Examination of Skewed Health Expenditure Data from the Medical Expenditure Panel Survey (MEPS)

William W. Yu and Steven Machlin

Agency for Healthcare Research and Quality Working Paper No. 04002

October 2004

Suggested citation: Yu WW, Machlin S. Examination of Skewed Health Expenditure Data from the Medical Expenditure Panel Survey (MEPS). Agency for Healthcare Research and Quality Working Paper No. 04002, October 2004, http://www.ahrq.gov.

AHRQ Working Papers provide preliminary analysis of substantive, technical, and methodological issues. The papers are works in progress and have not undergone a formal peer review. They are distributed to share valuable experience and research. Comments are welcome and should be directed to the authors. The views expressed are those of the authors and no official endorsement by the Agency for Healthcare Research and Quality or the Department of Health and Human Services is intended or should be inferred.

Examination of Skewed Health Expenditure Data from the Medical Expenditure Panel Survey (MEPS)
William W. Yu and Steve Machlin
October 2004

ABSTRACT

The Medical Expenditure Panel Survey Household Component (MEPS-HC) is designed to provide nationally representative annual estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. The expenditure data from MEPS have been shown to exhibit a marked positive skewness, with a few high expenditure respondents and many low or zero expenditure respondents. As a consequence of this departure from the normal distribution, the frequency with which a conventional confidence interval for a MEPS expenditure estimate will not capture the true population parameter may be higher than the probability stated for the confidence interval. Based on repeated sample simulations using data from the 1996 to 2001 MEPS-HC, this paper evaluates and compares the actual probability achieved for confidence intervals derived from expenditure data by types of expenditure and varying sample sizes. The results are also compared to estimated confidence probabilities obtained from repeated sample simulations for other types of variables that do not exhibit as marked a positive skewness as health care expenditures.

William W. Yu
Statistician, Center for Financing, Access, and Cost Trends
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
E-mail: wyu@ahrq.gov

Steve Machlin
Senior Statistician, Center for Financing, Access, and Cost Trends
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
E-mail: smachlin@ahrq.gov

Examination of Skewed Health Expenditure Data from the Medical Expenditure Panel Survey (MEPS)

Introduction

The Medical Expenditure Panel Survey (MEPS) is designed to provide nationally representative annual estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. It is co-sponsored by the Agency for Healthcare Research and Quality (AHRQ) and the National Center for Health Statistics (NCHS).

The expenditure data from MEPS have been shown to exhibit a marked positive skewness, with a few high expenditure respondents and many low or zero expenditure respondents. As a consequence of this departure from the normal distribution, the frequency with which a conventional confidence interval for a MEPS expenditure estimate will not capture the true population parameter may be higher than the probability stated for the confidence interval.

Based on repeated sample simulations using data from the 1996 to 2001 MEPS, this paper evaluates the "actual" probability achieved for confidence intervals derived from expenditure data by types of expenditure and varying sample sizes. The results are also compared to estimated confidence probabilities obtained from repeated sample simulations for estimated proportions that do not exhibit as marked a positive skewness as health care expenditures.

MEPS Household Component

The core survey for MEPS is the Household Component (HC). The MEPS-HC collects data through an overlapping panel design. In this design, data are collected through a series of five rounds of interviews over a period of two and a half years. Interviews are conducted with one member of each family who reports on the health care experiences of the entire family. Two calendar years of medical expenditure and utilization data are collected in each household and captured using computer-assisted personal interviews. This series of data collection rounds is launched again each subsequent year on a new sample of households to provide overlapping samples of survey data that provide continuous and current estimates of health care expenditures (Cohen JW, 1997).

The sampling frame for the MEPS-HC is drawn from respondents to the previous year's National Health Interview Survey (NHIS), conducted by NCHS. NHIS provides a nationally representative sample of the U.S. civilian noninstitutionalized population, with over sampling of Hispanics and blacks.

Source of Data

This study is based on six years of use and expenditure data from MEPS (1996-2001). Expenditures in MEPS are defined as the sum of direct payments for health care provided during the year, including out-of-pocket payments and payments by private insurance, Medicare, Medicaid, and other sources. Payments for over the counter drugs, alternative care services, and phone contacts with medical providers are not included in MEPS total expenditure estimates. Indirect payments unrelated to specific medical events such as Medicaid Disproportionate Share and Medicare Direct Medical Education subsidies also are not included (Cohen JW, Machlin SR, Zuvekas SH, et al., 2000).

The use and expenditure data included in this paper were derived from the MEPS-HC and Medical Provider Components (MPC). MPC data were collected for some office-based visits to physicians (or medical providers supervised by physicians), hospital-based events (e.g. inpatient stays, emergency room visits, and outpatient department visits), and prescribed medicines. HC data were collected for physician visits, dental and vision services, other medical equipment and services, and home health care not provided by an agency. Data on expenditures for care provided by home health agencies were collected only in the MPC. MPC data were used if complete; otherwise HC data were used if complete. Missing data for events where HC data were not complete and MPC data were not collected or not complete were derived through an imputation process (Machlin S. and Dougherty D., 2004).

Distribution of MEPS Expenditure Data

MEPS expenditure data, as show in Figure 1, exhibits a marked positive skewness, with a few high expenditure respondents and many low or zero expenditure respondents. Furthermore, this skewness or concentration of medical expenditures has also been shown to be consistent over time. Figure 2 (Berk ML and Monheit AC, 2001), updated with 2001 MEPS data, shows that the concentration of health care expenditures among the U.S. population has remained stable: the top 1% of the population accounts for 25-29% of total expenditures, the bottom 50% of the population accounts for only 3% of total expenditures, and this degree of concentration has been consistent over time.

Confidence Limits by Normal Approximation

In sample surveys, the normal approximation typically is used to calculate confidence limits. For example, 1-α (e.g., 95%) confidence limits are computed for the population mean $\overline{Y}$ by the normal approximation as follows:

Figure 1 - Distribution of Health Expenditures for the U.S. Population, 2001 MEPS

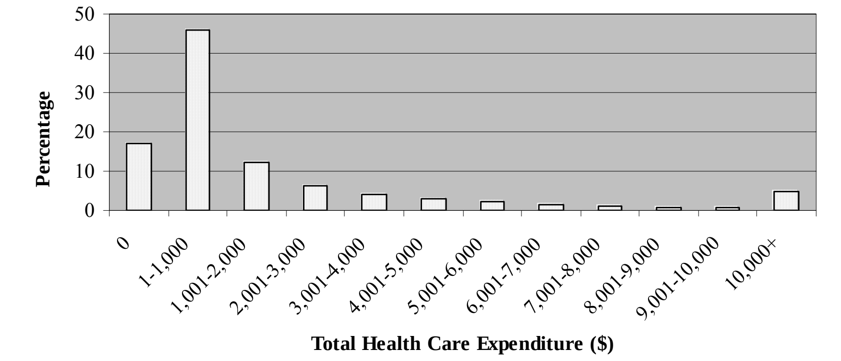
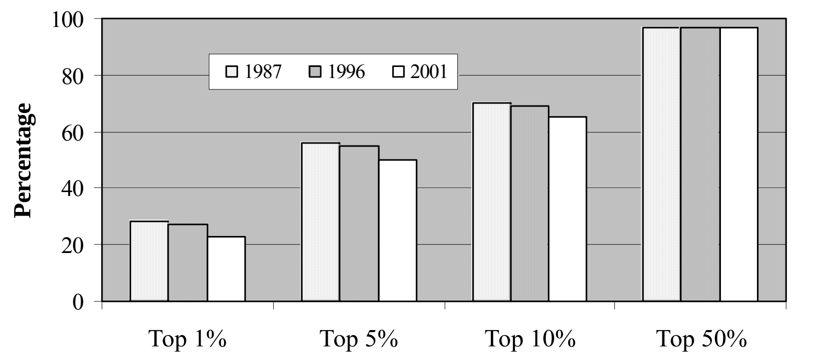


Figure 2 - Concentration of Medical Expenditures



$$\overline{y} - Z_{(1-\alpha/2)} S_{\bar{y}} < \overline{Y} < \overline{y} + Z_{(1-\alpha/2)} S_{\bar{y}} \qquad (1)$$

Another form of the normal approximation to 95% confidence limits for population proportion P is:

$$p \pm \{Z_{(1-\alpha/2)}\sqrt{1-n/N}\sqrt{pq/(n-1)}+\frac{1}{2n}\} \qquad (2)$$

where *q = 1-p, (1 - n/N)* is the finite population correction, and the last term on the right, *1/2n*, is a correction for continuity. With repeated sampling, we claim that statements of this kind will not hold for only 5% of the time. However, for highly skewed data, the probability that the statement above will not hold is often higher than 5% unless the sample size is very large.

Confidence that the normal approximation is adequate in most practical situations comes from a variety of sources (Cochran WG, 1963). It has been shown that for any population which has a finite standard deviation the distribution of the sample mean tends to normality as the sample size increases (Feller W, 1957). For populations in which the principal deviation from normality consists of marked positive skewness, Cochran recommends the following rule on minimum sample size for use of the normal approximation in computing confidence limits:

$$n > 25\ G_1^2 \qquad (3)$$

where G_1 is Fisher's measure of skewness.

$$G_1 = \frac{E(y_i - \overline{Y})^3}{\sigma^3} = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(y_i - \overline{Y})^3$$

(4)

This rule is designed so that 95% confidence limits will not hold for not more than 6% of the time. Application of this rule to compute 95% confidence limits on MEPS total expenditures requires a sample size of ~ 4,000. While annual MEPS sample sizes are substantially larger than 4,000, many of MEPS analytic and policy relevant subpopulations of interest are smaller than 4,000.

Evaluation of Impact of Skewness on Confidence Probability

We designed a simulation study to evaluate coverage error of confidence intervals constructed with the normal approximation method. We constructed a hypothetical population based on five years of MEPS data (1996-2000) with 124,564 records. Four variables listed in Table 1 with a wide range of skewness measures were selected for the study:

Table 1 – Variables Analyzed

Variables analyzed	Hypothetical Population Mean	Hypothetical Population Skewness
Total expenditures	$2,040	16.17
Rx expenditures	$294	9.43

Proportion with inpatient expenses	0.07	3.22
Proportion with dental visits	0.38	0.48

We selected 10,000 repeated samples of varying sizes (25 – 5,000) with replacement from the hypothetical population using a SAS uniform random number generator "ranuni (seed)." For each sample, confidence intervals about the means and proportions were computed based on (1) and (2), respectively at three different levels of α (.01, .02, .05), to determine if they cover the target hypothetical population parameters. The results are presented in Figures 3 to 6 for mean annual total health expenditures, mean annual Rx expenditures, the proportion with inpatient expenses, and the proportion with dental visits, respectively.

As shown in Figure 3, the actual coverage of confidence intervals containing the hypothetical population mean of total health expenditures (skewness=16.17) was far from the stated coverage. For example, at a sample size of 500, the actual coverage was at 95.4%, 93.9%, and 90.6% for the stated coverage of 99% (α=.01), 98% (α=.02), and 95% (α=.05), respectively. The coverage did not get close (within 1%) to the stated coverage until the sample size approached 4,000.

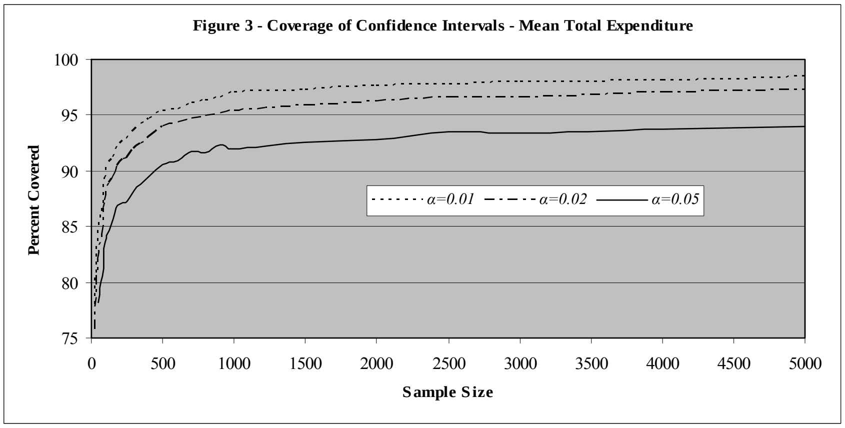


Figure 4 presents the simulation results for mean Rx expenditures which has a moderate skewness measure of 9.43. Compared to the coverages shown in Figure 3, the actual coverage increased to 97.7%, 96.5%, and 93.4% for the stated coverage of 99%, 98%, and 95% respectively and the coverage started to get close (within 1%) to the stated coverage when the sample size approached 1,000.

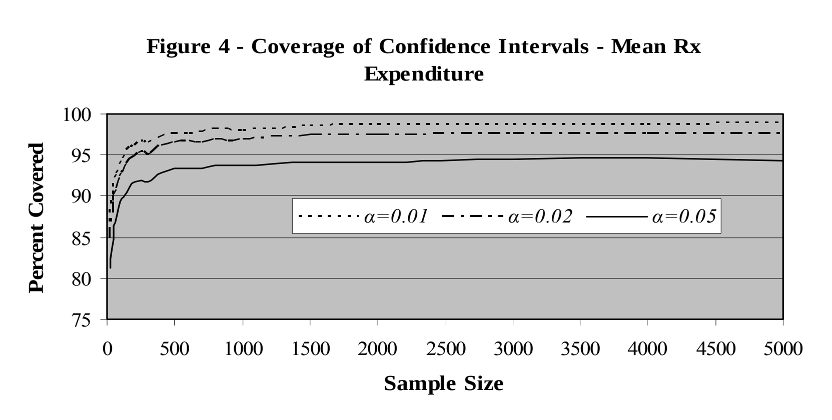


Table 2 – Sample Size Requirements to Approach True 95% Coverage for Expenditure Variables

Sample Size	Coverage of 95% C.I. for Mean Total Expenditures	Coverage of 95% C.I. for Mean Rx Expenditures
25	72.0%	81.2%
100	83.7%	89.3%
500	90.6%	93.4%
1,000	92.0%	93.7%
3,000	93.4%	94.4%
5,000	94.0%	94.3%

Comparing the simulation results for mean total expenditures and mean Rx expenditures, as shown in Figures 3, 4 and Table 2, we conclude that for estimates based on MEPS expenditure data with marked positive skewness, a large sample size (e.g., >1,000 for mean Rx expenditures or > 4,000 for mean total expenditures) may be needed to maintain validity of the normal approximation used to calculate confidence limits.

Simulation results for proportions with inpatient expenses and proportions with dental visits are shown in Figures 5 and 6, respectively. The sample size requirements to achieve the validity of normal approximation for these estimates were much smaller. For proportions with inpatient expenses (skewness=3.22), a sample size of 400 is needed to satisfy the requirement, while a sample size of 75 was sufficient for proportions with dental visits (skewness=0.48).

Table 3 – Sample Size Requirements to Approach True 95% Coverage for Proportions

Sample Size	Coverage of 95% C.I. for Proportion with Inpatient Expenses	Coverage of 95% C.I. for Proportion with Dental Visits
25	85.5%	93.5%
50	89.6%	93.2%
75	92.1%	94.0%
100	94.1%	94.8%
250	92.5%	95.3%
500	94.7%	95.0%

Comparing the simulation results for proportion with inpatient expenses and proportion with dental visits, as shown in Figures 5, 6 and Table 3, we conclude that for confidence intervals on estimates of proportions, sample sizes of about 100 appear sufficient for normal approximation to calculate confidence limits.

Figure 5 - Coverage of Confidence Intervals - Proportion with Inpatient Expenses

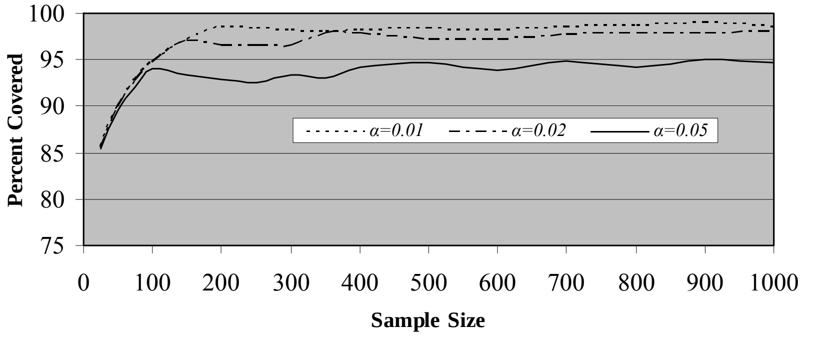
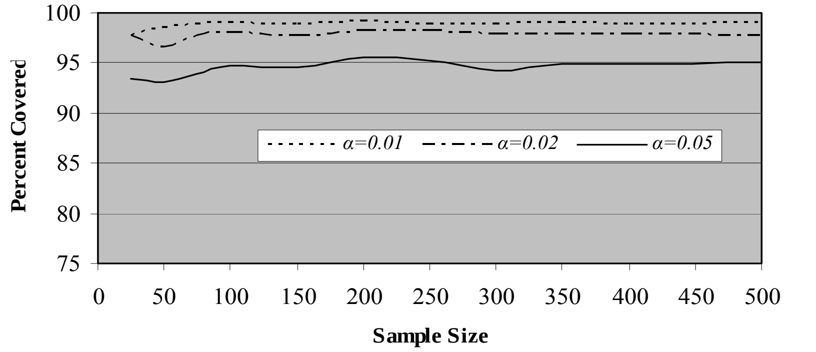


Figure 6 - Coverage of Confidence Intervals - Proportion with Dental Visits



Conclusions

- MEPS expenditure data are highly skewed. The extent of skewness varies by type of expense and population subgroup. This raises questions about the validity of the normal approximation used to compute confidence intervals because confidence levels (e.g., 95%) for intervals based on relatively large samples may be substantially overstated.

- Options to improve normal approximation:

    a. Increase sample size (pool multiple years of data);
    b. Reduce α level when constructing intervals. For example, Figure 3 shows that computing a 98% (α=.02) confidence interval for mean total expenditure with n=900 will result in a true 95% (α=.05) confidence interval.

- Sample sizes of 100 appear generally sufficient for less skewed distributions and normal approximation to binomial.

- Since the analysis is based on repeated simple random samples, results may understate the necessary sample sizes to achieve normality from complex sample data.

- Analysts should consider extent of skewness when interpreting estimates and making inferences based on MEPS expenditure data.

References

Cohen JW, “Design and methods of the Medical Expenditure Panel Survey Household Component.” Rockville (MD): Agency for Health Care Policy and Research; 1997. MEPS Methodology Report No.1. AHCPR Pub. No. 97-0026.

Cohen JW, Machlin SR, Zuvekas SH, *et al.*, “Health care expenses in the United States, 1996.” Rockville (MD): Agency for Healthcare Research and Quality; 2000. MEPS Research Findings 12. AHRQ Pub. No. 01-0009.

Machlin S. and Dougherty D., “Overview of methodology for imputing missing expenditure data in the Medical Expenditure Panel Survey.” 2004 Proceedings of the American Statistical Association, Section on Survey Research Methods, [CD-ROM], Alexandria, VA: American Statistical Association.

Berk ML and Monheit AC, “The concentration of health care expenditures, revisited.” Health Affairs 2001; 20: 9-18.

Feller W, “An introduction to probability theory and its applications.” John Wiley and Sons, New York, 1957; second edition.

Cochran WG, “Sampling Techniques.” John Wiley and Sons, New York, 1963; second edition.

Acknowledgements

The authors wish to thank Trena Ezzati-Rice and Jessica Banthin for their helpful reviews of the paper.