k - c_k \right) d_k}{\sum_{S'} \omega_k d_k}. \end{aligned} \tag{2.10}$$

It is here where the bias-correction terms serve an important purpose. If the logistic model, which was fit on the subsample of *all* adults, holds within the domain, then $\sum_{S'} \omega_k d_k \left(y_k - p_k \right) \big/ \sum_{S'} \omega_k d_k$ will be an estimate of zero, and the model-driven probability estimator, $\bar{y}_{P(d)}$ in equation (2.7), will be nearly unbiased. If the model does not hold in the domain (e.g., if males are more likely to have SMI than the model predicts), then the model-driven probability estimator can be significantly biased.

Adding the bias correction $\sum_{S'} \omega_k d_k \left(y_k - p_k \right) \big/ \sum_{S'} \omega_k d_k$ to $\bar{y}_{P(d)}$ produces an estimator that is nearly unbiased under survey-sampling theory. When the model holds in the domain, however, applying the correction will almost certainly result in a decrease in accuracy. A similar argument can be made about the appropriateness of adding the $\sum_{S'} \omega_k d_k \left(y_k - c_k \right) \big/ \sum_{S'} \omega_k d_k$ term in equation (2.10) to the cut point estimator, $\bar{y}_{C(d)}$, in equation (2.8).

Equations (2.4) and (2.5) can be viewed as special cases of (2.9) and (2.10), respectively, with $d_k \equiv 1$.

3 The MHSS subsample

3.1 About the MHSS subsample

The NSDUH is a stratified multi-stage probability survey. In 2008 through 2012, the MHSS subsample was drawn annually from adults responding to the corresponding NSDUH using Poisson sampling. Subsample selection probabilities were determined each year using an algorithm that tended to oversample adults with higher levels of psychological distress. The algorithm varied across the years. See Center for Behavioral Health Statistics and Quality (2014, Chapter 3) for more details.

A respondent subsample size of roughly 750 was targeted for 2008 while respondent subsamples of 500 each for the 2009 and 2010, and 1,500 each for the 2011 and 2012 were likewise targeted. A data set combining all the respondent from 2008 to 2012 was created for modeling SMI. Weights for modeling were developed assuming that the same model held across all the years. As a result, more weight was given to the samples from 2011 and 2012 than to earlier years (Center for Behavioral Health Statistics and Quality, 2014; Chapter 5).

For our purposes, we treat those subsample weights and associated NSDUH weights as given and based on survey-sampling theory. We also treat the strata and two variance primary sampling units (PSUs) per each of the 50 variance strata developed for the MHSS subsample variance estimator as if they were the NSDUH variance strata and variance PSUs. Finally, we treat the NSDUH PSUs as if they were selected with replacement.

3.2 Variance estimation under survey-sampling theory

Since the bias-corrected estimated domain totals in equations (2.9) and (2.10) are nearly unbiased under survey-sampling theory, one can use linearization to estimate their variances. In what follows, we use variants of the bias-corrected estimators in equation (2.9) and (2.10) to simplify the variance estimation.

Recalling that $\omega_k = 0$ when $k \notin S'$, a variance estimator for the sample mean

$$\bar{y}_{z(d)} = \frac{\sum_S w_k z_k d_k}{\sum_S w_k d_k}, \tag{3.1}$$

under a stratified, multistage sample, where $z_k = p_k + (\omega_k / w_k)(y_k - p_k)$ is

$$v(\bar{y}_{z(d)}) = \frac{\sum_{h=1}^{50} \left[\sum_{k \in S_{h1}} w_k d_k (z_k - \bar{y}_{z(d)}) - \sum_{k \in S_{h2}} w_k d_k (z_k - \bar{y}_{z(d)})\right]^2}{\left(\sum_S w_k d_k\right)^2}, \tag{3.2}$$

where S_{hj} are the respondents in the j^{th} variance PSU and variance stratum h. It is also a variance estimator for the following asymptotically-identical variant of the bias-corrected probability estimator:

$$\bar{y}_{P-\mathrm{BC2}(d)} = \frac{\sum_S w_k p_k d_k}{\sum_S w_k d_k} + \frac{\sum_S \omega_k (y_k - p_k) d_k}{\sum_S w_k d_k}. \tag{3.3}$$

This is because the MHSS subsample is Poisson (and thus independent across adults as well as PSUs) and the first stage of the NSDUH sample is treated as if it were drawn with replacement.

Similarly, by redefining $z_k = c_k + (\omega_k / w_k)(y_k - c_k)$, a variance estimator for the sample mean in equation (3.1) is also an estimator for that variance of this variant of the bias-corrected estimator:

$$\bar{y}_{C-\mathrm{BC}2(d)} = \frac{\sum_S w_k c_k d_k}{\sum_S w_k d_k} + \frac{\sum_S \omega_k (y_k - c_k) d_k}{\sum_S w_k d_k}. \tag{3.4}$$

The variance estimation approach taken above assumes that the domain respondent subsample sizes are such that p_k / P_k and c_k / C_k can be treated as unity, where P_k and C_k are the limits of p_k and c_k, respectively, as the subsample (along with the NSDUH sample and population) grows arbitrarily large. In fact, all these ratios are assumed to be $1 + \mathrm{O}_P\left(1/\sqrt{n}\right)$, where n is the MHSS subsample size.

Consider now a computed bias-correction term, say $\sum_S \omega_k (y_k - p_k) d_k / \sum_S w_k d_k$ or $\sum_S \omega_k (y_k - c_k) d_k / \sum_S w_k d_k$. To assess whether the term is significantly different from zero, one can create an asymptotic $t-$ statistic in the usual fashion, dividing the term by its standard error.

When evaluating the estimators in Section 3.3, we will instead use the asymptotically equivalent:

$$BiasMeasure(\bar{y}_{P(d)}) = \sum_{S'} \omega_k (y_k - p_k) d_k / \sum_{S'} \omega_k d_k, \tag{3.5}$$

and

$$BiasMeasure(\bar{y}_{C(d)}) = \sum_{S'} \omega_k (y_k - c_k) d_k / \sum_{S'} \omega_k d_k \tag{3.6}$$

to create asymptotic $t-$ statistics for evaluating domain-level biases so that the DESCRIPT procedure in SUDAAN (RTI International, 2012) can be employed treating the p_k and c_k as fixed (similarly, that variance estimator for (3.3) in equation (3.2) can be computed using DESCRIPT). Moreover, since virtually all the sampling error in the bias-correction terms comes from the MHSS subsampling phase (even in 2011 and 2012, subsample was only 3% of the NSDUH adult sample), we treat the standard errors of the bias measures as if they were computed for a Poisson sample with ignorably small sampling fractions, which is equivalent to a with-replacement element sample for variance-estimation purposes. For example, for the variance estimator of $BiasMeasure(\bar{y}_{P(d)})$, we compute (using SUDAAN's DESCRIPT): $v[BiasMeasure(\bar{y}_{P(d)})] = \frac{n}{n-1} \sum_{S'} [\omega_k \{[y_k - p_k] - BiasMeasure(\bar{y}_{P(d)})\} d_k]^2 / \left(\sum_{S'} \omega_k d_k\right)^2$, where n is the sample size of S'.

3.3 Evaluating the estimators

The model used by SAMHSA to predict SMI from adult NSDUH respondents was a logistic model with five variables (Center for Behavioral Health Statistics and Quality, 2014; Chapter 7). Two of the variables were rescaled total scores from short forms that measure psychological distress and functional impairment due to distress. The third was a dichotomous $(0/1)$ variable created from the answers to a series of questions assessing whether the respondent had a major depressive episode in the previous year. The fourth was also $0/1$ and indicated whether the respondent seriously contemplated suicide in the past year, and the fifth was a linear function of age from 18 to 30 that stayed constant after 30. Details on how this model was selected can be found in Center for Behavioral Health Statistics and Quality (2015, Chapter 4).

We used that model to create a set of domain-level cut point and probability estimates from the combined 2008-2012 data sets and to evaluate their potential biases. Some of the results are displayed in Tables 3.1

and 3.2. These tables reviewed domain estimates based on personal characteristics rather than state of residence because it seemed more likely that significant biases would be found for the characteristics like these rather than for states. Moreover, sample sizes for characteristics tended to be larger than those for states.

Table 3.1 show that using the bias-corrected probability in equation (2.9) is usually slightly more efficient (has a smaller standard error) than the direct estimator $\bar{y}_{U(d)} = \sum_{S'} \omega_k d_k y_k \big/ \sum_{S'} \omega_k d_k$. The bias-corrected cut point estimator in equation (2.10) is sometimes more efficient than the direct estimator, sometimes not. The standard errors in Table 3.1 are the square roots of linearization variance estimators for the direct estimator $\bar{y}_{U(d)}$ above or the bias-corrected estimator in equation (3.1) with the appropriated defined nonrandom z_k, each computed as a stratified with-replacement sample of primary sampling units and a probability subsample of individuals within each PSU; that is, with equation (3.2). For $v(\bar{y}_{U(d)})$, $z_k - \bar{y}_{z(d)}$ is replaced by $y_k - \bar{y}_{U(d)}$.

Table 3.1
Nearly unbiased estimators with their standard errors

	Direct (eq. 2.6)		Bias-Corrected Cut Point (eq. 2.10)		Bias-Corrected Probability (eq. 2.9)	
	Estimate	SE	Estimate	SE*	Estimate	SE*
All Adults	3.93	0.29	3.96	0.26	3.91	0.23
Male	2.96	0.34	2.91	0.39	3.01	0.31
Female	4.84	0.46	4.93	0.39	4.74	0.36
Age: 18-25	3.77	0.62	3.97	0.48	3.66	0.52
Age: 26-34	4.35	0.68	4.29	0.61	4.37	0.57
Age: 35-49	5.74	0.57	6.15	0.52	5.87	0.50
Age: 50+	2.74	0.40	2.47	0.47	2.60	0.36
White, Not Hispanic	4.43	0.35	4.47	0.30	4.34	0.27
Black, Not Hispanic	3.28	0.54	3.62	0.42	3.38	0.40
Other, Not Hispanic	4.09	1.25	4.27	1.10	4.33	1.12
Hispanic	2.02	0.71	1.68	0.88	2.11	0.70
Northeast	2.80	0.51	3.59	0.49	3.25	0.47
North Central	4.17	0.49	3.99	0.53	4.13	0.37
South	3.74	0.49	3.93	0.51	3.65	0.45
West	5.04	0.84	4.26	0.57	4.62	0.57
Employed Full Time	2.36	0.29	2.36	0.28	2.32	0.25
Employed Part time	4.34	0.71	3.82	0.55	3.91	0.46
Unemployed	5.64	1.22	6.57	0.92	6.13	0.90
Other Employment Status	6.21	0.66	6.22	0.64	6.15	0.55
Less than High School	5.69	0.99	4.44	0.77	4.72	0.71
High School Graduate	4.05	0.57	4.08	0.57	4.14	0.44
Some College	4.14	0.57	4.31	0.44	4.18	0.40
College Graduate	2.88	0.52	3.27	0.46	3.01	0.46
Metro	3.78	0.45	3.96	0.39	3.74	0.37
Small Metro	4.15	0.47	3.60	0.44	3.96	0.29
Nonmetro	3.99	0.47	4.63	0.54	4.36	0.48
Health Insurance: Yes	3.57	0.31	3.83	0.26	3.65	0.24
Health Insurance: No	5.73	0.94	4.65	0.93	5.24	0.74
< 100% of Poverty Level	9.01	1.30	9.00	1.23	8.62	1.05
100%-199% of Poverty	5.61	0.85	4.72	0.63	4.88	0.52
100% of Poverty	2.59	0.28	2.64	0.28	2.61	0.23
Rec'd MH Treatment: Yes	18.84	1.57	19.69	1.29	19.00	1.32
Rec'd MH Treatment: No	1.54	0.18	1.42	0.20	1.46	0.15

* Standard error is the square root of variance estimate computed using equation (3.2) with appropriately defined z_k.

The table suggests that there is little gain to be had from model correction, and leads us back to the model-driven probability and cut point estimators in equations (2.7) and (2.8) unless they exhibit systematic biases. Table 3.2 (with bias measures and their standard errors computed as described in Section 3.2) strongly suggests that the probability estimator, although unbiased when estimating SMI prevalence among all adults, can be very biased at the domain level. The cut point estimator, by contrast, is significantly biased at the 0.1 level in only two domains and never at the 0.05 level. Since we computed two-sided $p-$values for 32 domains, finding two domains with $p-$values below 0.1 is about what one should expect under the null hypothesis that the cut point estimator is not biased at the domain level.

Table 3.2
Model-driven estimates and their bias measures

	Standard cut point (eq. 2.8)			Probability (eq. 2.7)		
	Estimate	Bias Measure	SE of Bias Measure	Estimate	Bias Measure	SE of Bias Measure
All Adults	3.95	-0.01	0.27	3.91	0.00	0.23
Male	2.99	0.08	0.42	3.18	0.17	0.34
Female	4.84	-0.10	0.34	4.58	-0.16	0.31
Age: 18-25	3.94	-0.02	0.55	3.59	-0.07	0.49
Age: 26-34	5.03	0.69	0.66	4.64	0.26	0.51
Age: 35-49	5.08	-1.10*	0.57	4.77	-1.15**	0.55
Age: 50+	2.84	0.37	0.42	3.21	0.61*	0.32
White, Not Hispanic	4.31	-0.17	0.33	4.18	-0.16	0.28
Black, Not Hispanic	3.14	-0.48	0.45	3.38	0.00	0.46
Other, Not Hispanic	3.14	-1.14	1.13	3.47	-0.86	1.08
Hispanic	3.31	1.63*	0.85	3.28	1.17	0.65
Northeast	3.55	-0.04	0.39	3.62	0.33	0.35
North Central	4.16	0.16	0.60	4.02	-0.10	0.40
South	3.80	-0.13	0.52	3.86	0.22	0.44
West	4.28	0.02	0.56	4.10	-0.56	0.55
Employed Full Time	2.76	0.38	0.33	3.09	0.75**	0.28
Employed Part time	4.19	0.39	0.59	4.05	0.15	0.47
Unemployed	6.61	0.03	0.75	5.48	-0.57	0.70
Other Employment Status	5.33	-0.93	0.66	4.91	-1.30	0.56
Less than High School	4.34	-0.11	0.90	4.15	-0.64	0.83
High School Graduate	4.09	0.01	0.59	3.92	-0.22	0.46
Some College	4.50	0.18	0.37	4.35	0.17	0.31
College Graduate	3.09	-0.16	0.46	3.36	0.33	0.40
Metro	3.63	-0.34	0.38	3.68	-0.06	0.35
Small Metro	4.35	0.73	0.49	4.20	0.23	0.35
Nonmetro	4.24	-0.38	0.59	4.09	-0.27	0.51
Health Insurance: Yes	3.67	-0.16	0.27	3.72	0.07	0.24
Health Insurance: No	5.39	0.72	0.86	4.89	-0.34	0.68
< 100% of Poverty Level	7.21	-2.07	1.27	6.13	-2.88**	1.16
100%-199% of Poverty	4.83	0.12	0.62	4.53	-0.38	0.55
100% of Poverty	2.98	0.32	0.28	3.24	0.61***	0.21
Rec'd MH Treatment: Yes	18.33	-1.37	1.31	13.97	-5.07***	1.26
Rec'd MH Treatment: No	1.62	0.20	0.23	2.28	0.81***	0.17

* Bias measure is significantly different from zero at the 0.1 level.
** Bias measure is significantly different from zero at the 0.05 level.
*** Bias measure is significantly different from zero at the 0.01 level.

One curious result bears a brief mention. The cut point estimator among all adults had very little bias (-0.01), so its estimated root mean squared error equaled the standard error of the bias-corrected cut point after rounding (0.26). Oddly, this value was less than the standard error of its bias measure (0.27). One possible reason for the difference between the two standard errors was that we used $\sum_S \omega_k (y_k - c_k) d_k \big/ \sum_S w_k d_k$ as the bias-correction term and $\sum_S \omega_k (y_k - c_k) d_k \big/ \sum_S \omega_k d_k$ as the bias measure within a domain; all adults being the special case where $d_k \equiv 1.$ Our analysis (not shown) was that the difference in the denominators had very little impact.

What has a greater impact was ignoring the stratification and clustering in the NSDUH sample when computing the standard errors of the bias measures. Unexpectedly, ignoring the clustering actually tended to increase standard errors. This may be because the clustering in the NSDUH has virtually no measurable impact on variance so that any difference between standard error estimates computed with and without clustering is attributable to random noise or to asymptotic biases that are not actually ignorable in finite estimates.

3.4 A hybrid cut point

Consider the following hybrid of the probability and standard cut point estimators. Suppose we sorted the NSDUH sample rather than just the MHSS subsample by the fitted p_k values, and established a cut point p_H such that

$$\sum_{\substack{k \in S \\ p_k \geq p_H}} w_k = \sum_{k \in S} w_k p_k \tag{3.7}$$

holds as closely as possible. Setting $h_k = 1$ when $p_k > p_H$ and 0 otherwise, the hybrid cut point estimator for SMI prevalence in a domain is

$$\bar{y}_{H(d)} = \frac{\sum_S w_k h_k d_k}{\sum_S w_k d_k}. \tag{3.8}$$

It is not hard to see that for all adults, if a p_H could be found that satisfied equation (3.7), then the hybrid cut point estimator would equal the probability estimator exactly. Failing that the hybrid cut point estimator for all adults would have a slight bias, which could be measured, squared, and then added to the standard error of the probability estimator to equal its root mean squared error. In this case, the hybrid SMI prevalence estimate for all adults rounded to 3.89. Its root mean squared error rounded to the same value as the standard error of the probability estimator (0.23).

Table 3.3 repeats much of Table 3.2 for the standard cut point but also displays analogous results for the hybrid

$$(BiasMeasure(\bar{y}_{H(d)})) = \sum_S \omega_k (y_k - h_k) d_k \big/ \sum_S \omega_k d_k. \tag{3.9}$$

Its standard error is computed analogously to those of $\bar{y}_{C(d)}$ and $\bar{y}_{P(d)}$. The two sets of cut point outcomes are similar, but the bias measure for the hybrid estimator was significantly different from zero at the 0.05 level in two domains (both with $p-$values of 0.043). Since there are 32 domains analyzed, this remains consistent with the null hypothesis of no bias at the domain level.

Table 3.3
The cut point estimators and their bias measures

	Standard Cut Point (eq. 2.8)			Hybrid Cut Point (eq. 3.8)		
	Estimate	Bias Measure	SE of Bias Measure	Estimate	Bias Measure	SE of Bias Measure
All Adults	3.95	-0.01	0.27	3.89	-0.10	0.27
Male	2.99	0.08	0.42	2.94	0.03	0.42
Female	4.84	-0.10	0.34	4.78	-0.21	0.33
Age: 18-25	3.94	-0.02	0.55	3.89	-0.03	0.55
Age: 26-34	5.03	0.69	0.66	4.97	0.68	0.66
Age: 35-49	5.08	-1.10*	0.57	5.02	-1.16**	0.57
Age: 50 or Older	2.84	0.37	0.42	2.79	0.22	0.41
White, Not Hispanic	4.31	-0.17	0.33	4.24	-0.22	0.33
Black, Not Hispanic	3.14	-0.48	0.45	3.10	-0.48	0.45
Other, Not Hispanic	3.14	-1.14	1.13	3.11	-1.14	1.13
Hispanic	3.31	1.63*	0.85	3.25	1.30	0.79
Northeast	3.55	-0.04	0.39	3.50	-0.05	0.39
North Central	4.16	0.16	0.60	4.12	0.07	0.59
South	3.80	-0.13	0.52	3.74	-0.29	0.51
West	4.28	0.02	0.56	4.23	0.01	0.56
Employed Full Time	2.76	0.38	0.33	2.71	0.36	0.33
Employed Part Time	4.19	0.39	0.59	4.16	0.37	0.59
Unemployed	6.61	0.03	0.75	6.43	-0.27	0.69
Other Employment Status	5.33	-0.93	0.66	5.27	-1.09*	0.65
Less than High School	4.34	-0.11	0.90	4.21	-0.14	0.90
High School Graduate	4.09	0.01	0.59	4.03	-0.26	0.56
Some College	4.50	0.18	0.37	4.45	0.18	0.37
College Graduate	3.09	-0.16	0.46	3.07	-0.17	0.46
Large Metro	3.63	-0.34	0.38	3.58	-0.36	0.38
Small Metro	4.35	0.73	0.49	4.27	0.58	0.47
Nonmetro	4.24	-0.38	0.59	4.19	-0.53	0.58
Health Insurance: Yes	3.67	-0.16	0.27	3.62	-0.20	0.27
Health Insurance: No	5.39	0.72	0.86	5.31	0.44	0.82
< 100% of Poverty Threshold	7.21	-2.07	1.27	7.12	-2.44**	1.21
100%-199% of Poverty	4.83	0.12	0.62	4.78	-0.01	0.61
> 200% of the Poverty	2.98	0.32	0.28	2.93	0.30	0.28
Rec'd MH Treatment: Yes	18.33	-1.37	1.31	18.19	-1.46	1.31
Rec'd MH Treatment: No	1.62	0.20	0.23	2.28	0.81	1.17

* Bias measure is significantly different from zero at the 0.1 level.
** Bias measure is significantly different from zero at the 0.05 level.

4 Some concluding remarks

Population mental health attributes have been estimated using data from a subsample selected from a large general survey that are given a clinical diagnostic assessment to develop prediction models that are then applied to the full sample (see, for example, Kessler, Abelson, Demler, Escobar, Gibbon, Guyer, Howes, Jin, Vega, Walters, Wang, Zaslavsky and Zheng, 2004). This is the methodology used by SAMHSA with the annual NSDUH and the adult MHSS subsample from 2008 through 2012.

We have shown with NSDUH/MHSS subsample data that when estimating the SMI prevalence for the full population, in this case adults, using the estimated probabilities of having SMI directly instead of the standard cut point methodology results in a lower standard error. Nevertheless, the so-called probability estimator of SMI prevalence can often be substantially biased at domain-level when the standard cut point is not.

We also investigated bias-corrected versions of the two estimators based on survey-sampling theory. Unfortunately, these estimators were only slightly more efficient than simply computing estimates from the MHSS subsample directly, especially for subpopulations.

We evaluated a hybrid cut point estimator that was slightly more efficient than the standard cut point in estimating SMI prevalence among all adults. It did not exhibit the large biases at the domain level that plagued the probability estimator, although whether or not it was free of domain-level bias was not completely clear.

In 2013, SAMHSA discontinued the collection of clinical interviews. Nevertheless, the agency continues to compute SMI prevalence estimates for adults based on the model and cutpoints developed using the 2008-2012 MHSS subsample. The standard cut point estimator which demonstrated smaller domain-level biases in our tables also continues to be used for SMI estimation at the state level as well as for the domains analyzed here.

A troubling question is how a standard cut point estimator for SMI derived from a logistic-model fit of MHSS subsample data can be less prone to bias than a probability estimator based on the sample model fit. We suspect that logistic regression does a reasonably good job at ordering the relative probabilities of adults having SMI but not at estimating individual probabilities, especially in the tails. Domains with unusually high and low prevalences, like adults receiving mental-health treatment (or not receiving treatment) were particularly prone to having biased probability estimates. It may be that the application of asymptotic theory is not appropriate in the tails.

The interested reader would likely want to know how one can compute standard errors for the cut point estimators at the domain level. Because a cut point estimator is not continuous, we attempted to compute standard errors for it using Fay's Balances repeated replications (BRR). Unfortunately, as explained in Center for Behavioral Health Statistics and Quality (2015, Chapter 2.4.2.), our variance estimators for domain estimates were not satisfactory in a modest simulation experiment.

Few MHSS subsample respondents were in the tails of the distribution (and even fewer realized values for the probability of having SMI given the covariates of the model). This frustrated our attempts at improving the probability estimator (for which domain-level standard errors can be measured via linearization) by adjusting tail probabilities (which did not fit very well under the logistic model). We will not discuss these attempts further here. More research is clearly needed to develop either good standard-error measures for standard cut point estimates or good estimates of SMI prevalence for which standard errors can be reasonably measured.

A final word about the "gold standard" is warranted. For our purposes, a clinical diagnosis of an adult having SMI has been treated as equivalent to the person actually having diagnosable seriously mental illness. In fact, diagnoses are more fluid. They may vary according to the clinician or the whim of the person answering the clinician's questions. CBHSQ (2014, Chapters 5 and 6) describes the effort expended on removing as much variation from the MHSS clinical diagnoses as possible. We assumed here that each clinical diagnosis was effectively unbiased, that is, the probability of a random diagnosis for an individual within a domain being a false positive equaled the probability of it being a false negative.

References

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* (4th Ed.). Washington, DC: American Psychiatric Association.

Center for Behavioral Health Statistics and Quality (2015). *Estimating Mental Illness among Adults in the United States: Revisions to the 2008 Estimation Procedures*. Substance Abuse and Mental Health Services Administration, Rockville, MD: http://www.samhsa.gov/data/sites/default/files/NSDUH-N8-EstimatingMI-2012.pdf.

Center for Behavioral Health Statistics and Quality (2014). *2012 National Survey on Drug Use and Health: Methodological Resource Book (Section 16a, 2012 Mental Health Surveillance Study: Design and Estimation Report)*. Substance Abuse and Mental Health Services Administration, Rockville, MD: http://www.samhsa.gov/data/sites/default/files/NSDUH2012MRB-Ammended/NSDUHmrbMHSS-DesignEst2012.pdf.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861-874.

Kessler, R.C., Abelson, J., Demler, O., Escobar, J., Gibbon, M., Guyer, M., Howes, M., Jin, R., Vega, W., Walters, E., Wang, P., Zaslavsky, A. and Zheng, H. (2004). Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMHCIDI). *International Journal of Methods in Psychiatric Research*, 13, 122-139.

Kessler, R.C., Chiu, W., Demler, O. and Walters, E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, 62, 617-627.

Lehtonen, R., and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24, 1, 51-55. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1998001/article/3909-eng.pdf.

RTI International (2012). *SUDAAN Language Manual, Volumes 1 and 2, Release 11*. Research Triangle Park, NC: Research Triangle Institute.

Särndal, C.-E., Swensson, B. and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator. *Biometrika*, 76, 527-537.