Confidence Intervals for Skewed Healthcare Expenditure Data from the Medical Expenditure Panel Survey (MEPS)

William W. Yu, Agency for Healthcare Research and Quality
540 Gaither Road, Rockville, MD 20850-6649

Key Words: MEPS, medical expenditures, skewness, confidence intervals, Lognormal, Gamma.

1. 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 extremely high expenditure cases and many low or zero expenditure cases. As a consequence of this departure from the normal distribution, confidence levels for conventional normal confidence intervals may be overstated even for relatively large samples. Alternative non-normal distributions (e.g., Lognormal, Gamma) may be appropriate for use to construct confidence intervals for MEPS expenditure data.

Based on repeated sample simulations using data from the 1996 to 2002 MEPS, this paper compares the coverage errors, interval width, and relative symmetry achieved for confidence intervals (CI) constructed under normal distribution and alternative distributional assumptions.

2. 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.

3. Source of Data

This study is based on seven years of expenditure data from MEPS (1996-2002). 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).

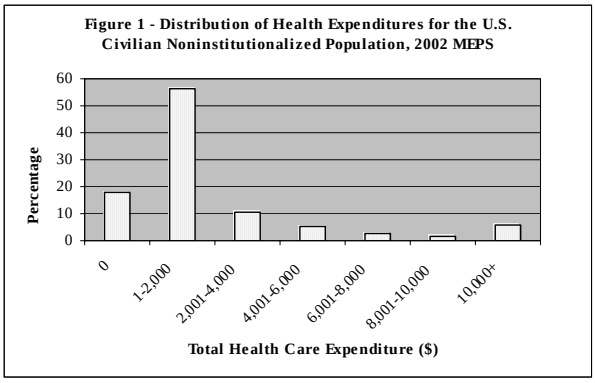
The expenditure data included in this paper were derived from the MEPS-HC and Medical Provider Components (MPC). MPC data were collected for some office-based visits to physicians (or medical providers supervised by physicians), hospital-based events (e.g. inpatient stays, emergency room visits and outpatient 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

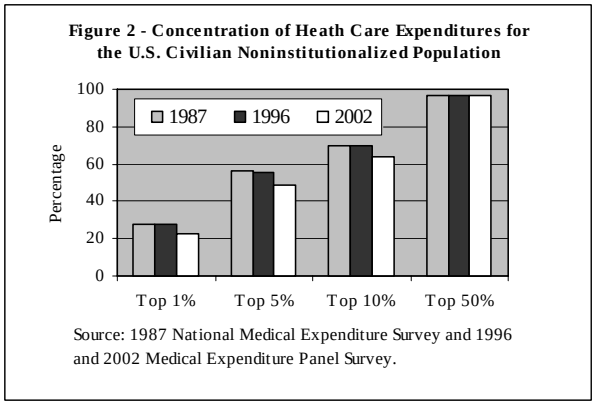
The views expressed in this paper are those of the author and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred. The author wishes to thank Steve Machlin and Joel Cohen for their helpful reviews of the paper.

complete and MPC data were not collected or not complete were derived through an imputation process (Machlin S. and Dougherty D., 2004).

4. Distribution of MEPS Expenditure Data

MEPS expenditure data, as shown in Figure 1, exhibit a marked positive skewness, with a few high expenditure cases and many low or zero expenditure cases. 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 2002 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 22-28% 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 except for a slight drop of concentration for the tail of the distribution in 2002.





5. Confidence Intervals by Normal Approximation

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

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

Another form of the normal approximation to 95% confidence intervals 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 these intervals will not capture the true population parameter 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 extremely large.

Rules for confidence that the normal approximation is adequate in most practical situations come 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 CIs:

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

where G_1 is Fisher's measure of skewness.

$$G_1 = \frac{E(y_i - \overline{Y})^3}{\sigma^3} = \frac{1}{N\sigma^3} \sum_{i=1}^{N} (y_i - \overline{Y})^3 \qquad (4)$$

This rule is designed so that 95% CIs will not contain the population parameter no more than 6% of the time. Application of this rule to compute 95% CIs on MEPS total expenditures requires a sample size of ~ 4,000.

A simulation study based on a hypothetical population with five years of MEPS data (1996-2000) concluded the following (Yu W and Machlin S, 2004):

- For MEPS estimates of proportions (e.g., proportion with inpatient expenses, skewness = 3.22), sample sizes of about 100 appear sufficient to maintain validity of normal approximation used to calculate CIs.

- For MEPS estimates of means (e.g., mean total healthcare expenditures, skewness = 16.17), a large sample size of ~ 4,000 is required to satisfy the requirement.

- 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.

- Probability levels for CIs on some MEPS estimates developed with normal approximation may be overstated.

Alternative non-normal distributions such as gamma and lognormal may be appropriate for use to construct CIs for MEPS expenditure estimates.

6. Gamma Confidence Intervals

If the total expenditures variable, x, has a gamma distribution, Γ(a,b), where a and b are the shape and scale parameters, respectively. Then E(x)=ab and Var(x)=ab^2. It follows that the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is distributed as Γ(na,b/n). Using the relationship between the gamma parameters (a,b) and the sample mean and standard deviation ($\overline{X}$,S), Baskin and Sommers (2005) derived a method of moment estimator, $\hat{a} = \overline{X}^2 / S^2$.

To construct the CI for mean total expenditures, let E(x)=μ and assume $\overline{X}/\mu$ ~ Γ(na,1/na). The following statement

$$P\{\Gamma(\alpha/2; n\hat{a}, 1/n\hat{a}) \le \overline{X}/\mu \le \Gamma(1-\alpha/2; n\hat{a}, 1/n\hat{a})\} \approx 1-\alpha$$

may be rearranged to produce the desired CI:

$$P\{\frac{\overline{X}}{\Gamma(1-\alpha/2; n\hat{a}, 1/n\hat{a})} \le \mu \le \frac{\overline{X}}{\Gamma(\alpha/2; n\hat{a}, 1/n\hat{a})}\} \approx 1-\alpha \quad (5)$$

7. Lognormal Confidence Intervals

If the total expenditures variable, x, has a lognormal distribution with location and scale parameters (μ,σ), then y=log(x) ~ N(μ,σ^2) with

E(x)=exp(μ+½σ^2), Var(x)=exp(2μ+½σ^2(exp(σ^2)-1)).

Since ($\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$) is

complete sufficient for (μ,σ^2), Cox and Land (1972) proposed the following CI for μ+½σ^2:

$$\overline{Y} + \frac{S^2}{2} \pm Z_{1-\alpha/2}\sqrt{\{\frac{S^2}{n} + \frac{S^4}{2(n-1)}\}} \quad (6)$$

which may then be converted into desired CI for exp(μ+½σ^2).

8. Evaluation of Confidence Intervals – (Normal, Gamma, Lognormal)

A simulation study was conducted to evaluate coverage probability, interval width, and relative symmetry achieved for 95% CIs constructed under normal, gamma, and lognormal distributions. The hypothetical population was constructed from seven years of MEPS data (1996-2002) with 194,104 records. Total healthcare expenditures and Rx expenditures were used in the study.

Ten thousand repeated samples of varying sizes ranging from 25 to 5,000 were selected with replacement from the hypothetical population using a SAS uniform random number generator "ranuni (seed)." For each sample, CIs about the means were computed based on (1), (5), and (6), respectively for α = .05, to determine if they cover the target hypothetical population means. The results are presented in simulations 1 - 4 for mean annual total health care expenditures (TOTEXP with and without $0) and mean annual prescribed medicine (Rx) expenditures (RXEXP with and without $0), respectively. TOTEXP and RXEXP were selected to represent variables with high and moderate skewness respectively. Lognormal based intervals were excluded from comparisons in simulations 1 and 3 where $0 expenditures were included.

In addition to coverage probability described above and average interval width, the tables also contain a measure of relative symmetry defined as

$$relative\ symmetry = \frac{|\%CI < pop\_mean - \%CI > pop\_mean|}{\%CI < pop\_mean + \%CI > pop\_mean}$$

where the denominator is the totality of coverage errors

and the numerator is the absolute value of the difference in coverage errors between the percentage of intervals falling below the population mean and the percentage of intervals falling above the population mean (Zhou X and Gao S, 1997.)

Simulation 1 – Total Expenditures (including 0's, normal vs. gamma)

As shown in Figure 3 and Table 1, the coverage of the simulated CIs for both gamma and normal were far from the stated coverage of 95% for sample sizes under 1,000. However, the coverage probability reached 94% (within 1% of the stated probability) at n ≈ 1,000 for gamma based CIs and at n ≈ 2,500 for normal based CIs. In general, gamma based CIs for mean total expenditures had better coverage probability than the normal based CIs. Table 1 also showed that for mean total expenditures, the gamma based CIs were wider at small sample sizes but converged to that of the normal based CIs as sample size increased. The gamma based CIs also were more symmetrical (smaller relative symmetry) in coverage errors than the normal based CIs.

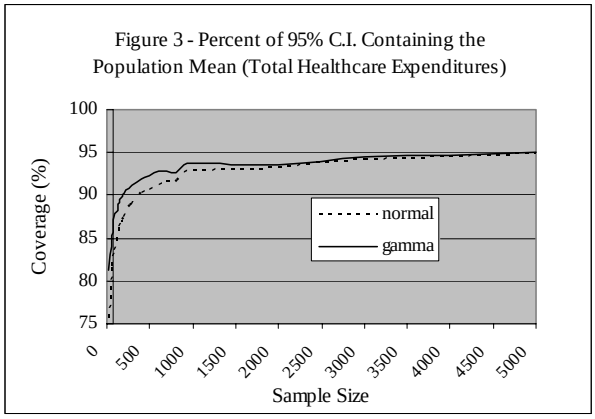


Table 1 – Comparison of simulated 95% CIs for mean total expenditures (including 0's)

{Population: Mean=2217.57, Std=7343, Skewness=14.71}

Sample Size	Assumed Distribution	Coverage Probability	Average Width	Relative Symmetry
25	normal	0.74	3,824	1.00
	gamma	0.81	7,874	0.50
100	normal	0.84	2,408	0.99
	gamma	0.88	3,016	0.55
500	normal	0.91	1,209	0.93
	gamma	0.92	1,264	0.54
1,000	normal	0.93	879	0.84
	gamma	0.94	899	0.43
5,000	normal	0.95	403	0.57
	gamma	0.95	405	0.30

Simulation 2 – Total Expenditures (excluding 0's, normal vs. gamma ns. lognormal)

Excluding cases with total expenditures = 0 from the simulations, figure 4 and table 2 showed that the lognormal based intervals had the best coverage probability for sample sizes < 1,000. However, as the sample size increased, the coverage rate for lognormal based intervals became worse. The gamma based CIs had slightly better coverage probabilities for large sample sizes (> 1,000) than normal based CIs. Table 2 also showed that the gamma based CIs generally had the best (smallest) relative symmetry. The differences in average width between the three alternatives decreased as sample size increased.

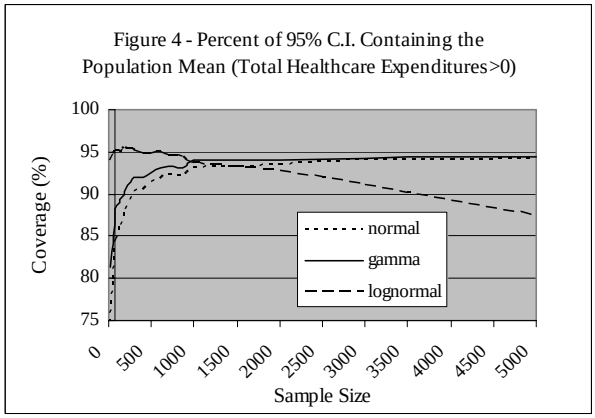


Table 2 – Comparison of simulated 95% CIs for mean total expenditures (excluding 0's)

{Population: Mean=2700.72, Std=8023, Skewness=13.56}

Sample Size	Assumed Distribution	Coverage Probability	Average Width	Relative Symmetry
25	normal	0.75	4,349	1.00
	gamma	0.81	8,039	0.51
	lognormal	0.94	9,347	0.84
100	normal	0.85	2,681	0.99
	gamma	0.89	3,220	0.59
	lognormal	0.95	3,150	0.39
500	normal	0.91	1,325	0.89
	gamma	0.92	1,373	0.47
	lognormal	0.95	1,300	0.40
1,000	normal	0.93	963	0.82
	gamma	0.94	980	0.42
	lognormal	0.94	913	0.61
5,000	normal	0.94	440	0.59
	gamma	0.94	442	0.37
	lognormal	0.88	404	0.94

Simulations 3 and 4 present results based on Rx expenditures data including 0's and excluding 0's, respectively.

Simulation 3 – Rx Expenditures (including 0's, normal vs. gamma)

The simulations based on Rx expenditures (including 0's) presented in figure 5 and table 3 show similar patterns as the ones observed in simulation 1 based on total expenditures.

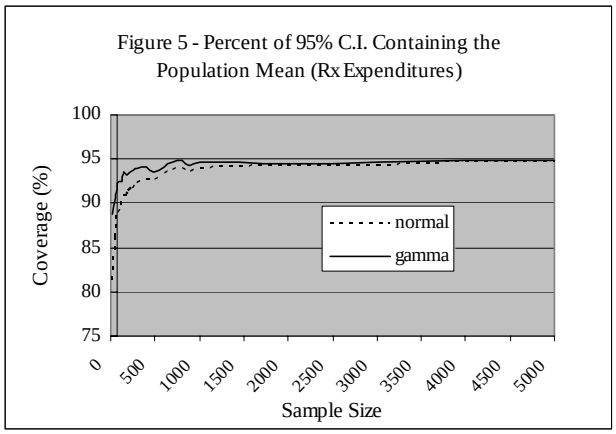


Table 3 – Comparison of simulated 95% CIs for mean Rx expenditures (including 0's)

{Population: Mean=357.08, Std=971.27, Skewness=9.51}

Sample Size	Assumed Distribution	Coverage Probability	Average Width	Relative Symmetry
25	normal	0.81	597	0.98
	gamma	0.89	1,013	0.17
100	normal	0.89	345	0.95
	gamma	0.92	400	0.19
500	normal	0.92	165	0.79
	gamma	0.94	170	0.23
1,000	normal	0.94	118	0.66
	gamma	0.95	120	0.23
5,000	normal	0.95	54	0.42
	gamma	0.95	54	0.19

Simulation 4 – Rx Expenditures (excluding 0's, normal vs. gamma vs. lognormal)

With the exception of a spike at sample size ~ 250, similar patterns observed in simulation 2 based on total expenditures (excluding 0's) were also observed in simulations based on Rx expenditures (excluding 0's) as shown in figure 6 and table 4. The lognormal based intervals had the best coverage probability for sample sizes < 50. However, as the sample size increased, the coverage rate for lognormal based intervals became worse. The gamma based CIs had the best coverage probability for small sample sizes (< 500) but similar as normal based CIs for sample sizes ≥ 500.

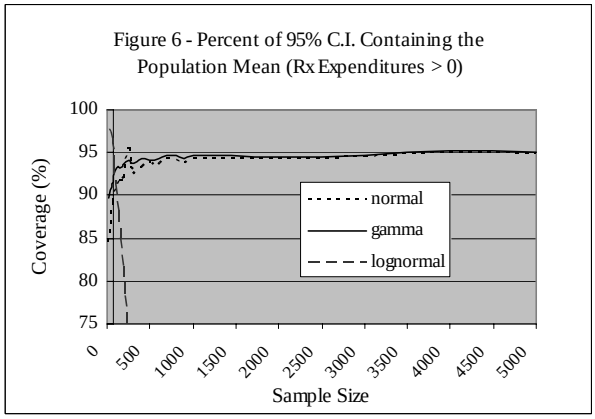


Table 4 – Comparison of simulated 95% CIs for mean Rx expenditures (excluding 0's)

{Population: Mean=582.37, Std=1186, Skewness=7.96}

Sample Size	Assumed Distribution	Coverage Probability	Average Width	Relative Symmetry
25	normal	0.85	775	0.97
	gamma	0.90	1,065	0.21
	lognormal	0.98	2,818	0.20
100	normal	0.90	431	0.89
	gamma	0.93	469	0.29
	lognormal	0.91	972	0.98
500	normal	0.94	202	0.73
	gamma	0.94	206	0.29
	lognormal	0.39	399	1.00
1,000	normal	0.94	145	0.63
	gamma	0.95	146	0.31
	lognormal	0.08	279	1.00
5,000	normal	0.95	65	0.37
	gamma	0.95	66	0.18
	lognormal	0.00	124	1.00

9. Conclusions

- MEPS expenditure data are highly skewed. This raises questions about the validity of the normal approximation used to compute CIs because confidence levels (e.g., 95%) for intervals based even on relatively large samples may be substantially overstated.

- Comparing simulations based on expenditure data including the 0's (normal vs. gamma):
  - the gamma based intervals had the highest coverage probability (for most sample sizes),

  - the differences in average interval width disappeared as sample size increased,
  - the gamma based intervals had the best relative symmetry in coverage.

- Comparing simulations based on expenditure data excluding the 0's (normal vs. gamma vs. lognormal):
  - the lognormal based intervals had the best coverage probability for small sample sizes, however, as the sample size increased, the coverage probability became substantially worse,
  - the differences in average interval width seemed to disappear as sample size increased,
  - the gamma based intervals had the best relative symmetry in coverage.

- Overall, the gamma based intervals appeared to have better coverage probabilities and the best relative symmetry.

- This analysis was based on repeated simple random samples. The effects of stratified multistage sampling which is more similar to MEPS design need to be studied.

10. 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.

Yu W and Machlin S, "Estimation of skewed health expenditure data from the Medical Expenditure Panel Survey (MEPS)." 2004 Proceedings of the American Statistical Association, Section on Survey Research Methods, [CD-ROM], Alexandria, VA: American Statistical Association.

Baskin RM and Sommers JP, "Confidence intervals for skewed data", Agency for Health Care Research and Quality, 2005. Draft report.

Land CE, "An evaluation of approximate confidence interval estimation methods for lognormal means." Technometrics, 1972; 14: 145-158.

Zhou X and Gao S, "Confidence intervals for the log-normal mean." Statistics in Medicine, 1997; 16:783-790.