



U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**



DOT HS 811 217

October 2009

Estimating Median Hospital Charge in the CODES Crash Outcome Data Evaluation System

This publication is distributed by the U.S. Department of Transportation, National Highway Traffic Safety Administration, in the interest of information exchange. The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the Department of Transportation or the National Highway Traffic Safety Administration. The United States Government assumes no liability for its contents or use thereof. If trade or manufacturers' names are mentioned, it is only because they are considered essential to the object of the publication and should not be construed as an endorsement. The United States Government does not endorse products or manufacturers.

1. Report No. DOT HS 811 217		2. Government Accession No.		3. Recipients's Catalog No.	
4. Title and Subtitle Estimating Median Hospital Charge in the CODES Crash Outcome Data Evaluation System			5. Report Date October 2009		
			6. Performing Organization Code NVS-412		
7. Author(s) Alexander Strashny, Ph.D.			8. Performing Organization Report No.		
9. Performing Organization Name and Address State Data Reporting Division, Office of Data Acquisition National Center for Statistics and Analysis National Highway Traffic Safety Administration NVS-412, 1200 New Jersey Avenue SE. Washington, DC 20590			10. Work Unit No. (TRAIS)n code		
			11. Contract of Grant No.		
12. Sponsoring Agency Name and Address State Data Reporting Division, Office of Data Acquisition National Center for Statistics and Analysis National Highway Traffic Safety Administration NVS-412, 1200 New Jersey Avenue SE. Washington, DC 20590			13. Type of Report and Period Covered NHTSA Technical Report		
			14. Sponsoring Agency Code		
15. Supplementary Notes I thank John Kindelberger, Robert Sivinski, Mike Fleming, and Mike McGlincy for their helpful comments.					
16. Abstract The Crash Outcome Data Evaluation System (CODES) is a data system that links crash data with medical and financial outcome data, such as hospital and emergency medical services data. This allows for improved studies of medical and financial outcomes of motor vehicle crashes. Because of the lack of personally identifiable information, crash and outcome data are linked probabilistically using a technique called multiple imputations. One of the variables in CODES is the hospital charge that was charged to each crash victim. This report derives the maximum likelihood estimator of the median hospital charge as well as its confidence interval. The estimator has desirable theoretical properties and performs well in simulations. The estimator is designed for non-negative positively skewed variables, such as cost, charge, price, income, wealth, and so on. It can be used with multiply imputed data, such as the data available in CODES, as well as with non-multiply imputed data.					
17. Key Words Record linkage, median, log-normal distribution, positively skewed, maximum likelihood estimator, Crash Outcome Data Evaluation System, multiple imputations, hospital charges			18. Distribution Statement Document is available to the public through the National Technical Information Service, Springfield, VA 22161		
19. Security Classif. (of this report) Unclassified		20. Security Classif. (of this page) Unclassified		21. No of Pages 20	22. Price

Table of Contents

Executive Summary	1
Introduction	2
Solution	3
Maximum Likelihood Estimator	3
Objections	5
Simulations	6
Demonstration.....	10
Conclusion	11
Appendix: Asymptotic MLE	12
References.....	14

List of Tables

Table 1	Properties of three point estimators.	8
---------	--	---

List of Figures

Figure 1	Distributions of the population used in the simulation.	7
Figure 2	The distributions of the following estimators of the median: MLE, asymptotic MLE, and sample median.....	8
Figure 3	The distributions of the lower and upper 95% confidence limits of the MLE. ..	9
Figure 4	The distributions of the lower and upper 95% confidence limits of the asymptotic MLE.....	10

Executive Summary

Objective. The Crash Outcome Data Evaluation System (CODES) is a data system facilitated by the National Highway Safety Traffic Administration that links State-level crash data with medical and financial outcome data, such as hospital and emergency medical services data. This allows for improved studies of medical and financial outcomes of motor vehicle crashes. Because of the lack of personally identifiable information, crash and outcome data are linked probabilistically using a technique called multiple imputations.

One of the variables often studied in CODES analyses is the hospital charge that was charged to each crash victim. The objective of this report is to develop a statistical technique for estimating the median of the hospital charge and other similar variables, as well as the confidence interval for the median.

Method. This report derives the maximum likelihood estimator (MLE) of the population median and its confidence interval. Note that the population median is the 50th percentile of the underlying probability distribution as opposed to the sample median, which is the 50th percentile of the observed data. The estimator is designed for non-negative positively skewed variables, such as cost, charge, price, income, wealth, and so on. It can be used with multiply imputed data, such as the data available in CODES, as well as with non-multiply imputed data, such as the data found in many other data systems. The report also derives the asymptotic MLE of the median and its confidence interval, though the MLE is preferred.

The report performs simulations to ascertain the performance of the above mentioned estimators. It also demonstrates the use of the estimators using Utah CODES data from 2003 to 2005.

Results. The derived estimators of the median and its confidence interval have desirable theoretical properties. For example, they minimize the root mean squared error, a common criterion used to define the “best” estimator. In simulations, the MLE performed well, outperforming even the sample median.

Conclusion. This report has developed a useful estimator of the median of hospital charges. The estimator gives both a point estimate as well as a confidence interval. It performs well in simulations. Notably, in simulations, the calculated confidence interval contained the true median with the appropriate probability. In application to real data, the estimator gives a result that may be very different from the sample median. This indicates that using the sample median to estimate the median may produce a large error.

Introduction

In many applications, there is the need to estimate the median of a non-negative positively skewed quantity, such as cost, charge, price, income, wealth, and so on. For example, the users of CODES health data have an interest in estimating the median charge for a hospital stay. The application to CODES is further complicated by the fact that CODES data consists of several multiply imputed datasets.

CODES is a data system facilitated by NHTSA that links State-level crash data with medical and financial outcome data, such as hospital and emergency medical services data. This allows for improved studies of medical and financial outcomes of motor vehicle crashes. Because of the lack of personally identifiable information, crash and outcome data are linked probabilistically using a technique called multiple imputations (McGlinchy, 2004).

CODES evolved as the result of the Intermodal Surface Transportation Efficiency Act of 1991, which provided funds to NHTSA to report to Congress on the benefits of seat belts and motorcycle helmets for people involved in motor vehicle crashes. To measure benefits in terms of reducing death, disability, and medical costs, NHTSA created CODES, a State-level data system that includes all people involved in police-reported crashes. The CODES report was delivered to Congress in February 1996 (NHTSA, 1996).

One of the variables often studied in CODES analyses is the hospital charge that was charged to each crash victim. The objective of this report is to develop a statistical technique for estimating the median of the hospital charge and other similar variables, as well as the confidence interval for the median.

The paper derives a useful estimator of the median of such a quantity along with its confidence interval. The estimator can be used both with multiply imputed data and with data that has not been multiply imputed. In other words, the estimator can be used with databases which do not consist of several multiply imputed datasets. The paper performs simulations to illustrate the performance of the estimator and demonstrates the use of the estimator on real data.

Various measures of central tendency of hospital charge, including the mean, the median, and the mode, could be interesting. It is not that the mean is “too high” and the median is “more stable”. Different measures of central tendency measure different things. The mean is useful in estimating the total charges. The mode is the most likely charge. Finally, the median is the charge paid by a “typical” person. These and other measures of central tendency could be interesting. This paper only addresses the estimation of the median.

Solution

Maximum Likelihood Estimator

Let's call the random variable that of interest X . For example, this could be the charge for a hospital stay. Suppose there are some observations of X , denoted by x_{ij} , where $j = 1, 2, \dots, k$ indexes the multiply imputed dataset, while $i = 1, 2, \dots, n$ indexes the observation within each dataset. Note that in CODES, $k = 5$. When dealing with data that is not multiply imputed, $k = 1$. Let m be the median of X .

The sample median is the method of moments estimator of the median. However, maximum likelihood estimators (MLEs) are better than method of moments estimators. MLEs are often considered the "best" estimators since they minimize the root mean squared error (RMSE). This section derives the MLE of m , denoted by \hat{m} , along with its confidence interval.

Let's assume that X follows the log-normal distribution. This assumption is equivalent to assuming that the logarithm of X follows the normal distribution. It is a typical assumption for variables such as cost, charge, price, income, wealth, and so on. The distributional assumption makes sense if the random variable X is non-negative and positively skewed, as is the case with hospital charges.

Assume further that, within each dataset j , X_i are independent and identically distributed. This means that all X follow the same probability distribution and that, within each dataset, they are independent of each other. Across the datasets j , X is assumed to be identically distributed but not necessarily independent. This reflects the fact that observations across multiply imputed datasets are related to each other.

Let $Y = \ln X$. Y has the normal distribution with parameters μ and σ^2 . Thus, the MLE of μ conditional on dataset j is

$$\hat{\mu}_j = \frac{1}{n} \sum_i y_{ij} = \frac{1}{n} \sum_i \ln x_{ij} \quad (1).$$

Conditional on data in all k datasets, the MLE of μ is

$$\hat{\mu} = \frac{1}{nk} \sum_j \sum_i \ln x_{ij} \quad (2).$$

Recall that MLEs have the property of asymptotic normality. This means that, as the amount of data approaches infinity, the distribution of an MLE approaches a normal distribution. However, in this particular case, since the logarithm of X is itself normally distributed, the above estimators of μ are also normally distributed even in small samples.

The MLE of the variance of Y conditional on dataset j is:

$$\hat{\sigma}_j^2 = \frac{1}{n} \sum_i (\ln x_{ij} - \hat{\mu}_j)^2 \quad (3).$$

Thus, the MLE of the variance of the estimator of μ conditional on dataset j is:

$$\hat{s}_j^2 = \frac{\hat{\sigma}_j^2}{n} = n^{-2} \sum_i (\ln x_{ij} - \hat{\mu}_j)^2 \quad (4).$$

Applying the law of total variance gives the estimator of the variance of the estimator of μ conditional on all the data:

$$\hat{s}^2 = \frac{1}{k} \left(\sum_j \hat{s}_j^2 + \sum_j (\hat{\mu}_j - \hat{\mu})^2 \right) \quad (5).$$

The above expression for variance takes into account both variance within each multiply imputed dataset and variance across the multiply imputed datasets. Note that the law of total variance does not assume independence, so there is no issue in applying it to multiply imputed data.

Since Y is normally distributed, so is the estimator of μ in equation 2. Thus, the 95% confidence interval for μ is the interval between $\hat{\xi}_Y(0.025)$ and $\hat{\xi}_Y(0.975)$, where $\hat{\xi}_Y(p)$ is the MLE of the p -th quantile of the estimator of μ :

$$\hat{\xi}_Y(p) = \hat{\mu} + \Phi^{-1}(p) \sqrt{\hat{s}^2} \quad (6),$$

where $\Phi^{-1}(p)$ is the standard normal quantile function. Again, since the estimator of μ is normally distributed even in small samples, this expression gives an exact confidence interval.

Note that the median of X is given by

$$m = \exp(\mu) \quad (7).^1$$

Also recall that MLEs have the property of functional invariance. This means that a function of an MLE is itself the MLE of the function. Symbolically, let θ be a random variable, let $f(\theta)$ be a function of that variable, and let hats denote MLEs. Then, the following equality is true:

$$\hat{f}(\theta) = f(\hat{\theta}) \quad (8).$$

Thus, the maximum likelihood estimator of the median of X is

¹ The median of a log-normal distribution is $\exp(\mu)$, while the expected value is $\exp(\mu + 0.5\sigma^2)$.

$$\hat{m} = \exp\left(\frac{1}{nk} \sum_j \sum_i \ln x_{ij}\right) \quad (9).$$

Not only does the above estimator minimize the RMSE, but, as the simulations below show, it is unbiased. The MLE of the quantiles of the estimator of the median is

$$\hat{\xi}(p) = \exp\left(\hat{\mu} + \Phi^{-1}(p)\sqrt{\hat{s}^2}\right) \quad (10).$$

The 95% confidence interval for estimator of the median of X is just the interval between $\hat{\xi}(0.025)$ and $\hat{\xi}(0.975)$.

Note that the derivation of the above estimators of the median and its confidence interval is not based on asymptotic arguments. They are the best estimators, in the sense that they minimize the root mean squared error, even for small samples.

In case one is dealing with data that is not multiply imputed, the above estimator simplifies by setting the number of imputations to $k = 1$. Thus, with non-multiply imputed data, the estimator becomes

$$\hat{\mu} = \frac{1}{n} \sum_i \ln x_i \quad (11)$$

$$\hat{m} = \exp(\hat{\mu}) \quad (12)$$

$$\hat{s}^2 = n^{-2} \sum_i (\ln x_i - \hat{\mu})^2 \quad (13)$$

$$\hat{\xi}(p) = \exp\left(\hat{\mu} + \Phi^{-1}(p)\sqrt{\hat{s}^2}\right) \quad (14).$$

Objections

Several objections may be raised against the MLE in equations 9 and 10.

Asymmetric confidence interval. One objection is that the confidence interval is not in the usual form of a point estimator plus or minus some multiple of the standard error. The response is that there is no reason that a confidence interval should take this usual form. What's more, in this particular case, it is good that it doesn't. This is because X , and thus its median, are non-negative. However, a confidence interval in the usual form might include negative values, which would be nonsensical. Since the exponential function is non-negative, the confidence interval as given above never includes negative values.

Unusual form of the estimator. Another objection is as follows. Schafer (1997, pp. 108-110), citing Rubin (1987), gives a procedure for combining estimators based on multiply imputed data. However, the MLE given above does not use this procedure.

The procedure, described in more detail in the appendix, can only be applied to estimators that are normally distributed. Thus, since MLEs are asymptotically normal, the procedure can be applied to MLEs, but only when the sample size approaches infinity. When this is the case, the procedure gives the asymptotic MLE. When the sample size is large, the asymptotic MLE is equivalent to, though it might be numerically slightly different from, the MLE. In other words, it is an approximation of the MLE that is valid only for large sample sizes. Since the MLE has been found, there is no reason to use an approximation of the MLE. See the appendix for more details. Also see the discussion of simulated estimator performance, below.

Potential bias due to record linkage. Yet another objection to the MLE is that, due to the record linkage procedure employed by CODES, the hospital charges, as imputed in the k datasets, might not be representative of the hospital charges in the population. Since the data that is used is skewed, the estimator presented above is not an accurate estimator of the population median.

It is true that the MLE estimator assumes, as do most statistical estimators, that the available data is representative of the population. We are not aware of any studies that demonstrate the presence of a skew in CODES linked data; nor of studies that demonstrate that such a skew does not exist. Thus, further investigation is certainly necessary.

While work on correcting estimators based on linked data exists (such as Larsen, 2004, and Lahiri & Larsen, 2000), implementing these corrections requires knowledge of linkage probabilities. Users of CODES data, who are the ones interested in estimating median hospital charges, typically do not know these probabilities. Properly correcting the estimator would have to be done during the linkage process itself. This is one possible direction for future work.

Finally, whether or not linked data is skewed, it would be skewed even more if only the high probability links are taken, while the lower probability links are discarded. The proper procedure used in CODES record linkage uses links of all probabilities.

Simulations

To illustrate the performance of the estimators derived above, the following simulations were performed.

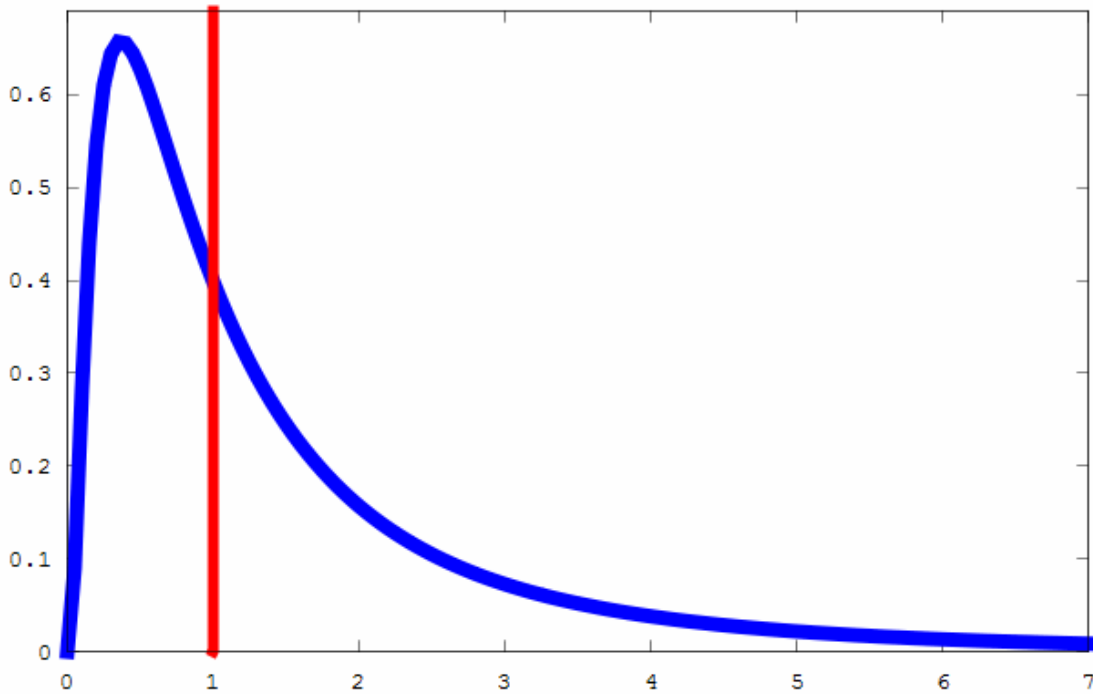
The simulated data was drawn from the log-normal distribution with the parameters of $\mu = 0$ and $\sigma^2 = 1$. Thus, the true median of the data was $m = \exp(\mu) = 1$. The expected value was $\exp(\mu + 0.5\sigma^2) \approx 1.6487$ while the mode was $\exp(\mu - \sigma^2) \approx 0.3679$.

The data was simulated such that it was independent across observations i and correlated across datasets j . In this way, the simulated data is similar to the multiply imputed data available in CODES. In each simulation, there were $k = 5$ datasets, which matches the number of multiply imputed datasets used in CODES. For each of the k datasets, $n = 1000$ observations were used. The number of observations available in CODES is typically much larger. Note that the more

observations, the better the estimator performance. A total of 10,000 simulations were performed.

Figure 1 shows the distribution of the population that was used in the simulation. As expected, the population is positively skewed. The population median of 1 is shown by a vertical line. As mentioned above, the mode is less than the median which is less than the mean. In the figure, the mode is the maximum of the distribution.

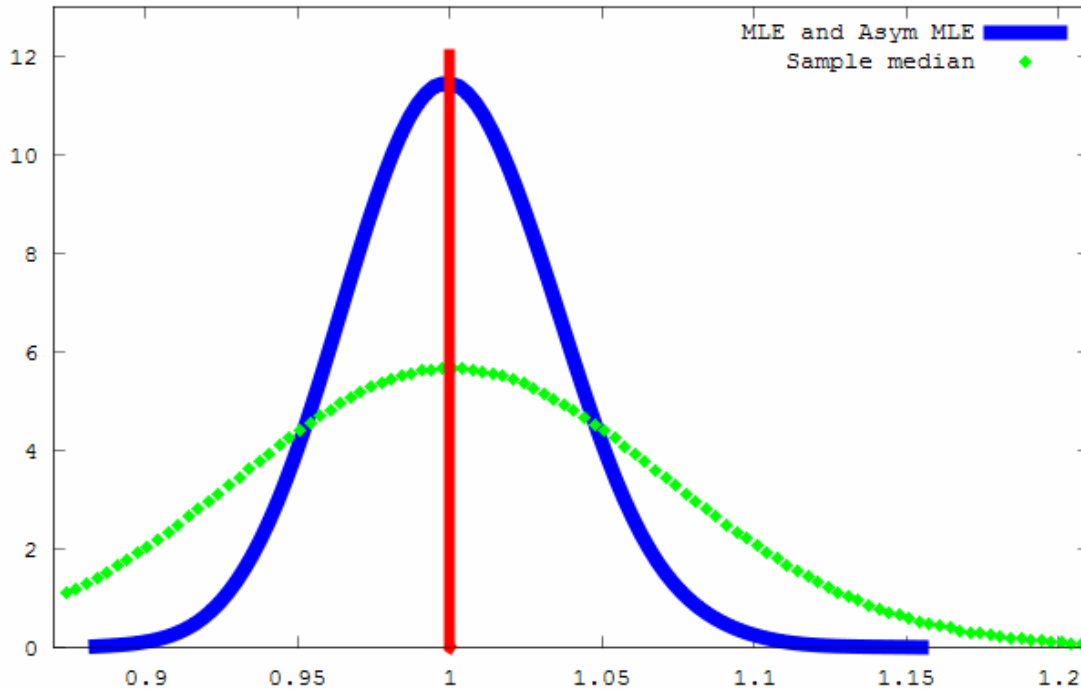
Figure 1 Distributions of the population used in the simulation



Blue line = distribution of the population.
Red vertical line = true median.

Figure 2 shows the distribution of the MLE (equation 9) and the asymptotic MLE (equation 22) of the median. The figure also shows the distribution of the sample median.

Figure 2 The distributions of the following estimators of the median: MLE, asymptotic MLE, and sample median



Solid blue line = MLE and asymptotic MLE of the median.

Dotted green line = sample median.

Solid red vertical line = true median.

The distributions of the MLE and the asymptotic MLE are virtually identical, so much so that they cannot be distinguished on the graph. Both distributions are centered around the true median of 1. The distribution of the sample median is also centered around the true median, though it is noticeably more spread out, indicating that using the sample median produces more error than either the MLE or the asymptotic MLE.

For each of the three estimators, Table 1 gives the root mean squared error (RMSE), bias, and standard error.

Table 1 Properties of three point estimators

Estimator	RMSE	Bias	Std Err
MLE	0.03208	0.00101	0.03207
Asym MLE	0.03208	0.00103	0.03207
Sample median	0.03745	0.00120	0.03743

RMSE is a measure of the typical distance between an estimator and the true parameter value. A common criterion for choosing the best estimator is to pick the one that minimizes RMSE. Maximum likelihood estimators minimize RMSE. Asymptotic MLEs minimize RMSE as the sample size approaches infinity.

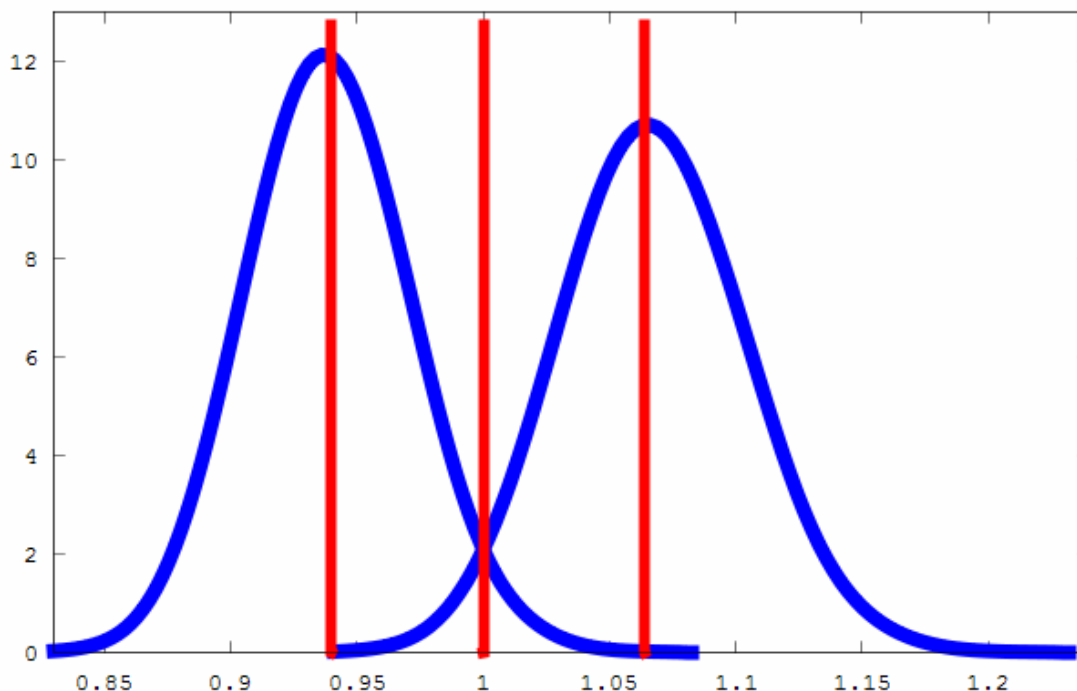
As Table 1 shows, in this simulation, the RMSE for both the MLE and the asymptotic MLE was about 0.03208. The RMSE was slightly lower for the MLE. This is consistent with the property that MLEs minimize RMSE. However, practically speaking, the two RMSEs were equal. The RMSE for the sample median was 0.03745, indicating that using the sample median to estimate the population median produced more error than using either the MLE or the asymptotic MLE.

Bias is the difference between the average estimate and 1, the true value of the median. In this particular simulation, the bias for all three estimators was about 0.1%, indicating that all three estimators are unbiased. Interestingly enough, though all biases were very small and about equal, MLE had the lowest bias.

Standard error measures the spread of the estimates. Again, standard error of MLE and asymptotic MLE were basically equal, though that of the MLE was slightly lower. The standard error of the sample median was higher.

Figure 3 shows the distributions of the 95% confidence limits of the MLE. Vertical lines indicate the population median and the confidence limits calculated by substituting population parameters into the equations.

Figure 3 The distributions of the lower and upper 95% confidence limits of the MLE



Blue lines = distributions of the lower and upper confidence limits.

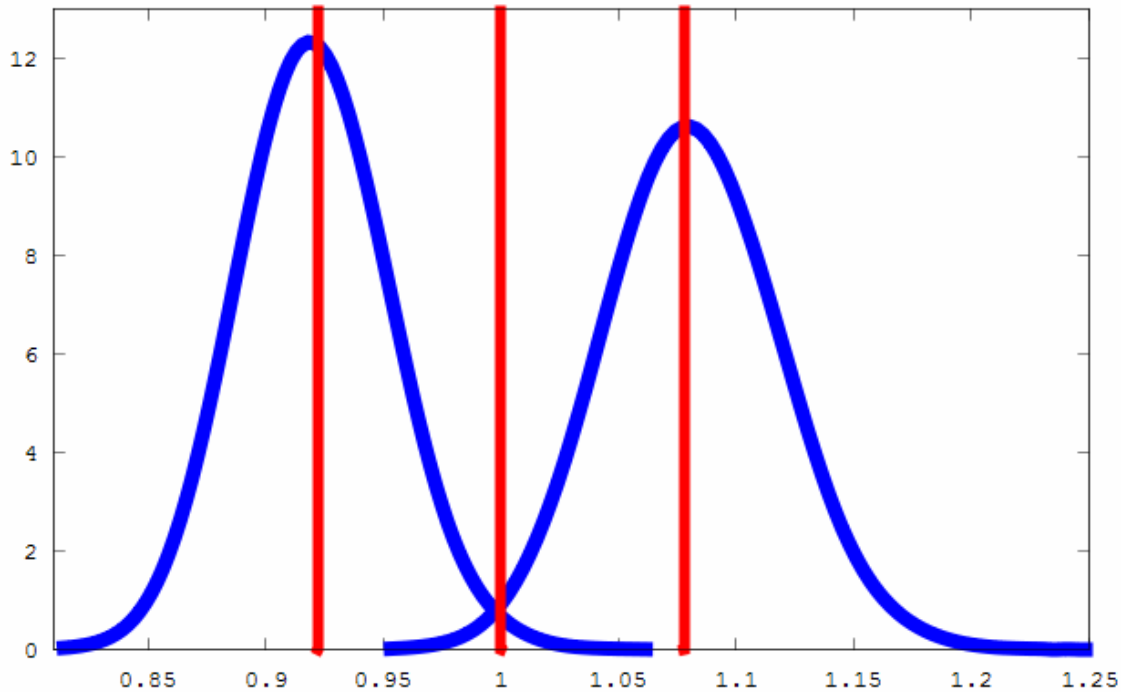
Red vertical lines = true confidence limits and the true median.

As expected, the MLE 95% confidence interval contained the true value of the median approximately 95% of the time. The distributions of the confidence limits centered around the true confidence limits of 0.940 and 1.064. Again, note that the MLE confidence interval is

asymmetric. The true interval was about 0.60 units below the true median and 0.64 units above the true median.

Figure 4 shows the distributions of the 95% confidence limits of the asymptotic MLE. The confidence interval was wider than the MLE confidence interval. The true interval ranged from 0.922 to 1.078. Though this is the 95% interval, it included the true value of the median about 99% of the time. Thus, in simulations, the asymptotic MLE confidence interval was wider than it should have been. The interval was symmetric, extending about 0.078 units to either side of the median.

Figure 4 The distributions of the lower and upper 95% confidence limits of the asymptotic MLE



Blue lines = distributions of the lower and upper confidence limits.

Red vertical lines = true confidence limits and the true median.

Demonstration

As a demonstration, the estimators derived above are applied to Utah CODES data for the years 2003 to 2005. Crash data was probabilistically linked to hospital data using $k = 5$ imputations.

Hospital charges were then adjusted for inflation to 2008 dollars. Of all the Consumer Price Indices (CPI) available, the most appropriate one for adjusting hospital charges for inflation is called Hospital Services (not to be confused with Hospital and Related Services) (Katz & Crawford, 2009). This is the CPI index that is used here. Missing hospital charges were then imputed using 1 imputation, thus preserving the total number of imputations at $k = 5$.

Note that charges were first adjusted for inflation and then imputed. This is the correct order, since it only makes sense to impute a variable if all of its values are in the same units. Adjusting for inflation causes all the values to be in the units of 2008 dollars.

The sample median was \$0.00. The estimates of the median, both using the MLE method and the asymptotic MLE method was \$0.01. The median was zero because about 86% of the people in the dataset were unlinked to a hospital record. Because of this, the imputation procedure for hospital charges assigned zero charges to them.

Next, let's consider only the people who were linked to a hospital record. The median estimates were as follows:

- MLE
 - Median: \$1049
 - 95% confidence interval: \$1036 to \$1062
- Asymptotic MLE
 - Median: \$1049
 - 95% confidence interval: \$1032 to \$1065

The sample median was \$791.

The sample median was much lower than the estimate of the population median. Thus, using sample median as an estimate of the population median would have produced a large error.

The confidence interval for the MLE was slightly asymmetric. The lower limit was \$12.92 below the point estimate, while the upper limit was \$13.08 above it. As discussed above, an asymmetric confidence interval for the median of hospital charges makes sense.

The confidence interval for the asymptotic MLE was wider than for MLE. This is consistent with the simulations discussed in the previous section.

Conclusion

This report has developed a useful estimator of the median of hospital charges. The estimator gives both a point estimate as well as a confidence interval. It performs well in simulations. Notably, in simulations, the calculated confidence interval contained the true median with the appropriate probability. In application to real data, the estimator gives a result that's very different from the sample median. This indicates that using the sample median to estimate the median would produce a large error.

Note again that when working with hospital charges, they should be adjusted for inflation before being imputed.

Appendix: Asymptotic MLE

This appendix details how to calculate the asymptotic MLE of the median. Since the asymptotic MLE is an approximation of the MLE that is only valid for large sample sizes, and since the actual MLE of the median has been derived above in equations 9 and 10, there is no need to use this asymptotic MLE in practice.

Schafer (1997, pp. 108-110), citing Rubin (1987), gives a procedure for combining estimators based on multiply imputed data. The procedure can only be applied to estimators that are normally distributed. Thus, since MLEs are asymptotically normal, the procedure can be applied to MLEs, but only when the sample size approaches infinity. In general, the procedure cannot be applied to MLEs when the sample size is small.

The procedure is as follows. Let θ be the parameter of interest. Let $\hat{\theta}_j$ be a normally distributed estimator of θ conditional on dataset j . Then, the point estimator of θ conditional on all the data is

$$\tilde{\theta} = \frac{1}{k} \sum_j \hat{\theta}_j \quad (15).$$

Note the use of a tilde to indicate an asymptotic MLE.

Likewise, let \hat{u}_j^2 be the estimated variance of $\hat{\theta}_j$. Then the estimated variance of $\tilde{\theta}$ is \tilde{u}^2 , where

$$\tilde{u}_w^2 = \frac{1}{k} \sum_j \hat{u}_j^2 \quad (16)$$

$$\tilde{u}_B^2 = \frac{k+1}{k(k-1)} \sum_j (\hat{\theta}_j - \tilde{\theta})^2 \quad (17)$$

$$\tilde{u}^2 = \tilde{u}_w^2 + \tilde{u}_B^2 \quad (18).$$

Finally, the $100(1-\alpha)\%$ confidence interval for θ under this procedure is

$$\nu = (k-1) \left[1 + \frac{\tilde{u}_w^2}{\tilde{u}_B^2} \right]^2 \quad (19)$$

$$CI = \tilde{\theta} \pm t(\nu, 1-0.5\alpha) \sqrt{\tilde{u}^2} \quad (20),$$

where $t(\nu, p)$ is the Student's t quantile function with ν degrees of freedom.

One objection to the estimator in equations 9 and 10 is that it is not of the form given in equations 15 and 20. The response is that the derivation of the MLE in equations 9 and 10 did

not rely on asymptotic arguments. Thus, that is the best estimator, even in small samples. The estimator in equations 15 and 20 is equivalent, but only as the sample size approaches infinity. In other words, it's an asymptotic MLE.

Though the MLE in equations 9 and 10 should be used to estimate the median, the asymptotic MLE based on equations 15 and 20 is given below, for reference.

Combining equations 1 and 7 gives the MLE of the median of X conditional on dataset j :

$$\hat{m}_j = \exp\left(\frac{1}{n} \sum_i \ln x_{ij}\right) \quad (21).$$

Note that the estimator of the median is clearly not normally distributed. To see this, consider the case when $n = 1$. In that case, the estimator has the same distribution as X , which is log-normal. However, since the estimator is an MLE, its distribution approaches the normal distribution as the sample size n approaches infinity.

Based on this and on equation 15, the asymptotic MLE of the median is

$$\tilde{m} = \frac{1}{k} \sum_j \exp\left(\frac{1}{n} \sum_i \ln x_{ij}\right) \quad (22).$$

According to David (1970), the following is an asymptotic approximation of the variance of the median

$$u^2 = \frac{1}{4nf^2(m)} \quad (23),$$

where $f(x)$ is the probability density function (pdf) of X . Recall that the pdf of the log-normal distribution is

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \quad (24).$$

Substituting equations 7 and 24 into equation 23 gives the asymptotic variance of the median of the log-normal distribution:

$$u^2 = 0.5 \frac{\pi}{n} \exp(2\mu) \sigma^2 \quad (25).$$

Substituting equation 3 into the above gives the asymptotic MLE of the variance of the estimator of median conditional on dataset j :

$$\tilde{u}_j^2 = 0.5 \frac{\pi}{n^2} \exp(2\hat{\mu}_j) \sum_i (\ln x_{ij} - \hat{\mu}_j)^2 \quad (26).$$

This is used to construct the confidence interval for the median as shown in equation 20 and preceding equations.

References

- David, H. A. (1970). *Order Statistics*. New York: John Wiley & Sons.
- Katz, S., and Crawford, M. (2009). CPI Detailed Report – Data for April 2009. Washington, DC: Bureau of Labor Statistics.
- Lahiri, P., and Larsen, M. D. (2000). “Model-based analysis of records linked using mixture models.” *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Larsen, M. D. (2004). “Record Linkage Using Mixture Models.” In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. New York: John Wiley & Sons.
- McGlinchy, M. H. (2004). “A Bayesian Record Linkage Methodology for Multiple Imputation of Missing Links.” *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- NHTSA. (1996). Benefits of Seat belts and Motorcycle Helmets. Report to Congress. February 1996. Based on Data from The Crash Outcome Data Evaluation System (CODES). Washington, DC: National Highway Traffic Safety Administration.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.

DOT HS 811 217
October 2009



U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

