CS-BIGS 6(2): 17-35

http://www.csbigs.fr

# Medicare charges and payments : data analysis, Benford's Law and imputation of missing data

**John Quinn**
*Bryant University, Smithfield, RI, USA*

**Phyllis Schumacher**
*Bryant University, Smithfield, RI, USA*

**Alan Olinsky**
*Bryant University, Smithfield, RI, USA*

*Covered charges (Charges) and Medicare reimbursements (Payments) for various diagnoses were recently made available on the Web. This paper provides an analysis of the relationship between these two variables. In addition, we investigate the conformity of the variables to Benford's Law and its generalizations which are a set of rules that describe the distribution of digits in numbers for many types of data sets. Benford's Law is often used to help detect data that are fraudulently created. We also investigate whether adding a subset of randomized values will alter its distribution of digits as analyzed through various Benford tests.*

Keywords *: Healthcare Data, Medicare Data, Benford's Law, Missing Data, Imputation.*

## 1. Introduction

The Centers for Medicare and Medicaid Services (CMS) have recently made datasets publicly available for analysis and research. The dataset that is utilized in this paper is "Medicare Provider Charge Data: Inpatient" (Centers for Medicare and Medicaid Services, 2013). Our goal in this research, which is an extension of an earlier paper (Quinn et al., 2014), is to analyze two specific variables within this dataset. These two variables are Average Covered Charges (Charges) and Average Total Payments (Payments). More specifically, the charges consisted of "hospital-specific charges for the more than 3,000 U.S. hospitals that receive Medicare Inpatient Prospective Payment System (IPPS) payments for the top 100 most frequently billed discharges, paid under Medicare based on a rate per discharge using the Medicare Severity Diagnosis Related Group (MS-DRG) for Fiscal Year (FY) 2011. These DRGs represent almost 7 million discharges or

60 percent of total Medicare IPPS discharges. Hospitals determine what they will charge for items and services provided to patients and these charges are the amount the hospital bills for an item or service." Meanwhile, the payments are the "average of Medicare payments to the provider for the DRG including the DRG amount, teaching, disproportionate share, capital, and outlier payments for all cases. Also included in Total Payments are co-payment and deductible amounts that the patient is responsible for and payments by third parties for coordination of benefits." There are 163,065 observations for each variable.

The initial phase of the research is to perform a data analysis of the variables Charges and Payments, and then to investigate the relationship of these two variables with respect to multiple other variables in the dataset, such as geographic region (State, City), diagnosis (DRG) and provider centers (Providers). Next, we consider the distributions of the two principal variables with respect to Benford's Law, a set of rules that apply to a variety of data and are often used to confirm the authenticity of the values. The final portion of the research is to investigate the potential impact of missing data and how it affects the conformance to Benford's Law.

## 2. Data

The data are represented below by two dotplots (Figures 1 and 2). These graphs were generated using Minitab Software (Minitab Inc, 2010). The distributions of the variables Charges and Payments are both positively skewed, but more skewed for Charges, which can be observed in these two graphs. There are also two frequency tables (Tables 1 and 2), one for Charges and one for Payments, to provide the counts represented by the dotplots. The class intervals are different between the two tables since in general, the payments are much smaller than the charges.

**Table 1:** Frequency Distribution for the Charges.

| Dollar Range | Charges |
|---|---|
| 0-99999 | 155152 |
| 100000-199999 | 6795 |
| 200000-299999 | 863 |
| 300000-399999 | 185 |
| 400000-499999 | 40 |
| 500000-599999 | 24 |
| 600000-699999 | 4 |
| 700000-799999 | 0 |
| 800000-899999 | 0 |
| 900000-999999 | 2 |
| Total | 163065 |

**Table 2:** Frequency Distribution for the Payments.

| Dollar Range | Payments |
|---|---|
| 0-9999 | 112164 |
| 10000-19999 | 39130 |
| 20000-29999 | 6521 |
| 30000-39999 | 3232 |
| 40000-49999 | 1239 |
| 50000-59999 | 458 |
| 60000-69999 | 200 |
| 70000-79999 | 72 |
| 80000-89999 | 29 |
| 90000-99999 | 11 |
| at least 100000 | 9 |
| Total | 163065 |

As expected, Charges, with a mean value of \$36,134 (s.d. = \$35,065), has a much higher average than Payments with a mean of \$9,707 (s.d. = \$7,665). The magnitude of the mean difference between the two variables was a little surprising, though it was expected that in general Medicare reimbursements would only cover a portion of the provider charges. In fact, the minimum value of Charges is \$2,459 and the maximum is \$929,119 whereas the variable Payments had a minimum of \$2,673 and a maximum of only \$156,158.

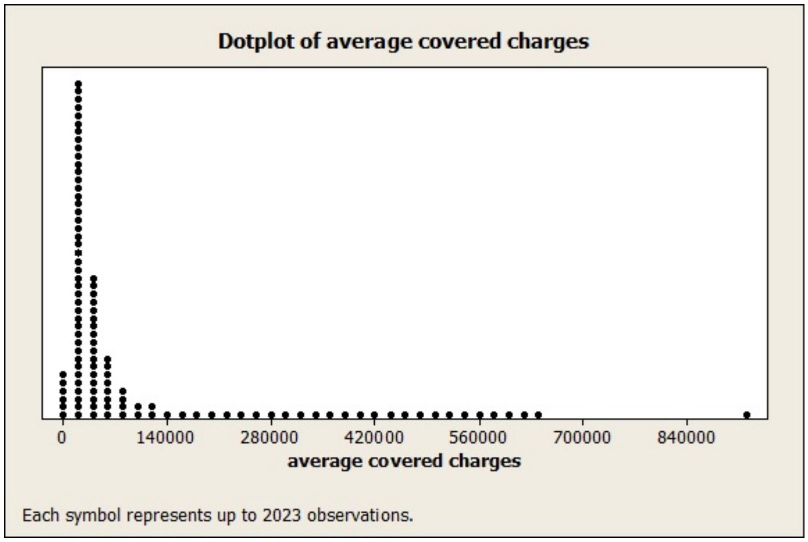



**Figure 1:** Dotplot of Charges.

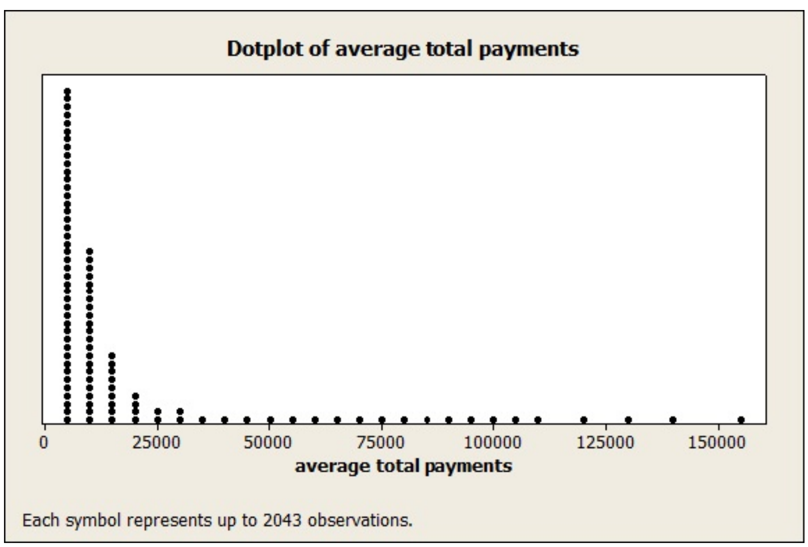



**Figure 2:** Dotplot of Payments.

To analyze the linear relationship between Charges and Payments, scatter plots and regressions were run using (Tableau Software, 2014). The results for the regressions for the Sum of Charges and the Sum of Payments which were aggregated by DRGs, State, City, and Provider are shown in Figures 3 through 6 below.

The relationship between Charges and Payments is examined further by first summing the values for each variable by every DRG and running a linear regression on the aggregated variables. The graph in Figure 3 indicates that when one regresses the Sum of the Average Total Payments by the Sum of the Average Covered Charges, a very strong linear relationship exists ($R^2$=0.988, b = .279; p<.0001). Based on this model, for each additional $1,000,000 in charges, approximately $279,000 is reimbursed by Medicare payments for every DRG specified.

When the same variables (Sum of Average Covered Charges and Sum of Average Total Payments) are plotted for each state rather than each DRG, the linear relationship still exists but is not as strong ($R^2$= 0.809298, b = 0.2330, p<0.0001). Unlike the case for the DRGs, this graph reveals some interesting outliers, at either extreme of almost total reimbursement and of very low reimbursement. For example, Maryland (MD) has Charges of 44.5M and 42.0M of Payments which translates into a 94.4% reimbursement. Similarly, though not quite as good, Massachusetts (MA) has Charges of approximately 78.9M and receives Payments of 39.5M for 50.1% reimbursement. Conversely, Nevada (NV) has Charges of 73.4M but Payments of only 12.4M which is a 16.9% reimbursement. Likewise, New Jersey (NJ) has a similar ratio of roughly Charges of 319.1M and Payments of 51.5M (16.1% reimbursement).

At a more disaggregated level, a regression model was run for the same variables by each city. The results were very similar to the model based on each state ($R^2$= 0.821641, b = 0.2395, p<0.0001). Not unexpectedly, since Maryland fared well in the previous model, Baltimore did extremely well here with Charges of 16.7M and Payments of 15.7M or 94.0% reimbursement. Likewise, since Nevada had very small proportional payments, Las Vegas did poorly with Charges of 49.2M but payments of only 7.4M or 15% reimbursement.

At the most basic level available, a regression model was run for the Charges and Payments based on each individual hospital/medical center provider. In this case, the model was not as good a fit as the earlier models ($R^2$ = 0.707331, b = 0.1868, p < 0.0001). This would be expected since the data are more variable since they are disaggregated. As a result, there are many points that are far from the estimated line, when compared to the previous models. As before, some of these points represent cases where the charges are almost fully reimbursed, (e.g., UMD Medical Center, with Charges of 1.89M and Payments of 1.78M or 94.2% reimbursement and Johns Hopkins Hospital, with Charges of 2.3M and Payments of 2.1M or 94.3% reimbursement) while others indicate situations where a very small proportion of the charges are paid (e.g., Crozer Chester Medical Center with Charges of 13.3M and Payments of 1.1M or 8.3% reimbursement, and Bayonne Hospital Center with Charges of 8.85 M and Payments of 600 thousand or only 6.8% reimbursement). In fact, this issue of high charges has been previously alluded to by (Herman, 2013)

> "The results relate back to May, when HHS and CMS released a trove of data on hospital inpatient charges. The data showed charges vary wildly across the country at different hospitals for the same procedure. For example, at Upper Chesapeake Medical Center in Bel Air, Md., the average Medicare charges for a major joint replacement with major complications and comorbidities totaled a little more than $23,000. At Crozer Chester Medical Center in Upland, Pa., a roughly hour-drive east, the average charges for the same procedure cost almost $322,000. However, hospital and health system executives have argued the charges do not reflect what they are actually reimbursed, although the OIG's report paints a much different picture."

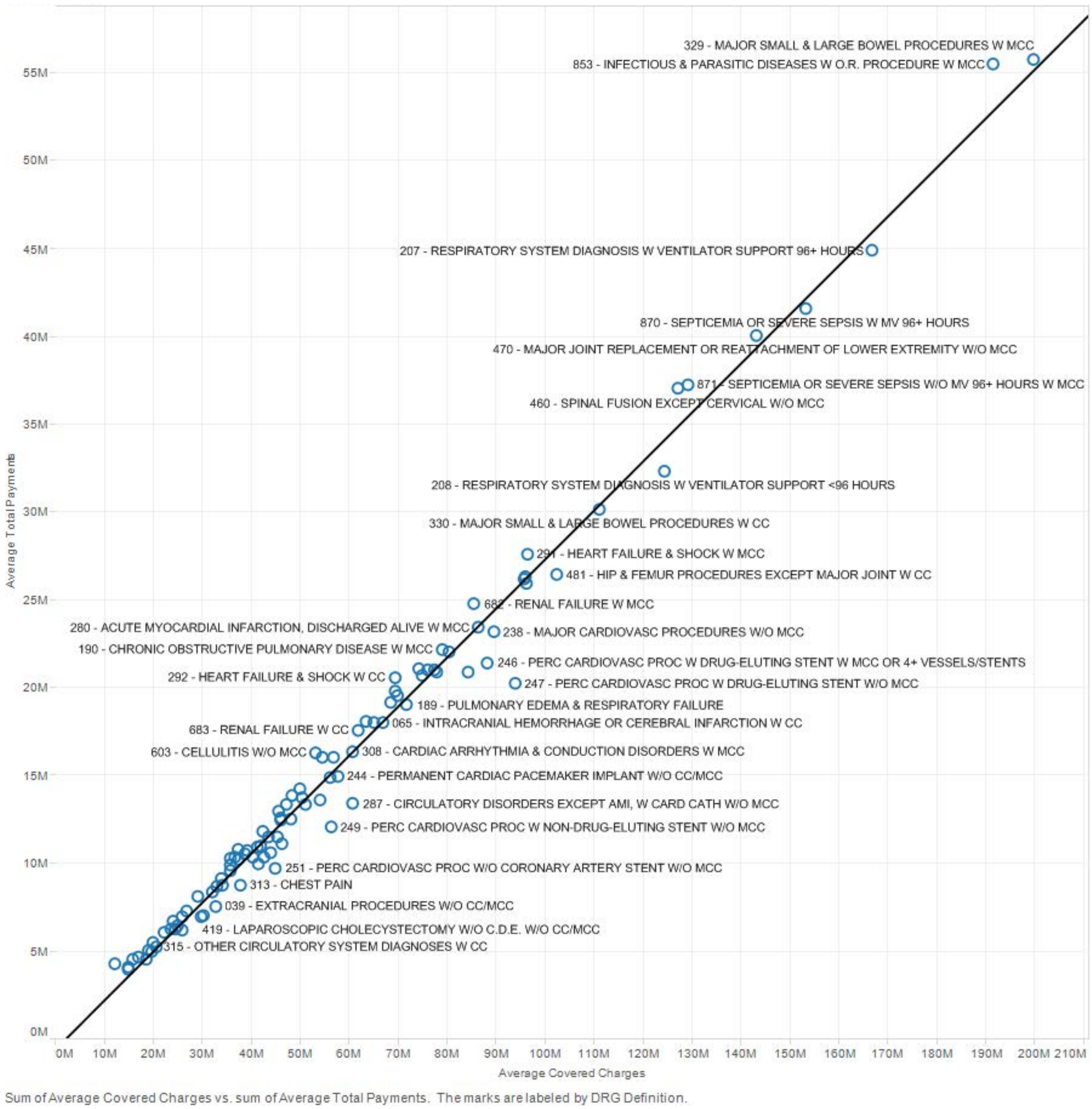
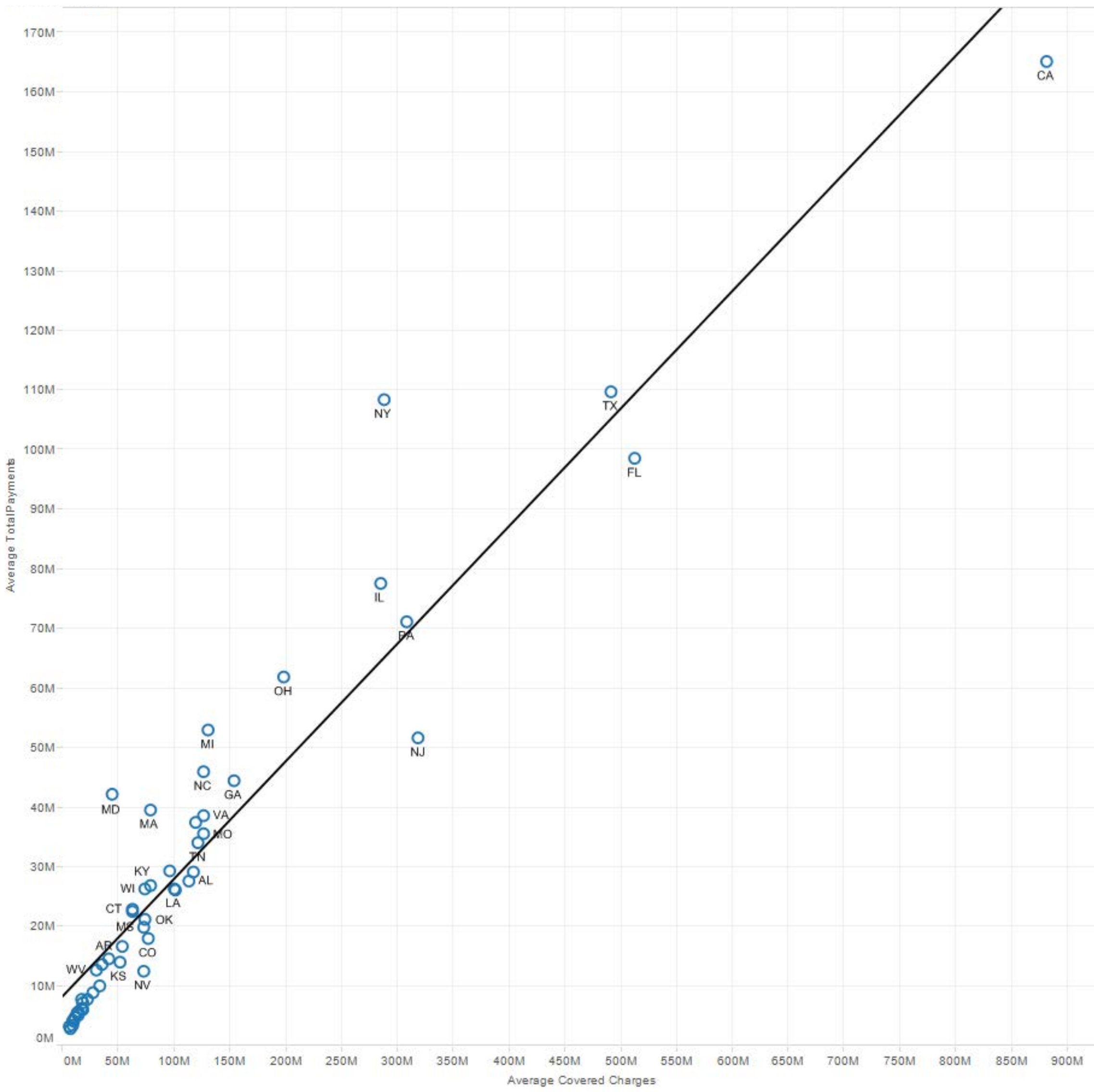



**Figure 3:** Scatter Plot of Payments by Charges Aggregated by DRGs.




Sum of Average Covered Charges vs. sum of Average Total Payments. The marks are labeled by Provider State.

**Figure 4:** Scatter Plot of Payments by Charges Aggregated by State.

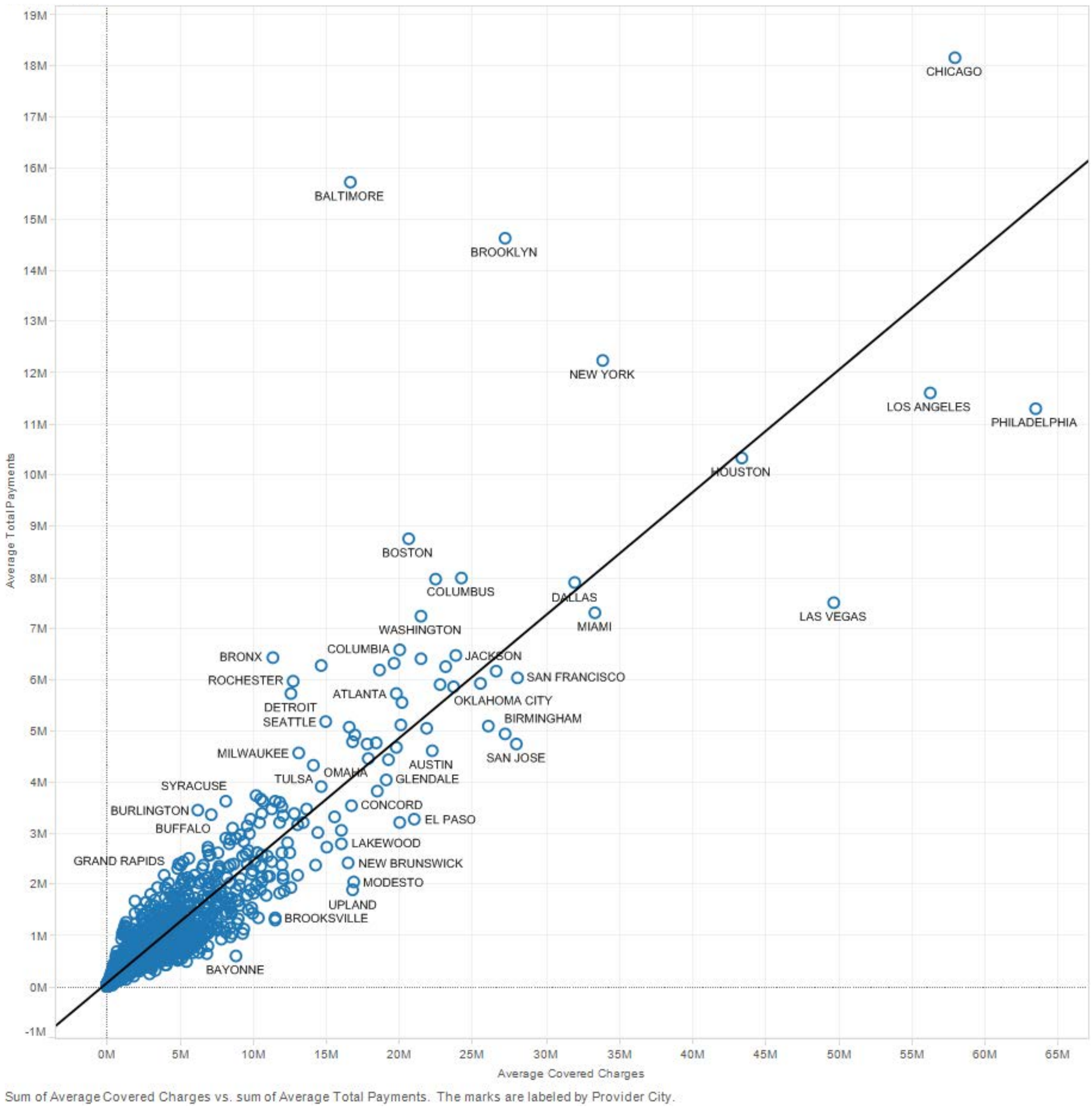



**Figure 5:** Scatter Plot of Payments by Charges Aggregated by City.

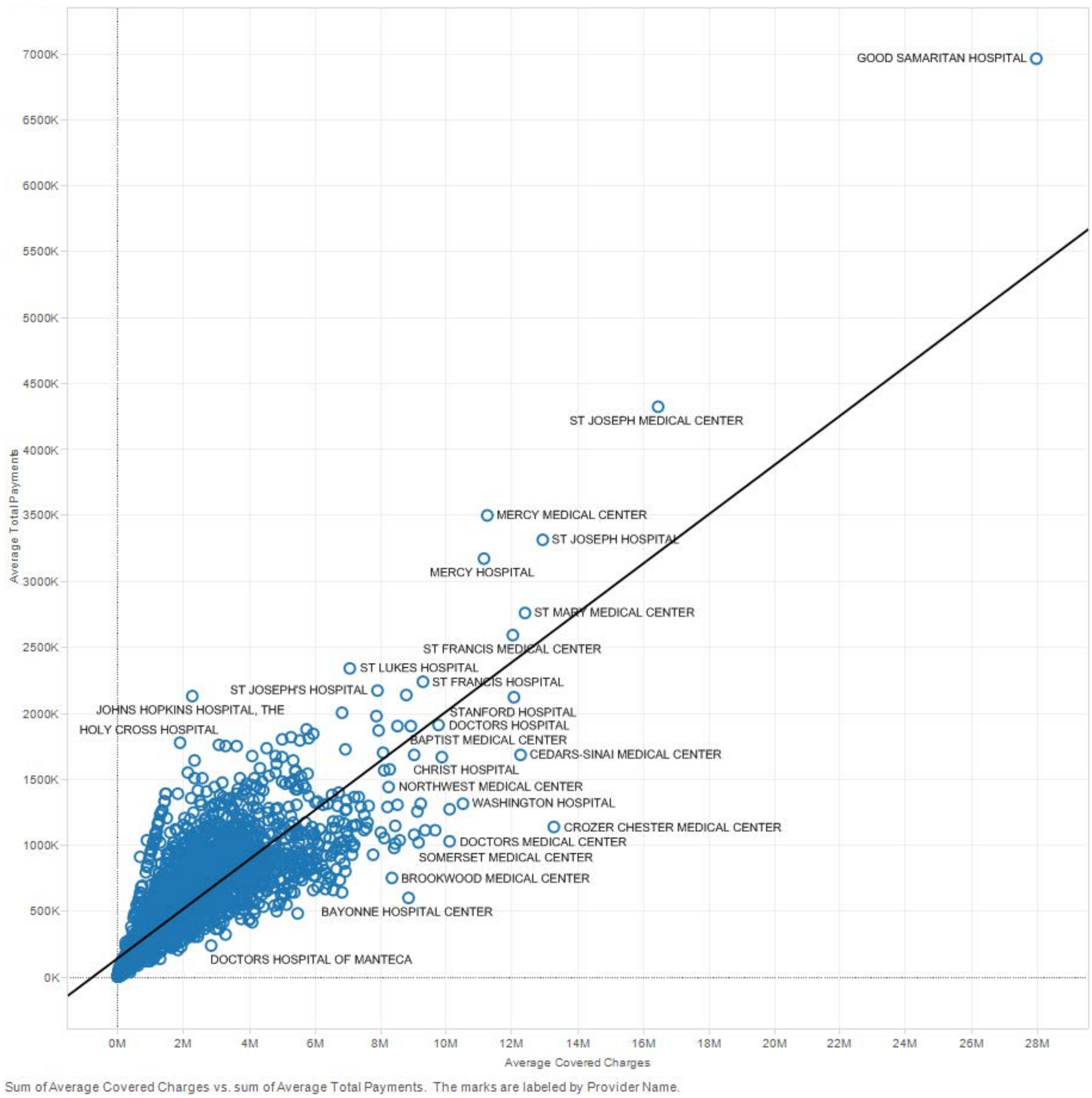



**Figure 6:** Scatter Plot of Payments by Charges Aggregated by Provider.

## 3. Benford's law background

Benford's Law describes the non-uniform distribution of digits in data values that many types of processes generate, such as financial data, values from some naturally occurring events and numbers that are produced which span several orders of magnitude. It is often used to help detect fraud in forensic accounting, health insurance and various other fields where data are expected to follow Benford's Law (Durtschi et al., 2004; Lu and Boritz, 2005; Maher and Akers, 2002; Nigrini, 2012, 2014). If deviations from the Law exist, then the implication is that the data might be fraudulent or artificially generated, as opposed to being a result of an open, ordinary process.

This interesting distribution of the digits 0 to 9 was first discovered in 1881 by an American astronomer, Simon Newcomb, who observed, "That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones." (Newcomb, 1881, p.39). These would be the pages where the numbers began with the digit 1. It was rediscovered in 1938 by Frank Benford, a physicist, who investigated the distribution of digits for various sets of scientific data and other types of values: "It has been observed that the first pages of a table of common logarithms show more wear than do the last pages, indicating that more used numbers begin with the digit 1 than with the digit 9. A compilation of some 20,000 first digits taken from widely divergent sources shows that there is a logarithmic distribution of first digits when the numbers are composed of four or more digits." (Benford, 1938).

Mark Nigrini also studied the distributions of the digits. In his studies of Benford's Law, he provides formulas for obtaining the probability distributions of the digits, and discusses several tests for whether or not a data set follows the distributions suggested by Benford's Law. Nigrini divides the tests into the Primary Tests and the Advanced Tests. The Primary Tests include the distribution of the first digits, the second digits and the first-two digits (also known as the first-order test). The Advanced tests include the second-order test and the summation test. The second-order test considers the distribution of the first-two digits of the differences between consecutive numbers in the data set when it is sorted in ascending order. In the summation test, the sum of all the values beginning with each of the first-two digits (10, 11, ..., 98, 99) is divided by the grand sum of all the values in the data set, and each fraction (of the total sum) is compared to the theoretical Benford proportions. We give the formulas for the primary tests below:

First Digit probabilities:

$$P\left(D_1 = d_1\right) = \log_{10}\left(1 + \frac{1}{d_1}\right) \quad \textit{where } d_1 \in \{1,\ 2,\ \ldots,\ 9\}$$

Second Digit probabilities:

$$P\left(D_2 = d_2\right) = \sum_{d_1=1}^{9} \log_{10}\left(1 + \frac{1}{d_1 d_2}\right) \quad \textit{where } d_2 \in \{1,\ 2,\ \ldots,\ 9\}$$

First-Two Digits probabilities:

$$P\left(D_1 D_2 = d_1 d_2\right) = \log_{10}\left(1 + \frac{1}{d_1 d_2}\right) \quad \textit{where } d_1 d_2 \in \{10,\ 11,\ \ldots,\ 99\}$$

These formulas are used to obtain the proportion of times a given digit appears in a given position. For example, the estimated probability or frequency of a number beginning with the digit 1 is:

$$\begin{aligned} P\left(D_1 = 1\right) &= \log_{10}\left(1 + \frac{1}{1}\right) \\ &= \log_{10}(2) \\ &\cong 0.30103. \end{aligned}$$

These calculations for the probabilities or frequency distributions of the digits can be represented in tabular form. Calculations for the probabilities of the digits 0 to 9 as they should appear in the first, second and third places of data which are expected to follow Benford's Law, are provided below in Table 3 (Nigrini, 2011, 2012).

Two observations can be made about these distributions, which are presented for the first three digit positions in Table 3. We note first that in all places, as the value of the digit increases, they occur with decreasing probabilities. However, the lower the order of the digit (for example, the third place digit or fourth place digit), the more uniform the distribution becomes.

**Table 3:** Proportions for the First, Second and Third Positions of Digits in Numbers from Benford's Law.

| Digit Position in Number | | | |
|---|---|---|---|
| **Digit** | **First** | **Second** | **Third** |
| 0 | | .11968 | .10178 |
| 1 | .30103 | .11389 | .10138 |
| 2 | .17609 | .10882 | .10097 |
| 3 | .12494 | .10433 | .10057 |
| 4 | .09691 | .10031 | .10018 |
| 5 | .07918 | .09668 | .09979 |
| 6 | .06695 | .09337 | .09940 |
| 7 | .05799 | .09035 | .09902 |
| 8 | .05115 | .08757 | .09864 |
| 9 | .04576 | .08500 | .09827 |

A common problem in data analysis is that there are missing values in many data sets. This can be due to a variety of reasons, based on the nature of the study. It could be intentional, such as the case with census data, where participants might refuse to divulge information that they consider sensitive, or with certain college admissions data, like SAT and ACT scores, which might not be provided to schools if they are score-optional institutions. The data could also be missing because of unforeseen circumstances, as when instruments fail, like an anemometer measuring wind speed at a weather station. No matter what the reasons for the missing data, a decision has to be made to either leave the records out of the analysis or to use an imputation method to fill in the gaps. In one approach to dealing with missing data when applying Benford's Law, Lu and Boritz provide an algorithm which offers a modified set of Benford frequencies which take into account the fact that there are missing values (Lu and Boritz, 2005).

It was thought that the average covered charges, which are the hospital bills, might follow Benford's law. On the other hand, since the payments by Medicare follow a schedule, with specific limits, we suspected that these average payments would not likely follow Benford's law.

## 4. Benford's law methodology

Our intention in this part of the paper is to investigate what happens when missing data are replaced by imputed values. More specifically, we are interested if replacing missing data by imputed values alters the relationship of a data set to Benford's Law. This is important for data sets that follow Benford's Law as well as for those that do not. If there is a large percentage of missing data that is replaced by these imputed values, does it alter the distribution of digits so that it no longer follows Benford's Law? Changes in relationship to Benford's Law might inadvertently lead to allegations that non-fraudulent data are fraudulent and vice-versa.

To investigate how imputation affects the distribution of data, we will use two variables, Charges and Payments, defined earlier. We will show that one, Charges, basically follows Benford's Law while the other, Payments, does not. We will remove 20% of the observations in both variables at random and replace these missing values using two types of imputation. We chose to make 20% of the data to be missing since that generally represents a substantial part of the data set. Saunders et al. (2006) reference studies where a small amount of missing data could be considered 5% or less, or even 20% or less. While Acuna and Rodriguez (2004, pg 1.) state:

> Missing data is a common problem in statistical analysis. Rates of less than 1% missing data are generally considered trivial, 1-5% manageable. However, 5-15% require sophisticated methods to handle, and more than 15% may severely impact any kind of interpretation.

To start, we will use the most common imputation method of replacing missing values with the mean. Then we will also impute missing values with uniform random numbers since it is possible that those fraudulently generating numbers would do so by randomizing the entries. This is reinforced by (Chang, 2013, pg. 13):

> Benford's Law has been used in fraud detection. As early as 1972, Hal Varian suggested that the law could be used to detect possible fraud in lists of socio-economic data submitted in support of public planning decisions. Based on the plausible assumption that people who make up figures tend to distribute their digits fairly uniformly, a simple comparison of first-digit frequency distribution from the data with the expected distribution from Benford's law ought to highlight any anomalous results.

Then we will test the modified data sets to see how they conform to the distribution of digits specified by Benford's Law. We will utilize Microsoft Excel to perform our analysis (Microsoft, 2010). SAS$^{TM}$ has also been used to test whether data follow Benford's Law (Smith, 2002). However, Excel is very easy to implement, particularly when used in conjunction with a template developed by Nigrini (2014). Finally, we will present our results graphically.

## 5. Results

Dotplots and frequency tables were provided earlier which illustrated the skewness of both the charges and the payments. A third frequency table is provided below, for these two variables, which is more disaggregated than the previous tables and contains unequal class widths. These intervals are chosen in order to facilitate the examination of the first digit of the dollar values so that they can help illustrate the level of conformity of the charges and the payments to Benford's Law.

**Table 4:** Frequency Table of Charges and Payments for illustrating the first digit.

| Dollar Range | Charges | Payments |
|---|---|---|
| 2000-2999 | 8 | 192 |
| 3000-3999 | 49 | 10655 |
| 4000-4999 | 280 | 24675 |
| 5000-5999 | 695 | 22074 |
| 6000-6999 | 1347 | 20259 |
| 7000-7999 | 2127 | 14617 |
| 8000-8999 | 2771 | 10940 |
| 9000-9999 | 3687 | 8752 |
| 10000-19999 | 49936 | 39130 |
| 20000-29999 | 35016 | 6521 |
| 30000-39999 | 21278 | 3232 |
| 40000-49999 | 13577 | 1239 |
| 50000-59999 | 8891 | 458 |
| 60000-69999 | 5976 | 200 |
| 70000-79999 | 4240 | 72 |
| 80000-89999 | 3017 | 29 |
| 90000-99999 | 2257 | 11 |
| 100000-199999 | 6795 | 9 |
| 200000-299999 | 863 | 0 |
| 300000-399999 | 185 | 0 |
| 400000-499999 | 40 | 0 |
| 500000-599999 | 24 | 0 |
| 600000-699999 | 4 | 0 |
| 700000-799999 | 0 | 0 |
| 800000-899999 | 0 | 0 |
| 900000-999999 | 2 | 0 |
| Total | 163065 | 163065 |

Having examined the data, we now test whether our expectations of Charges basically fitting Benford's Law and Payments not fitting are justified. The plots presented below were obtained using Nigrini's Excel template (2014). We have included graphs for the first digit test, the first-two digits test and the second-order test for Charges (Figures 7 to 9), and Payments (Figures 10 to 12).

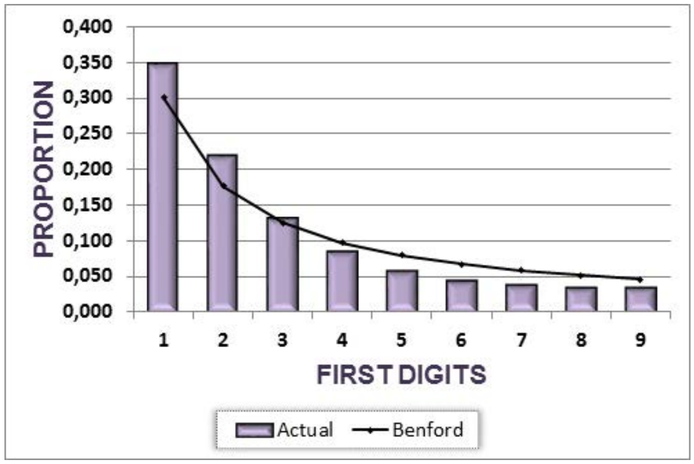



**Figure 7:** Charges: First Digit Test.

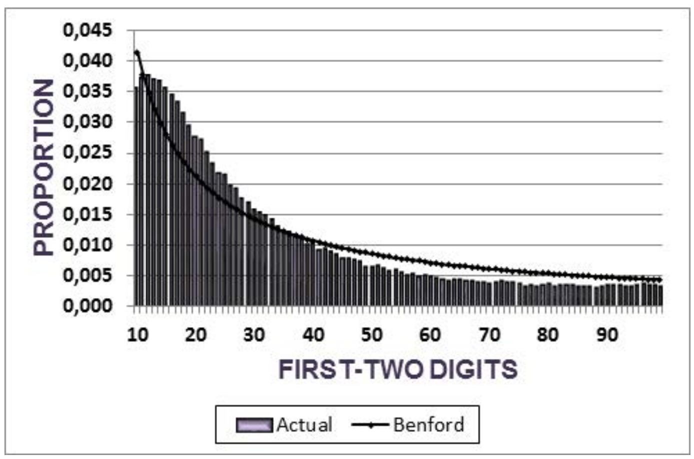



**Figure 8:** Charges: First-Two Digits.

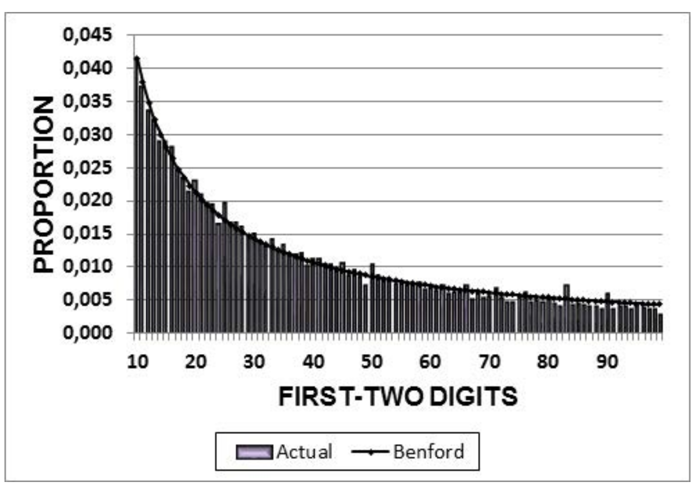



**Figure 9:** Charges: Second-Order Test.

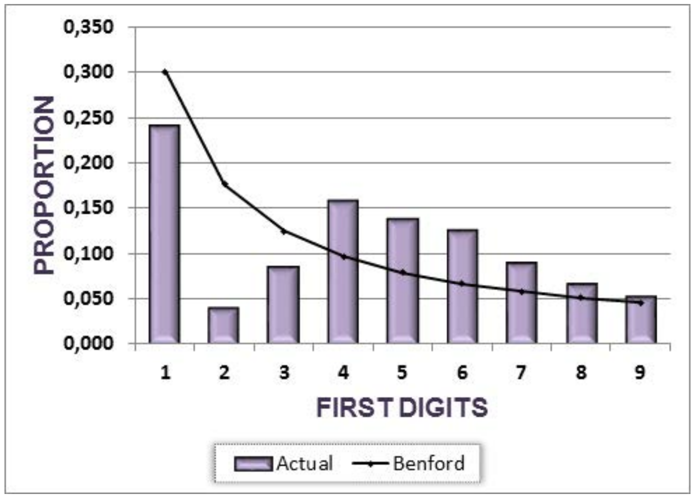



**Figure 10:** Payments: First Digit Test.

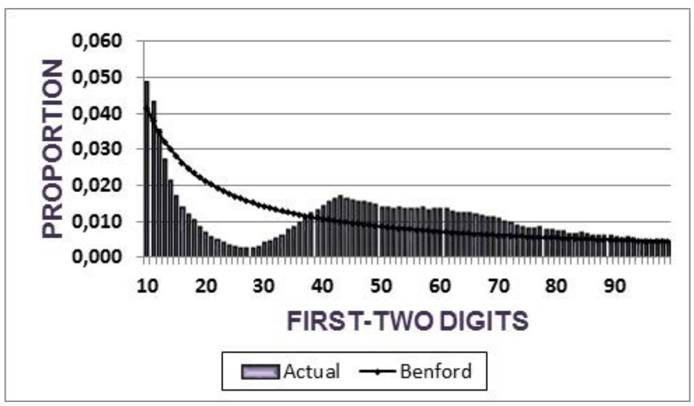



**Figure 11:** Payments: First-Two Digits Test.

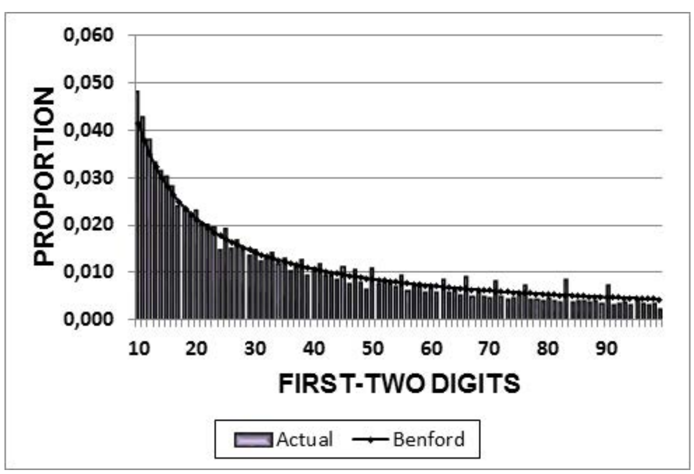



**Figure 12:** Payments: Second-Order Test.

With reference to the first digit test (Figure 7) and first-two digits test (Figure 8), it seems to be true that the Charges do in fact follow Benford's Law fairly closely. The Payments, however, do not appear to follow Benford's Law. For both the first digit test (Figure 10) and the first-two digits test (Figure 11) for the Payments, there is clearly a dip around the 2-3 or 20-30 digits. This coincides with our expectation since Payments are structured around the DRG procedures, with specified payment limits and thus are not expected to follow Benford's Law. Also, this decrease in payments beginning with a first digit of 2 or 3, can be seen in Table 4 where the number of payments from $2000 to $2999 (192) and from $3000 to $3999 (10655) are much smaller than in the ensuing payment intervals, like $4000 to $4999 (24675) and $5000 to $5999 (22074), and so forth. We have not included the second digits test or the summations test for any of the scenarios since they weren't very informative.

The second-order test, which is represented in Figures 9 (Charges) and 12 (Payments), analyzes the first-two digits of the numbers resulting from taking the differences in successive values when the data are sorted in ascending order. For this test, both Charges and Payments follow the predicted Benford proportions pretty closely, which might be due to the number of observations being so large. Nigrini states, "Let $x_1$, $x_2 \ldots, x_N$ be a data set comprising records drawn from a continuous distribution, and let $y_1$, $y_2 \ldots, y_N$ be the $x_i$'s in increasing order. Then, for many data sets, for large $N$, the digits of the differences between adjacent observations ($y_{i+1}$- $y_i$) is close to Benford's Law." (Nigrini, 2012).

For the mean imputation, we have only included one plot (Figure 13) since it fully represents the distortion of the distribution caused by replacing 20% of the missing values with one constant.

As can be seen in Figure 13, mean imputation clearly changes the distribution of the data for the average charges. Therefore, it would no longer satisfy Benford's Law. The spike in the plot at the mean is very evident and appears in the plots for all scenarios for both Charges and Payments.

The same three tests were then run with Uniform Random Imputation for Charges (Figures 14 to 16) and Payments (Figures 17 to 19). There were some interesting results when the random uniform imputation was applied to the Charges data set. For example, though the original Charges data were a fairly good fit for Benford's for the first digit and first-two digits tests (Figures 7 and 8), there were generally slightly higher proportions of the lower digits and lower proportions of the higher digits than Benford's expected fractions. When 20% of the values were replaced by random uniform values, the first digits and first-two digits became more spread out which adjusted these fractions of the digits, so that the imputed data set looks like a better fit (Figures 14 and 15) than the original. In other words, if we were to consider the first digits and first-two digits, we could be fooled if the imputed data were actually fraudulent. However, this is not the case with the second-order test. In this instance, the first-two digits for the original set (Figure 9) follow the expected theoretical Benford proportions very closely. The same test for the imputed charges data set (Figure 16) also generally follows the Benford proportions with one important exception, there are prominent spikes when the second digit is 0 (i.e., at 10, 20, . . . , 90). This could be due to rounding of the random uniform values. Nigrini points out, "A pattern of spikes at the prime first-two digits *10, 20, . . . , 90* will occur if these differences are drawn from data from a discrete distribution." (Nigrini, 2012, p. 99).

For the original Payments data set, only the second-order test indicated any similarity to Benford's (Figure 12) but that doesn't necessarily signify much. As alluded to above, non-Benford sets, like many uniform density distributions, can match Benford's with the second-order (differencing) test (Nigrini, 2012,

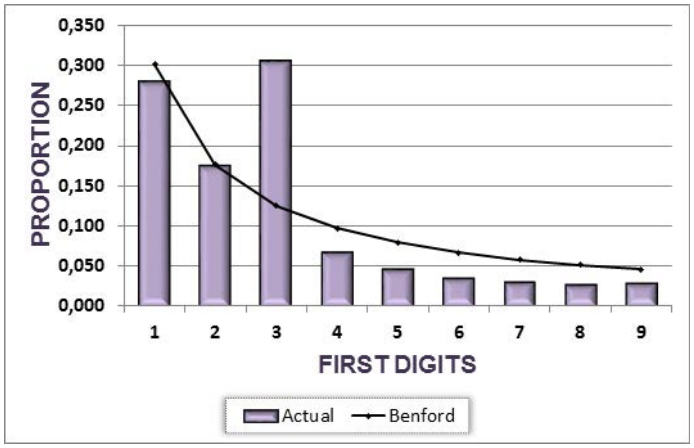



**Figure 13:** Charges: Mean Imputation First Digit Test.

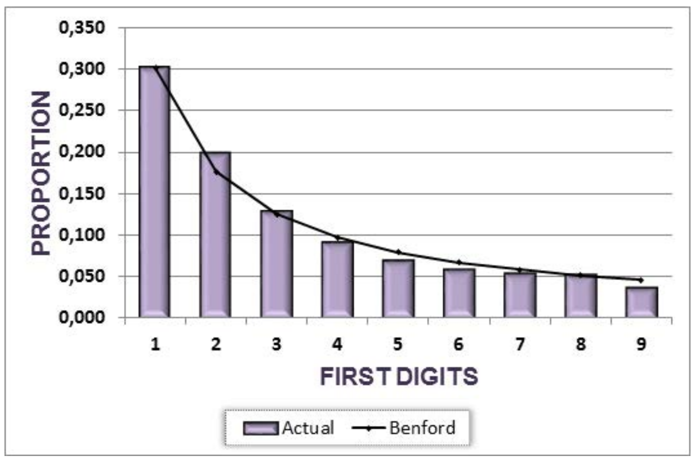



**Figure 14:** Charges: Random Uniform Imputation First Digit Test.

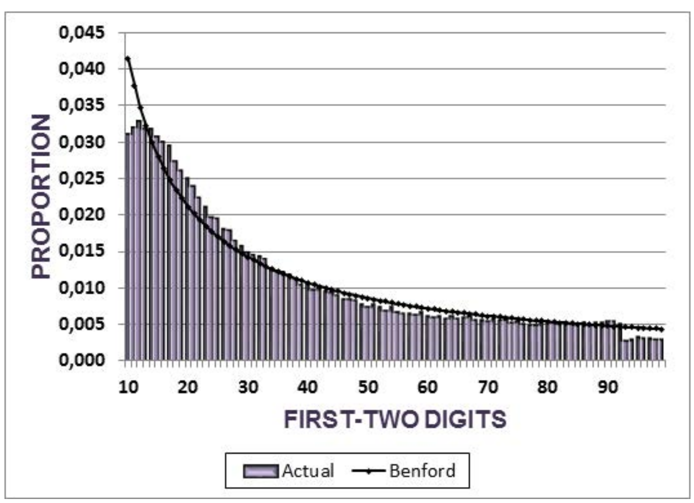



**Figure 15:** Charges: Random Uniform Imputation First-Two Digit Test.

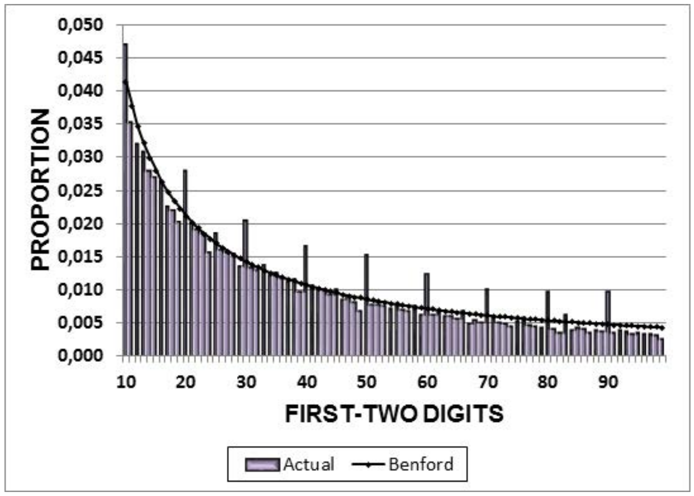



**Figure 16:** Charges: Random Uniform Imputation Second-Order Test.

p. 99). With Payments being non-Benford, it might have seemed that replacing 20% of the data with uniform random values would have spread out the data and made the set more Benford-like (by helping fill the gap for the lead digits of 2 and 3 mentioned above). However, the imputed first digit and first-two digits tests (Figures 17 and 18) closely resemble the corresponding graphs for the full Payments (Figures 10 and 11). The imputed second-order test (Figure 19) has similar spikes to what occurred with the imputed Charges (Figure 16).

## 6. Conclusions

The analysis of the Medicare Charges and Payments by DRG, State, City and Provider resulted in several interesting conclusions. Overall, Payments were approximately one quarter of the Charges. This is not surprising since in general, health insurance allows a limited amount for specific medical services and providers correspondingly charge considerably more than they expect to be reimbursed. This was closely observed in the plots and regression model results of the Payments by Charges aggregated by DRG. However, we observed many outliers in the other models. For instance, by State, Maryland was almost fully reimbursed, MA was reimbursed by more than 50%; whereas, New Jersey and Nevada was reimbursed for about 1/6 of their covered charges. When the data were examined by City, it turns out that Baltimore had almost 100% reimbursement; whereas Los Vegas had approximately 1/6 of the charges reimbursed. These outliers can be further seen when disaggregating the data by provider. Within Baltimore Maryland, the provider that stands out is Johns Hopkins Hospital with charges of 2.26 million and payments of 2.13 million or 94.3% reimbursement and the provider, Bayonne Hospital Center, located in Bayonne, New Jersey had charges of 8.85 million with payments of only 600 thousand or only 6.8%.

After analyzing the data, we addressed the question as to whether or not either of the two main variables, Charges and Payments, satisfied Benford's Law. There are five Benford tests (three primary and two advanced). We have included the results from three of these five tests (First Digit, First-Two Digits and Second-Order test). We determined that the variable Hospital Medicare Average Covered Charges (Charges) is approximately a Benford Set while the Hospital Medicare Average Total Payments (Payments) variable is not. We then investigated what happens if 20% of the observations are replaced with imputed values, to decide whether it significantly affected the distribution of digits. One common method of im-

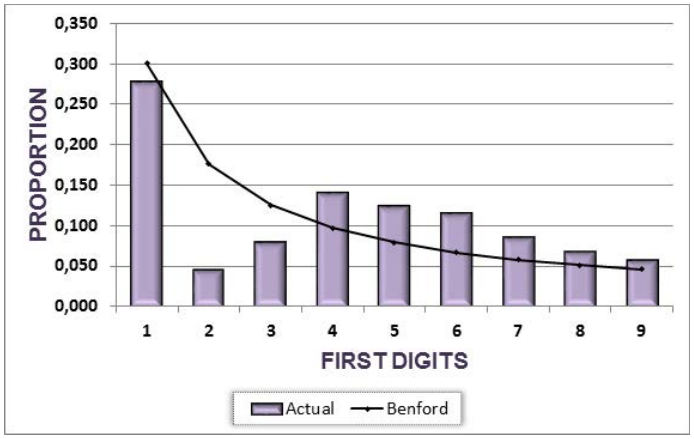



**Figure 17:** Payments: Random Uniform Imputation First Digit Test.

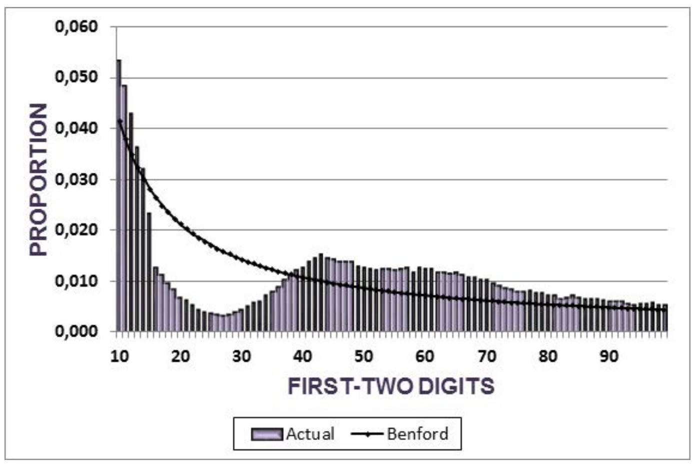



**Figure 18:** Payments: Random Uniform Imputation First-Two Digit Test.

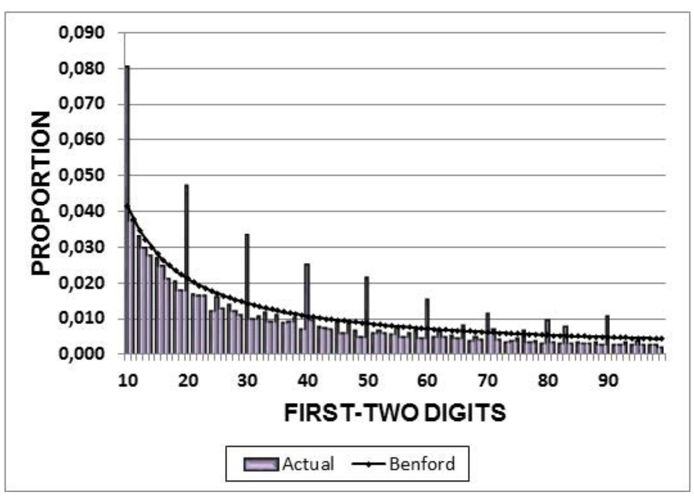



**Figure 19:** Payments: Random Uniform Imputation Second-Order Test.

putation is to use the mean value of the data. This resulted in an expected spike in the First Digit test since we were replacing 20% of the data with one constant. We didn't proceed any further with that analysis. We then tried a different imputation method, in which we replaced 20% of the observations with uniform random values, with the idea that people tend to randomize numbers when asked to make up values. The results were mixed in that the Charges imputed data were as close or closer to the Benford proportions for two tests (First Digit and First-Two Digits) than the original data set. As such, based on the graphs, it could be easy to disguise the fact that 20% of the data were artificially generated. On the other hand, the imputed values for the Payments variable did not seem to have much of an effect based on the Benford analysis. There was a gap in the original data with numbers beginning with a 2 or 3, and that did not change much when 20% of the values were replaced. That is, the imputed set was also non-Benford. These results imply that imputation of missing values with uniform random numbers does not significantly change how well the data set fit Benford's Law.

## References

Acuna, E. and Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications*, pages 639–647. Springer.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572.

Centers for Medicare and Medicaid Services (2013). *Medicare Provider Utilization and Payment Data: Inpatient*. Retrieved July 2, 2014 on https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient.html.

Chang, M. (2013). *Paradoxes in scientific inference*. Boca Raton, FL: CRC Press.

Durtschi, C., Hillison, W., and Pacini, C. (2004). The effective use of benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting*, 5(1):17–34.

Herman, B. (2013). OIG: High Hospital Charges Lead to Excessive Medicare Outlier Payments. *Becker's Hospital Review*. Retrieved May 30, 2014 on http://www.beckershospitalreview.com/finance/oig-high-hospital-charges-lead-to-excessive-medicare-outlier-payments.html.

Lu, F. and Boritz, J. E. (2005). Detecting fraud in health insurance data: Learning to model incomplete benford's law distributions. In *European Conference on Machine Learning*, pages 633–640. Springer.

Maher, M. and Akers, M. (2002). Using benfords law to detect fraud in the insurance industry. *International Business & Economics Research Journal (IBER)*, 1(7):1–11.

Microsoft (2010). Microsoft excel. Redmond, Washington: Microsoft.

Minitab Inc (2010). *The Minitab 16 software*. State College, PA.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4(1):39–40.

Nigrini, M. (2011). *Forensic analytics: methods and techniques for forensic accounting investigations*. John Wiley & Sons, Inc., Hoboken, NJ.

Nigrini, M. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud detection*. John Wiley & Sons, Inc., Hoboken, NJ.

Nigrini, M. (2014). *Nigrini Excel Template*. Retrieved on July 2, 2014 on http://www.nigrini.com/benfordslaw.htm.

Quinn, J., Schumacher, P., and Olinsky, A. (2014). Benford's law and imputation of missing data. In *Proceedings of the Northeast Decision Sciences Institute Annual Conference, Philadelphia, PA, May 27-29*.

Saunders, J. A., Morrow-Howell, N., Spitznagel, E., Doré, P., Proctor, E. K., and Pescarino, R. (2006). Imputing missing data: A comparison of methods for social work researchers. *Social work research*, 30(1):19–31.

Smith, C. (2002). Detecting anomalies in your data using benford's law. *Proceedings of the SAS Users Group, SUIG27*. Paper 249-27.

Tableau Software (2014). *Tableau, Version 8.1*. Seattle, WA.

Correspondence: aolinsky@bryant.edu.