Social Science & Medicine 356 (2024) 117155

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Assessing the accuracy of self-reported health expenditure data: Evidence from two public surveys in China

Zhuang Hao, Xudong Zhang, Yuze Wang *

College of Economics and Management, Huazhong Agricultural University, Wuhan, 430070, China

ARTICLE INFO

Handling Editor: Joanna Coast

JEL classification:
C19
C81
C83
I10

Keywords:
Health expenditure
Data quality
Benford's law
Public survey

ABSTRACT

This paper utilizes Benford's law, the distribution that the first significant digit of numbers in certain datasets should follow, to assess the accuracy of self-reported health expenditure data known for measurement errors. We provide both simulation and real data evidence supporting the validity assumption that genuine health expenditure data conform to Benford's law. We then conduct a Benford analysis of health expenditure variables from two widely utilized public datasets, the China Health and Nutrition Survey and the China Family Panel Studies. Our findings show that health expenditure data in both datasets exhibit inconsistencies with Benford's law, with the former dataset tending to be less prone to reporting errors. These results remain robust while accounting for variations in survey design, recall periods, and sample sizes. Moreover, we demonstrate that data accuracy improves with a shorter time interval between hospitalization and interviews, when the data is self-reported as opposed to proxy responses, and at the household level. We find no compelling evidence that enumerators' assessments of respondents' credibility or urgency to end interviews are indicative of data accuracy. This paper contributes to literature by introducing an easy-to-implement analytical framework for scrutinizing and comparing the reporting accuracy of health expenditure data.

1. Introduction

Over the past twenty years, there has been a remarkable surge in demand for medical services around the world, attributable to various factors such as the pandemics, the gradual expansion and increased generosity of health insurance coverage, rising income levels, an ageing population, the widespread adoption of medical technology, and shifts in the disease spectrum. In 2020, global health spending reached US$ 9 trillion, which accounted for more than 10% of global GDP.[1] Publicly accessible survey data have emerged as the primary data source for research related to the utilization and expenditure of health care and services, providing a foundation basis for evaluating health policies and aiding government oversight of the healthcare system.

In the Chinese context, per capita medical expenditure has significantly increased, rising from US$ 111 in 2000 to US$ 671 in 2021.[2] Health economics and policy studies on rapidly growing health expenditure and utilization have relied on large-scale surveys in China and two prominent publicly accessible datasets are the China Health and Nutrition Survey (CHNS) and the China Family Panel Studies (CFPS). These surveys offer nationally representative and comprehensive information on individual utilization and expenditure of health care and services compared to other public surveys.[3] The CHNS is a comprehensive database that aims to examine the impact of social and economic transformation of Chinese society on the health and nutritional status of its population. It collects data on participants' health expenditures over the previous four weeks, including precise amounts spent

* Corresponding author.
E-mail addresses: zhuang.hao@mail.hzau.edu.cn (Z. Hao), xvdong_z@webmail.hzau.edu.cn (X. Zhang), wangyz@mail.hzau.edu.cn (Y. Wang).

[1] Global health spending was retrieved from the World Health Organization report of Global Spending on Health at https://www.who.int/publications/i/item/9789240064911.

[2] Annual health expenditure per capita was inflation-adjusted to the current US dollar of 2021 and retrieved from the World Health Organization Global Health Expenditure database at https://apps.who.int/nha/database.

[3] There are other public surveys available for health-related information, but they are either less representative of the entire population or record less information about health expenditure and utilization. Specifically, China Health and Retirement Longitudinal Study (CHARLS) and Chinese Longitudinal Healthy Longevity Survey (CLHLS) focus on the middle-aged and elderly population; China Household Finance Survey (CHFS) records less information about the utilization of healthcare services.

https://doi.org/10.1016/j.socscimed.2024.117155
Received 6 January 2024; Received in revised form 30 June 2024; Accepted 18 July 2024
Available online 26 July 2024
0277-9536/

on health care and services. The survey also gathers information on participants' socioeconomic status, demographic characteristics, and geographic location, all essential for analyzing expenditure data. In comparison, the CFPS collects data on a broader array of topics from a greater number of provinces but lacks the same granularity as the CHNS concerning health expenditures. For example, the CFPS only captures data related to hospitalization expenses over the preceding 12-month period, excluding outpatient care and preventative care costs. Additionally, the CFPS employs well-designed sample weights, enhancing the representativeness and generalizability of estimates and findings using its data.

Researchers have used these datasets to enhance comprehension of the factors influencing healthcare spending (Shi et al., 2021; Fu et al., 2022; Si and Chu, 2022). Moreover, the CHNS and CFPS have become the most prevalent datasets to quantify disparities in health expenditure across the lifespan (Feng et al., 2015), residential status and income (Yip et al., 2019), and the impacts of various health conditions such as influenza (Liu et al., 2012), depression (Hsieh and Qin, 2018), and extreme temperature (Li et al., 2023). Furthermore, these datasets have also been employed to evaluate the effects of health insurance coverage and generosity on demand for healthcare services (for examples, see Lei and Lin, 2009; Liu and Zhao, 2014; Huang and Gan, 2017; He and Nolen, 2019; Zhang et al., 2019; Zhao, 2019; Sun, 2020; Huang and Liu, 2023). However, despite the extensive use of self-reported health expenditure data, existing research has not given adequate attention to their accuracy, the extent to which the reported data aligns with the actual expenses it is designated to measure.

Self-reported survey data is well known to be susceptible to measurement error issues, with recall error being an extensively studied reason. In the context of health expenditure data, the presence of recall error can be suggested by a notable data pattern where reported values disproportionately concentrate on certain heaped figures, such as 500 and 1000. This specific type of recall error that mistakenly rounds or approximates values is referred to as the "heaping error".[4] Recall errors are particularly relevant when considering self-reported health expenditure data, which often stems from complex past processes involving different parties, procedures, and locations. Aside from recall error, misunderstandings of survey questions or reluctance to provide accurate responses can also introduce measurement errors into the data (Rosenman et al., 2011). The presence of measurement errors in survey data will typically cause biased and inconsistent estimates of regression model parameters (see Bound et al. (2001) and Schennach (2016) for more complete reviews). This is especially troublesome when such estimates are used to formulate health and other social policies that have far-reaching effects on the broader population and the allocation of substantial budgetary resources. Estimation methods that provide consistent estimates for both linear and nonlinear models with measurement errors have been devised in the measurement error literature. However, these methods rely on strong assumptions about the properties of the error, and require extra information in the form of either instrumental variables or auxiliary samples to overcome the loss of information induced by the presence of measurement error (Bound et al., 2001; Chen et al., 2011).

For health expenditure data, its accuracy is typically assessed using external reference sources for validation, which is a benchmark approach. An ideal approach would involve meticulous monitoring of individual expenditures across all levels of healthcare providers, including hospitals, clinics, and pharmacies (Lavado et al., 2013). However, employing and integrating such rigorous external data sources is often infeasible due to the associated costs. Designs of specific surveys might allow for examining the data reliability using multiple indicators of variables measured with error. For example, taking advantage of the total health expenditure as well as specific expenses on detailed health services in the World Health Survey, Xu et al. (2009) used the deviation of reported total expenditure from the aggregated expenditure to assess the expenditure data quality.

In the absence of an external gold standard or internal survey information for data validation, Benford's law has been utilized as a benchmark to evaluate the accuracy of self-reported data (Kaiser, 2019; Dang and Owens, 2020). Benford's law posits that the first significant digit in many naturally occurring numerical sequences is more likely to be a lower number than a larger one, and it provides specific expectations for the distribution of these digits (Benford, 1938). According to this law, there is an expectation that natural numbers commencing with digit 1 will manifest approximately 30% of the time. Similarly, numbers starting with digit 2 are projected to occur for about 18% of the time. The prevalence of leading digits progressively diminishes, with numbers commencing with digit 9 anticipated to manifest for less than 5% of the time. With recent advances in the knowledge about Benford's law, Villas-Boas et al. (2017) demonstrated that a diverse set of data generated from the economic behavioral systems is widely accepted to conform to Benford's law.

In this paper, we utilize Benford's law to analyze the accuracy of self-reported health expenditure data within the context of public surveys. We first demonstrate the applicability of Benford's law to health expenditure data with two pieces of evidence, one using real data and the other using simulated data, supporting the validity assumption that genuine health expenditure data should conform to Benford's law. Using the CHNS and CFPS as illustrative datasets, we show that health expenditure data reported in neither dataset is completely consistent with Benford's law. However, overall, the data in the CHNS tends to be less prone to reporting errors. These results remain robust even after accounting for various factors, including survey design, recall periods, and sample sizes. We further examine whether data accuracy can be indicated by other information available in the datasets. Analysis results suggest that a longer time interval between hospitalization and interview, a proxy response by other family members, and individual-level rather than household-level reported data are associated with lower accuracy of data, and enumerators' opinions about respondents cannot indicate data accuracy.

This study makes several contributions to the literature. First, it bridges the gap between Benford and measurement error literature by demonstrating that deviations from Benford's law indicate biases in regression model parameter estimates. While most Benford studies attempt to show their hypothetical true data should follow Benford's law using various methods, few have tested whether the Benford analysis results correlate with parameter estimates of most empirical researchers' interest. This may limit the impact of the Benford analyses in applied economics and policy studies. Our simulation results show that greater deviations from Benford's law are positively correlated with larger estimation biases of regression model parameters. Second, to the best of our knowledge, this is the first study to employ Benford's law as a screening tool for assessing health data. This complements health literature by addressing challenges in acquiring reliable external validation data and facilitating comparisons across datasets with different designs, such as survey scales and recall periods. The assessment framework can be particularly useful in less developed countries where obtaining accurate administrative records for validation is often less feasible. Finally, this study provides new evidence on potential indicators of misreporting in health expenditure data, which may inform future survey designs for health expenditure.

The paper is organized in the following way. Section 2 describes the analysis strategy. Section 3 documents health expenditure measures in the CHNS and CFPS. Section 4 discusses the main results, checks for robustness, and examines potential indicators of data accuracy. Section 5 concludes the paper.

[4] It is well documented that heaping is a common type of recall error in self-reported consumption and expenditure data, of which many values disproportionately concentrate on certain figures, such as multiples of numbers with the FSD of 5 (Browning et al., 2003; Paulin and Krishnamurty, 2018).

2. Benford Analysis

2.1. Analytical challenges

Ideally, an accurate, uniform external data source would serve as a gold standard for health data validation.[5] Such sources could include detailed insurance claims data or hospital administrative records, which are less prone to reporting bias. For example, Kjellsson et al. (2014) compared the self-reported number of hospitalized nights to registered data for assessing the impact of the length of recall periods on data quality. However, obtaining registered data serving as a gold standard can be prohibitively costly in most cases. To evaluate the accuracy of health expenditure data without a gold standard, researchers often turn to features of the survey design to assess data quality based on internal consistency. For example, Xu et al. (2009) utilized two types of test-retest procedures to evaluate the consistency of reported health expenditure within the World Health Survey. One test-retest compared the reported expenses in the original survey to the reported expenses in a follow-up survey, and the other compared the reported total health expenditure to the calculated total health expenses by summing reported expenses on detailed categories. However, these approaches rely heavily on the unique structure and design of the survey.

When working with datasets like the CHNS and CFPS, which cover multiple provinces and years, it is challenging to obtain a uniform external data source for validation. Furthermore, the design of both surveys does not readily support test-retest procedures for assessing internal consistency in health expenditure data. Moreover, the two datasets differ in terms of the survey areas, years and recall periods used for health expenditure data, which makes it less intuitive to compare the two datasets directly. These challenges call for alternative approaches for evaluating data accuracy.

2.2. Benford's law

In light of these analytical challenges, we turn to Benford's law, which has been used to assess data quality, particularly its accuracy, in various fields. In his seminal work, Frank Benford (1938) stated that for many real-world sets of numerical datasets, the first significant digit (FSD) is more likely to be small rather than large, following a specific distribution (hereafter, denoted by Benford distribution). The probability of observing the first significant digit, d, should be approximately equal to $\log_{10}(d+1) - \log_{10}(d)$. According to the law, the number 1 appears as the first significant digit about 30.1% of the time, while the number 9 appears as the maximum significant digit for only 4.6% of the time. Benford distribution is presented in Panel A of Table A4. Several satisfactory explanations of Benford's law were detailed in the work of Boyle (1994) and Hill (1995), and a recent summary of the mathematical intuition underlying Benford's law was discussed by Dang and Owens (2020).

Benford's law was initially applied to naturally generated data sequences, such as address numbers, molecular weights, and areas of rivers (for a survey, see Miller, 2015). By contrast, certain types of numerical sequences disobey Benford's law. Such examples include data with limited values (e.g., telephone numbers in a given region with the same area codes), sequentially assigned numbers (e.g., Student ID), and human thought-influenced numbers (e.g., the price set by psychological threshold like $9.99) (Durtschi et al., 2004; Schräpler, 2011). With recent advances in the knowledge about Benford's law, its applications have been extended to economics and social sciences. Villas-Boas et al. (2017) demonstrated that Benford's law is frequently observed in economic and behavioral systems, particularly in tax returns (Nigrini, 1996), macroeconomic measures (Gonzalez-Garcia and Pastor, 2009), household income and individual earnings (Kaiser, 2019), expenses and revenues of organizations (Dang and Owens, 2020), industrial output at the city level (Huang et al., 2020), and household food consumption (Abate et al., 2023). Numerical sequences that result from the mathematical combination of numbers (e.g., the product of price and quantity) and the transaction level data (e.g., payments) would be particularly expected to conform to Benford's law.

Existing studies have provided various explanations and proofs of the convergence of a random variable to Benford distribution. Boyle (1994) showed that a random variable asymptotically converges to Benford distribution as a consequence of underlying multiplicative operations. Hill (1995) outlined conditions for the application of Benford's law to data and emphasized that for samples randomly taken from a set of random distributions, the resultant FSD of all numbers would follow the Benford distribution. Wallace (2002) discussed the statistical criterion of adherence to Benford's law that the data is right-skewed and the mean is higher than the median. It follows that the larger the ratio of the mean divided by the median, the more closely the dataset will follow Benford's law. Clementi and Gallegati (2005) stated that among common distributions, the log-normally distributed economic data (e.g., income or expenditure) tends to conform to Benford's law.

2.3. Testing Benford's law

Two primary methods are commonly employed to assess the accuracy using Benford's law in the existing literature: the statistical testing of the hypothesis concerning the equality between the observed FSD distribution and the Benford distribution, and the statistical measure of the deviation of the observed FSD distribution from the Benford distribution. More specifically, the first method involves conducting inferential tests to determine the extent to which the observed FSD distribution conforms to Benford's law. The null hypothesis of the joint test is as follows:

$$H_0 : p^*(d_i) = p(d_i), \text{for } d_i \in \{1, 2, \dots, 9\}$$

where, $p^*(d_i)$ denotes the observed probability of the FSD, d_i, and $p(d_i) = \log_{10}(d_i + 1) - \log_{10}(d_i)$ is the theoretical probability of d_i under Benford's law.[6] If the null hypothesis gets rejected, the reliability of the data warrants further consideration.

We first employ two classic goodness-of-fit tests: the Pearson's Chi-square (χ^2) test and the Kuiper's modified Kolmogorov-Smirnov test. The χ^2 statistic is calculated as

$$\chi^2 \equiv N \sum_{d_i=1}^{9} \frac{[p^*(d_i) - p(d_i)]^2}{p(d_i)}$$

where, N is the total number of non-zero health expenditure reported in the analysis sample and the degree of freedom is eight. χ^2 test serves to evaluate the goodness of fit between the observed FSD distribution and the theoretical Benford distribution. However, it is sensitive to sample size, meaning that in cases of a large sample, even minor differences

[5] Accuracy of data refers to the extent to which the reported data aligns with the true data. A uniform external data source ensures standardized data records across different contexts or institutions (e.g., all levels of healthcare providers, including hospitals, clinics, and pharmacies), enabling reliable analyses and providing comprehensive information.

[6] The application of Benford's law has been extended to the first two significant digits (Barney and Schulzke, 2016). Like many others, we focus on testing the FSD mainly because of the limitation of the sample size posed by the first two-digit test. Nigrini (2012) suggested a "general rule" for testing the first two digits that at least 1000 observations are required for "good conformity" and 3000 should provide a "good" fit. Testing FSD can be performed with a sample size of around 100, which will greatly enhance the applicability of the Benford analysis. Given that in our surveys, the recall period of health expenditure data is only two or four weeks, the low incidence of health expenditure will limit the application of testing the first two digits.

between the two distributions can appear statistically significant.[7] The critical values at 1%, 5%, and 10% significant levels for χ^2 test are 20.09, 15.51, and 13.36. In line with Judge and Schechter (2009), we also employ Kuiper's modified Kolmogorov-Smirnov (V_N^*) test. This test, while similarly sensitive to sample size, accounts for the ordinality of the data. V_N^* is calculated as the sum of the two maximum deviations of the observed cumulative distribution function of d_i above and below the cumulative distribution function of Benford distribution, multiplied by an adjustment factor as

$$V_N^* \equiv \left(\max_{d_i\in\{1,2,\ldots,9\}}\left\{\sum\nolimits_{i=1}^{d_i}[p(d_i)-p^*(d_i)]\right\}+\max_{d_i\in\{1,2,\ldots,9\}}\left\{\sum\nolimits_{i=1}^{d_i}[p^*(d_i)-p(d_i)]\right\}\right)\left(N^{\frac{1}{2}}+0.155+0.24N^{-\frac{1}{2}}\right)$$

V_N^* test was initially designed to assess the goodness-of-fit of an observed distribution to a continuous distribution, while Benford distribution is categorical. This discrepancy leads to a lower p-value in the V_N^* test, rendering the test results conservative. Morrow (2014) has provided the asymptotically valid critical values for V_N^* test with eight degrees of freedom at 1%, 5%, and 10% significance levels, which are 1.58, 1.32, and 1.19, respectively. In addition to these two tests, which are designed for cases where the null distribution exhibits a linear support, we also adopt the Freedman-Watson U-square (U^2) test. This test explicitly accounts for the circular support of distributions, and it is more appropriate in the context of Benford's law, where the FSD grows from 1 to 2 … to 9 to 1 and around again. As in Qu et al. (2020), the U^2 is calculated as

$$U^2 \equiv N\sum\nolimits_{d_i=1}^{9}t_{d_i}\left\{\sum\nolimits_{i=1}^{d_i}[p^*(d_i)-p(d_i)]-\left(\sum\nolimits_{d_i=1}^{9}t_{d_i}[p^*(d_i)-p(d_i)]\right)\right\}^2$$

where, for $d_i \in \{1,2,\ldots,8\}$, $t_{d_i} = [p(d_i)+p(d_i+1)]/2$, and for $d_i = 9$, $t_{d_i} = [p(9)+p(1)]/2$. Lesperance et al. (2016) have demonstrated that the asymptotic critical values for inference using U^2 test with eight degrees of freedom are 0.304, 0.205, and 0.163 at significance levels of 1%, 5% and 10% respectively. To evaluate the conformity of the observed FSD distribution to Benford distribution, we conduct all three tests, and our empirical results show a high degree of consistency across the tests.

The other approach is to adopt appropriate statistical measures that can quantify the extent of deviation between the observed FSD distribution and the Benford distribution. Such measures are less susceptible to different sample sizes, thereby mitigate concerns about the impact of sample size on test results. Importantly, these measures enable a comparison of the degree of deviation across datasets, especially when alternative test results disagree. Following Nigrini (2012), we use the Mean Absolute Deviation (*MAD*), a widely recognized measure in the context of Benford's law. *MAD* is computed as the average of the absolute differences between the observed probability of each FSD and the probability predicted by Benford's law and it is expressed as follows

$$MAD \equiv \frac{1}{9}\sum\nolimits_{i=1}^{9}|p^*(d_i)-p(d_i)|$$

A larger *MAD* signifies a more pronounced deviation of the observed FSD distribution from the Benford distribution.[8] Notably, it's essential to acknowledge that *MAD* is not entirely independent of the sample size, unless the sample size is sufficiently large. Recognizing this limitation, Barney and Schulzke (2016) proposed another measure known as the Excess Mean Absolute Deviation (*EXMAD*). *EXMAD* accounts for sample size, applying a diminishing rate of penalization for larger samples, as illustrated in Figure A1. It is written as

$$EXMAD \equiv MAD - E(MAD|N)$$

where, for $N \le 500$, $E(MAD|N)$ can be calculated using the formula $\sum\nolimits_{d_i=1}^{9}\sum\nolimits_{j=0}^{N}\binom{N}{j}p(d_i)^j[1-p(d_i)]^{N-j}\left(\left|\frac{j}{N}-p(d_i)\right|/9\right)$. For practical applications with $N > 500$, and $E(MAD|N)$ can be approximated as $1/\sqrt{18.13N}$.[9] *EXMAD* plays a crucial role in evaluating deviations, especially when dealing with small sample sizes, in which case *MAD* is at the higher risk of report false positives. An *EXMAD* value smaller than 0 indicates a close conformity of the observed FSD distribution to Benford distribution. Conversely, a larger *EXMAD* value indicates a more significant deviation from Benford distribution. Another statistical measure of deviation proposed by Cho and Gaines (2007) is the normalized Euclidean Distance (d^*), which calculates the Euclidean distance in a nine-dimensional space occupied by an FSD vector, comparing the observed FSD distribution with Benford distribution. It is calculated as follows

$$d^* \equiv \frac{1}{M}\sqrt{\sum\nolimits_{i=1}^{9}[p^*(d_i)-p(d_i)]^2}$$

where, M is a normalization factor, defined as $M \equiv \sqrt{\sum\nolimits_{i=1}^{8}[p(d_i)]^2+[p(9)-1]^2}$, approximately equal to 1.03631. This factor assures that d^* falls within the bounded range of 0–1.[10] A higher value of d^* indicates a more substantial difference between the observed FSD distribution and Benford distribution.

2.4. Statistical tests vs. deviation measures

Statistical tests and deviation measures should be considered complementary tools for data screening purposes. Both a larger test statistic and a greater deviation measure indicate a greater divergence from Benford's law. Statistical tests are widely used in Benford studies because of their ease of use and practical interpretation. Furthermore, these tests are particularly useful for testing whether the data statistically conforms to Benford's law. For example, a χ^2 statistic smaller than 13.36 suggests a higher likelihood of data conforming to Benford's law. Deviation measures, being less sensitive to sample size, offer more informative insights when comparing data quality across datasets, especially when sample sizes vary significantly. Notably, for MAD, Nigrini (2012) proposes a close conformity threshold of less than 0.006 and a nonconformity threshold of larger than 0.015. However, these cutoff values are to be used cautiously, especially when dealing with sample sizes smaller than 1,000, as argued by Barney and Schulzke

[7] There is no universal consensus regarding the minimum sample size for χ^2 test to be considered valid in Benford analysis. Michalski and Stoltz (2013) demonstrated through simulation studies that testing Benford's law requires at least 110 data points for the test to be powerful. Dang and Owens (2020) showed that the lower bound of the sample size should be around 100 based on the asymptotic property of χ^2 distribution. We do not explicitly discuss sample size considerations in this paper, given that all our analysis samples consist of more than 110 observations.

[8] Nigrini (2012) updated experienced cut-off scores for the degrees of conformity to the Benford's distribution: close conformity ($0 < MAD \le 0.006$), acceptable conformity ($0.006 < MAD \le 0.012$), marginally acceptable conformity ($0.012 < MAD \le 0.015$), or nonconformity ($MAD > 0.015$).

[9] While Barney and Schulzke (2016) proposed the formula of $E(MAD|N)$ for the first two significant digits, it is straightforward to derive the formula for the first significant digit. The calculation of $E(MAD|N)$ for small sample size is extremely computational demanding and we calculated $E(MAD|N)$ in Python. Calculated values of $E(MAD|N)$ by sample size are listed in Table A3 and plotted in Figure A1.

[10] For details about the calculation of the normalization factor M, refer to Campanelli (2022).

(2016). Although less sensitive to sample size, no agreed-upon threshold exists to determine conformity to Benford's law for EXMAD and d^*.[11]

When comparing datasets with similar sample sizes, both smaller test statistics and deviation measures may indicate a lesser deviation from the Benford distribution. However, in cases involving significantly different sample sizes, EXMAD becomes particularly useful, as it adjusts for larger sample sizes. Given the significant variance in sample sizes between the CHNS and CFPS, we prefer EXMAD among the three deviation measures.

2.5. *Benford's law applied to health expenditure data*

While, as described in Section 2.2**,** Benford's law has been used to assess the accuracy of various economic variables, its application to health expenditure remains unexplored. To employ Benford's law for assessing health expenditure data, it is essential to justify the foundational validity assumption that the genuine data conforms to Benford's law. Indeed, accurately reported health expenditure data is likely to conform to this law. First, the accurate health expenditure data aligns with the conditions outlined in Hill (1995). The actual health expenditure is realized by complex, unobservable processes involving interactions among various stakeholders: healthcare providers, patients, their families, and insurers. These interactions are inherently stochastic and are known only by those directly involved during the healthcare process, leading to diverse data generating processes that govern the ultimate health expenditure. For example, the distribution of overall health expenditure differs between healthier, more educated patients and their sicker, less educated counterparts. Moreover, the distribution of patients' overall health expenditure varies across different departments within the same hospital and across different types of healthcare facilities. Hill (1995)'s conditions are likely to apply to the actual expenditure data of the survey sample, randomly taken from these combined distributions. Secondly, as shown by Duan et al. (1983), the positive health expenditure data are typically modeled assuming they follow a right-skewed log-normal distribution, which tends to conform to Benford's law (Fang and Chen, 2020; Scott and Fasli, 2001). This characteristic aligns with the nature of healthcare spending, where the majority of individuals incur minimal costs, while a minority faces significantly high expenses. Both the sampling process and statistical reality underscore the rationale for expecting the conformance of genuine health expenditure data to Benford's law.

In addition to theoretical rationales, we further apply Benford's law to two sets of hospital administrative records to validate the applicability of Benford analysis to actual health expenditure data, which are known for their relatively lower susceptibility to reporting errors compared to self-reported data.[12] The first dataset, collected on October 22, 2022, comes from a large regional oncology hospital in Hubei province China, detailing out-of-pocket payments for 145 hospitalized patients diagnosed with lung cancer, stomach cancer, liver cancer, breast cancer, or colorectal cancer. Among these patients, 120 recorded non-zero out-of-pocket payments, forming the basis for our Benford analysis.[13] The second dataset, comprising 2,622,129 patient-level records of total charges during the year 2012, is publicly available on the New York State Government website.[14] Table 1 presents the results of the Benford analysis on both hospital administrative datasets. As shown in Panel A, the three statistical tests fail to reject that the out-of-pocket inpatient expenditure conforms to Benford's law, indicating the conformance of genuine health expenditure data to Benford's law. In Panel B, although tests reject the conformance null hypothesis, the MAD of 0.0034, aligned with Nigrini (2012)'s cutoff scores, implies a "close conformity" of the data to Benford's law. Additionally, the significantly lower EXMAD in Panel B compared to Panel A suggests a heightened likelihood of conformity to Benford's law for the NY State hospital dataset. These findings not only suggest the conformance of genuine health expenditure data to Benford's law but also advocate for the combined utilization of statistical tests and deviation measures, particularly in scenarios characterized by large sample sizes.

2.6. *Self-reported health expenditure data with measurement error*

Like any other consumption and expenditure data derived from surveys, self-reported health expenditure data is likely to suffer from measurement error issues. Errors can be introduced into data at any stage during the survey conduction (survey designs and data collection) or by any party involved (interviewers and respondents), resulting in various types of measurement errors (Biemer, 2009). Some errors may be classical, meaning that they are uncorrelated with the true value of any variables of interest. For example, illiteracy or innumeracy can lead to significant, but random measurement errors in survey data (De Groote and Traoré, 2005). However, in far more cases, health expenditure is associated with respondent characteristics such as age, income, or family support, and thus it is more likely to contain non-classical errors (Cohen and Carlson, 1994; Hernan and Cole, 2009). Furthermore, respondents often round and heap their answers on certain values, particularly in consumption and expenditure surveys (Browning et al., 2003; Paulin and Krishnamurty, 2018). Consequently, measurement error is of great importance for studies that estimate a relationship between health expenditures and respondent characteristics or policy welfare. The presence of measurement errors will typically cause biased and inconsistent estimates of regression model parameters (see Bound et al. (2001) and Schennach (2016) for more complete reviews). In this section, we review three common types of measurement errors in health expenditure data, examine their potential correlation with estimate biases of regression parameters, and assess the performance of Benford analysis under different situations by simulation.[15]

We begin by outlining our setup. Following the measurement error model specification by Bound et al. (2001), we assume the true model is $y^* = \alpha^* + \beta^* x^* + \varepsilon$, in which we are interested in the ordinary least square estimate of β^*. We assume that the error term ε is uncorrelated with x^*. Instead of x^* and y^*, we observe x and y from the data with measurement errors μ and υ, where $x = x^* + \mu$ and $y = y^* + \upsilon$.

Case I. classical measurement error exists only in x. In this case, the model is written as $y^* = \alpha + \beta x + \varepsilon$ and the assumption that μ is uncorrelated with x^* holds. In this case, the proportional bias of the estimate β_{y^*x} is well defined as

$$\beta_{y^*x} = \beta^* \left[1 - \frac{\sigma_\mu^2}{\sigma_{x^*}^2 + \sigma_\mu^2}\right]$$

where, σ_μ^2 and $\sigma_{x^*}^2$ are variances of μ and x^*. Thus, the classical mea-

[11] Experienced cutoff values for d^* vary widely in the literature, ranging from (Cho and Gaines, 2007)'s 0.024 to (Goodman, 2016)'s 0.25, making it challenging to draw informative conclusions.

[12] As pointed out by an anonymous reviewer, it is ideal to justify the applicability of Benford's law by showing the conformity of genuine data while highlighting deviations in inaccurate data, however, the with-error version of these data is unavailable.

[13] The dataset with identifiers removed is available upon request.

[14] The dataset was available through the website: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/u4ud-w55t/data.

[15] The authors thank an anonymous referee for suggesting the examination of Benford analysis performance under different types of measurement errors, a consideration that has been overlooked in the Benford literature.

Table 1
Results of benford analysis on two hospital administrative datasets.

	Statistical Tests			Deviation Measures			*Observations*
	χ^2	*Vn**	U^2	*MAD*	*EXMAD*	*d**	
Panel A. Records from a Hubei Oncology Hospital							
OOP Inpatient Expenditure	11.67	1.0433	0.1411	0.0255	0.0042	0.1065	120
Panel B. Records from NY State Hospitals							
Total Charge	3279.26***	25.0475***	100.4362***	0.0034	0.0033	0.0152	2,622,129

Notes: * indicates 90 percent, ** indicates 95 percent, and *** indicates 99 percent significantly different from the Benford distribution.

surement error attenuates the estimate of β^*, and this attenuation bias is largely determined by σ_μ^2. In the simulation study, we generate a log-normally distributed variable x^* with a mean of 3517 and a standard deviation of 15,221 which follows the Benford distribution.[16] We draw μ from the normal distribution $N\left(0, \sigma_\mu^2\right)$ with σ_μ^2 ranging from 10 to 100 by an increment of 10. We set the sample size to 10,000. We iterate the process by 10,000 times for each σ_μ^2 value and record the mean test statistics and deviation measures from the Benford analysis.

Case II. non-classical measurement error added to y. In this case, the model is written as $y = \alpha + \beta x^* + \varepsilon$ and the assumption that υ is correlated with x^* holds. According to Blattman et al. (2016), we get the expected bias of the estimate β_{yx^*} as

$$E\left(\beta_{yx^*} - \beta^*\right) = \gamma$$

where, γ is the slope parameter in $\nu = \delta + \gamma x^* + \epsilon$ and ϵ is the error term. In the simulation study, we generate the log-normally distributed y^* with the same mean and standard deviation as in the previous simulation study. We set the sample size to 10,000. We draw x^* from the uniform distribution $U(-1, 1)$ and draw ϵ from the standard normal distribution. For simplicity, we set $\delta = 0$. We choose γ between 20 and 200 with an increment of 20. We iterate the process by 10,000 times for each γ value and record mean test statistics and deviation measures from the Benford analysis.

Case III. heaping errors in y. Another common type of non-classical measurement error for consumption and expenditure variables is the heaping error. In this case, the model of interest is still written as $y = \alpha + \beta x^* + \varepsilon$, with values of y disproportionately concentrating on certain heaped points. According to Ahmad et al. (2024), we generate y with heaping errors based on the following latent heaping process

$$y = \begin{cases} A_{j^*}, \text{if } B < \max\limits_j HP_j \\ y^*, \text{if } B \geq \max\limits_j HP_j \end{cases}$$

where, $A_1 < A_2 < \ldots < A_j < \ldots < A_k$ are the heaped points, B follows the uniform distribution between 0 and 1, $HP_j(y^*) = \exp\left(-\sqrt{\left|y^* - A_j\right|}/I\right)$ is the heaping function with $j^* = \underset{j}{\operatorname{argmax}} HP_j$, and I is the predetermined heaping intensity. While there is no analytical form for the bias of parameter estimate in the heaping case, Ahmad et al. (2024) argues that with the heaping intensity increasing, the bias in β_{yx^*} grows theoretically, which is also shown by simulation results. We generate y based on a log-normally distributed y^* with the same mean and standard deviation as in the previous simulation study. We set the sample size of 10,000. We pick 6 heaped points including 100, 500, 1,000, 5,000, 10,000, and 50,000 and choose the heaping intensity I between 0.2 and 2 with an increment of 0.2. We iterate the process by 10,000 times for each I value and record mean test statistics and deviation measures from the Benford analysis.

As shown in Fig. 1, both test statistics and deviation measures of Benford analysis increase as the measurement error and estimate bias of regression model parameters increase in all three cases. These results are robust across different cases of measurement error and test statistics. Therefore, using health expenditure data deviating from the Benford distribution will potentially take the risk of estimation bias in the regression model, and using Benford's law to identify data inaccuracy is a potentially profitable approach.

3. Health expenditure data in the CHNS and CFPS

3.1. CHNS and CFPS

The Chinese Health and Nutrition Survey (CHNS) is a longitudinal, comprehensive survey that was designed to assess the impact of health, nutrition, family planning policies, socioeconomic transformations on the health and nutrition of the Chinese population. This ongoing collaborative project, conducted by the University of North Carolina at Chapel Hill and the Chinese Center for Disease Control and Prevention, spans ten published waves (1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, 2011, and 2015). The most recent 2015 wave sampled 7319 households, comprising over 20,914 individuals from 360 communities across 15 provinces in China using a multi-stage, random cluster process.[17] The dataset includes detailed information on individual health, socioeconomic status, demographics, as well as household and community variables.

The China Family Panel Studies (CFPS) is a nationwide sociological survey project implemented by Peking University's Institute of Social Science Survey. Designed to collect data related to social, economic, demographic, educational, and health aspects, the CFPS serves as a valuable database for both academic and policy research. Commencing in 2010, the CFPS maintains follow-up assessments every two years, with data collected in six waves (2010, 2012, 2014, 2016, 2018, and 2020). In the most recent wave of 2020, the CFPS surveyed a total of 11,620 households and 28,530 individuals in 31 provinces.[18] It includes information on economic activities, education, family dynamics, population movement, and health-related variables.

[16] We set the mean and variance of the simulated sample as those of the CFPS data as in Table 2.

[17] The original 9 sample provinces and autonomous regions in the CHNS were Heilongjiang, Liaoning, Shandong, Henan, Hubei, Hunan, Jiangsu, Guizhou, and Guangxi. 3 municipalities including Beijing, Chongqing, and Shanghai were added in 2011. Another 3 provinces including Shaanxi, Yunnan, and Zhejiang were added in 2015, although data is not available yet.

[18] The original 25 sample provinces, municipalities, and autonomous regions in the CFPS were Hebei, Shanxi, Liaoning, Jilin, Heilongjiang, Jiangsu, Zhejiang, Anhui, Fujian, Jiangxi, Shandong, Henan, Guangdong, Hunan, Hubei, Sichuan, Guizhou, Yunnan, Shaanxi, Gansu, Guangxi, Beijing, Tianjin, Shanghai, and Chongqing. Three provinces and autonomous regions including Qinghai, Ningxia, and Xinjiang were added in 2012. Two provinces and autonomous regions including Hainan and Inner Mongolia were added in 2014. The autonomous region of Tibet was added in 2016. Data of Shanghai, Liaoning, Henna, Gansu, and Guangdong were provincially representative, which allows for province-level inferences and cross-province comparisons (Xie and Hu, 2014).

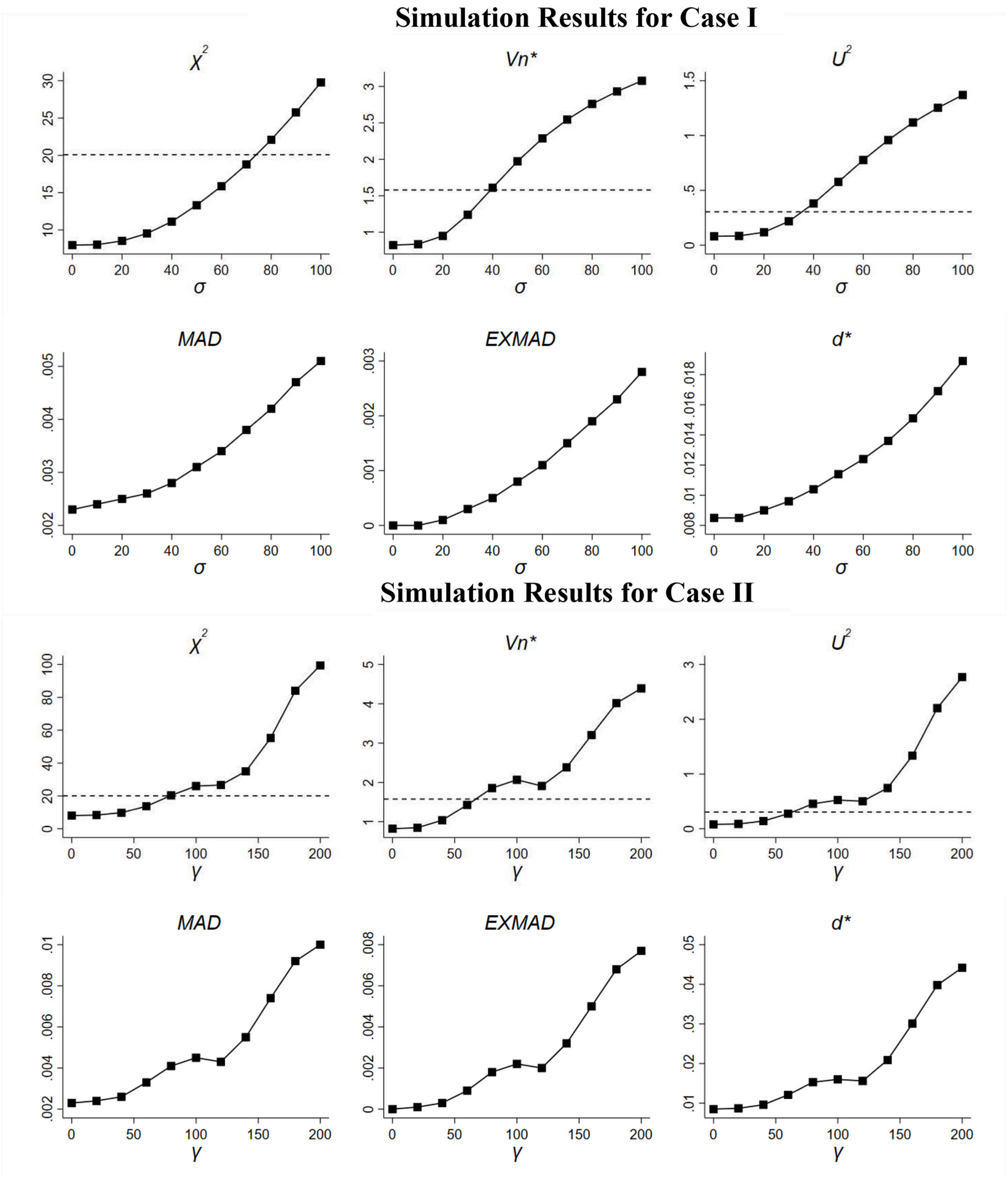

Fig. 1. Benford analysis results and estimate bias with measurement Error

Notes: The figure plots the values of the statistical tests and deviation measures for each case of measurement error. The horizontal axis represents the degree to which measurement errors contaminate the data. And the horizontal line indicates the critical value (20.09, 1.58, 0.304) of the χ^2 test, V_N^* test and U^2 test for 99% confidence level.

1-1. Simulation results for Case I

1–2. Simulation results for Case II1–3. Simulation results for Case III.

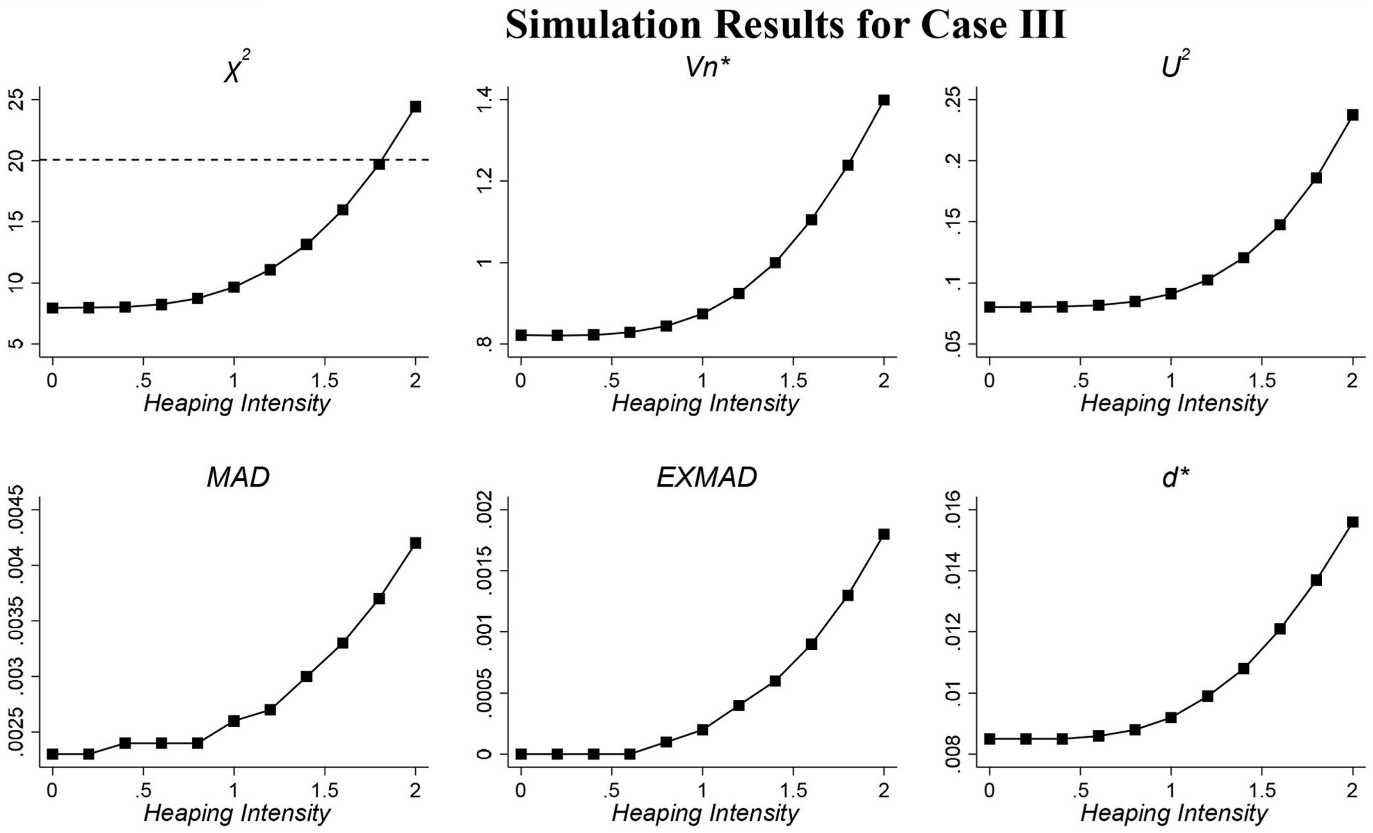

Fig. 1. (*continued*).

3.2. Survey questions

In the CHNS, respondents are asked about their expenses on health care and services in the previous four weeks if they had any health expenditure. Detailed categories encompass self-treatment, outpatient care, inpatient care, preventive care, and other disease or injury-related issues. Additionally, respondents are asked about the proportion of expenses covered by their health insurance for outpatient, inpatient, and preventive care. Respondents were also asked about the proportion in the format of percentage of the expenses covered by their health insurance, which can be used to calculate their out-of-pocket expenses in these categories. Respondents who replied “unknown” or “if insurance covers all expenses” in health expense-related questions are excluded from the data.[19]

The CFPS records an individual’s health-related expenses for the past 12 months.[20] Adult respondents were asked about the total inpatient care expenses. However, the pattern of health expenditure questions in the CFPS has evolved since 2012. In the 2010 wave, respondents only disclosed inpatient care expenses, rendering the calculation of total health expenditure infeasible. Subsequently, from 2012 onwards, respondents reported additional expenses in addition to inpatient care, enabling the estimation of total health expenditure. Information on out-of-pocket inpatient expenses is available exclusively for the 2010 and 2012 waves. Respondents who reported unknown health expenditure are dropped from the data. Detailed survey questions are listed in Table A1 and Table A2.

3.3. Variables

In this study, we focus on examining the reporting accuracy of four variables of health expenditure that are widely used in health economics literature. These variables are available in both CHNS and CFPS. The first variable we assess is *total inpatient expenditure*, inclusive of both out-of-pocket and insurer payments for inpatient care. The data of this variable is consistently available in all waves of both datasets. The second variable under examination is *out-of-pocket inpatient expenditure*, which quantifies the financial burden shouldered by patients. The data of this variable is available for all waves of the CHNS and the 2010 and 2012 waves of the CFPS.

Two additional variables, *total health expenditure* and *out-of-pocket health expenditure*, are constructed to account for overall expenses related to various types of health care and services. In the CHNS, *total health expenditure* encompasses expenses across inpatient care, outpatient care, preventive care, self-treatment, and any other expenses on disease or injury-related issues. The *out-of-pocket health expenditure* in the CHNS is derived by summing the reported out-of-pocket expenses within three available categories including inpatient care, outpatient care, and preventive care. Due to the difference in designed questions, the calculation of *total health expenditure* is slightly different in the CFPS. There are only two specific categories available in the CFPS, that is, inpatient care and other health services, and we construct *total health expenditure* by combining expenses for these two categories. The *out-of-pocket health expenditure* in the CFPS is determined by summing the

[19] Approximately 13% of respondents reported utilization of health care and services but answered “unknown” or “if insurance covers all expenses” to the question regarding the associated expenses. With these answers, we are not able to infer their exact expenses.

[20] Since the 2015 wave, respondents in the CFPS were asked questions about their health care and service utilization and expenditure “during the past 12 months” instead of “during the past year” as in previous survey waves.

reported total expenses within these two categories, as paid by the respondent and their family.[21] The congruence in survey levels and definitions of the four variables between the CHNS and CFPS enables a direct assessment of reporting accuracy of health expenditure variables between them.

3.4. Descriptive Statistics

To assess the overall accuracy of health expenditure data, we pooled data across all the waves of the CHNS and CFPS separately, excluding zero expenditures which are not applicable for Benford analysis. This cross-section data offers a substantial advantage by strengthening the statistical power of our analysis through the inclusion of a maximal number of non-zero observations. The practice of pooling data is not uncommon in the Benford literature, with Dang and Owens (2020) having surveyed a series of studies that pooled samples by year, by organization and year, or by industry and year. Pooling the data across all the waves, we get around 15,000 and 130,000 observations with positive health expenditure in the CHNS and CFPS, respectively.[22]

Descriptive statistics are presented in Table 2. As expected, the health expenditure variables in both datasets meet the prerequisites necessary for Benford analysis. The mean-to-median ratios for all variables exceed 2. This ratio is notably larger for health expenditure variables than inpatient expenditure variables, mainly because the median expenditure for respondents with any health expenditure is substantially lower than inpatient expenditure. While all variables exhibit right-skewed distributions, total expenditure is more right-skewed. Notably, there are demographic and socioeconomic differences within the analysis samples of the CHNS and CFPS. The CHNS sample appears to consist of more elderly, female, healthier, educated, and rural respondents. Moreover, the CHNS has a significantly smaller proportion of individuals covered by health insurance, reflecting the historical reality of fewer people having health insurance in earlier periods.

3.5. Simulation of conformity to Benford's law

The validity of applying Benford analysis to self-reported health expenditure data hinges upon the foundational assumption that genuine data conforms to Benford's law. While we have presented supportive arguments grounded in statistical distributions, random sampling processes, and empirical evidence from two hospital administrative datasets, we aim to further mitigate the risk of falsified inference by a Monte Carlo simulation. Following Kaiser (2019), we construct hypothetical log-normally distributed samples based on the first and second moments of health expenditure variables sourced from the CHNS and CFPS.[23] In each of the 10,000 iterations, we generate a hypothetical sample mirroring the mean, variance, and size of the original datasets, subsequently computing the χ^2 value for a Benford analysis. Fig. 2 illustrates the distribution of 10,000 calculated χ^2 values for each variable within both datasets, with a vertical line denoting the critical value (i.e., 20.09) for the 99% confidence level of the χ^2 distribution with 8 degrees of freedom. Most simulated χ^2 values fall below this critical threshold, providing robust reinforcement for the assumption that Benford analysis is applicable to genuine health expenditure data.

4. Empirical results

4.1. Main results of Benford Analysis

4.1.1. Data comparison with Benford Distribution

In Fig. 3, we plot FSD distributions of *total inpatient expenditure*, *out-of-pocket inpatient expenditure*, *total health expenditure*, and *out-of-pocket health expenditure* by dataset and we add Benford distribution for comparison.[24] An overview of the graphs suggests that the data is overall close to Benford distribution with specific patterns of FSDs. However, FSD of 5 is considerably more common than predicted by Benford distribution.[25] This is consistent with previous studies indicating that people tend to round reported numbers to certain heaped values when they cannot recall the exact amount.[26] More specifically, as shown in Table A4, the probability of observing an FSD of 5 in the context of *total health expenditure* is 0.148 within the CFPS dataset, nearly twice the expected probability according to Benford's law. FSD of 3 is generally over-reported for health expenditure data as well. FSD of 9, however, is reported less frequently than expected. The reported probabilities of *total health expenditure* and *out-of-pocket health expenditure* in the CFPS are both around 0.009, as opposed to the expected probability of 0.046 by Benford's law. FSDs of 4 and 7 are also generally reported less frequently than Benford's law.

While FSD distributions of all variables deviate from Benford's law, the deviation of the out-of-pocket expenditure variables appears to be overall smaller than the total expenditure variables. This discrepancy might be due to the complexity of memorizing and reporting total expenses. When visually comparing the FSD distributions between the CHNS and CFPS, *total inpatient expenditure*, *total health expenditure*, and *out-of-pocket health expenditure* in the CHNS are generally closer to the Benford distribution than the CFPS. Specifically for *total health expenditure* and *out-of-pocket health expenditure*, as shown in Table A4, FSD of 5 in the CFPS is more frequently reported than the CHNS by 3.7 and 4.7 percentage points, respectively. It's also worth noting that we cannot determine whether the CHNS or CFPS has more accurately reported *out-of-pocket inpatient expenditure* through visual inspection.

4.1.2. Statistical tests and deviation measures

In addition to visual analysis, we conduct statistical tests to examine whether FSD distributions of health expenditure data recorded in the CHNS and CFPS statistically conform to Benford's law, and compute deviation measures to compare the goodness-of-fit of the data to Benford distribution. The test results and deviation measures that suggest the degree of deviation are presented in Table 3. Our statistical tests reject the hypothesis that the health expenditure data in the CHNS and CFPS conform to the Benford distribution, supplementing our visual findings of inaccuracies in self-reported health expenditure data. An exception to note is that both χ^2 and U^2 test fail to reject the conformance hypothesis for *out-of-pocket inpatient expenditure* in the CHNS. Nevertheless, this

[21] Our primary focus centers on the reported incurred expenditure, as opposed to other variables, such as whether a respondent had any health expenditure. This latter variable is also of importance to researchers in the field of health economics and policy studies. The main reason is that the utilization of Benford analysis in this paper necessitates its application to the first significant (non-zero) digit(s) of reported numbers. The other reason is that the act of accurately reporting whether one has been hospitalized in the past is generally less memory-demanding compared to reporting the specific amount of expense incurred, and thus, it is likely to be more accurate.

[22] Following the prevailing approach in the Benford literature, we did not adjust the expenditure inflation. The analysis is primarily centered on assessing the accuracy of reporting, rather than on the absolute magnitude of the expenditure.

[23] Refer to Section 4 in (Kaiser, 2019) for detailed information regarding the generating process of the hypothetical log-normally distributed sample. While the variables generated in Kaiser's paper pertain to income and those in our study concern health expenditure, the underlying algorithm should remain similar.

[24] Table A4 presents detailed Benford distribution and FSD distributions of the health expenditure variables.

[25] Judge and Schechter (2009) used Benford's law to examine household agricultural production data and likewise found that the data with an FSD of 5 is much over-reported.

[26] Figure A2 and A3 plot the time trends of EXMAD measures for data in the CHNS and CFPS.

Table 2
Descriptive statistics.

	CHNS						CFPS					
Panel A. Health Expenditure Variables (in Chinese Yuan)												
	Mean	*S.D.*	*Median*	*Skewness*	*Observations*	*Mean/ Median*	*Mean*	*S.D.*	*Median*	*Skewness*	*Observations*	*Mean/ Median*
Total Inpatient Expenditure	6853	15,012	2000	5.45	1145	3.43	12,383	24,877	5000	7.23	20,163	2.48
OOP Inpatient Expenditure	5969	15,908	2000	9.13	605	2.98	6231	12,700	2600	5.81	5229	2.40
Total Health Expenditure	1766	10,109	75	9.13	15,537	23.56	3517	15,221	600	67.08	130,936	5.86
OOP Health Expenditure	1677	7535	100	10.03	4377	16.77	2626	9263	500	17.24	103,785	5.25
Panel B. Demographic and Socioeconomic Characteristics												
	Mean	*S.D.*	*Min*	*Max*	*Observations*		*Mean*	*S.D.*	*Min*	*Max*	*Observations*	
Age	44.42	23.30	0	100	15,537		40.54	22.10	0	104	130,916	
Whether Female	0.55	0.50	0	1	15,537		0.52	0.50	0	1	130,930	
Self-rated Health	2.83	0.86	1	5	7111		3.20	1.24	1	5	114,559	
Educational Years	6.52	4.44	0	18	13,765		6.25	5.09	0	18	126,049	
Whether Having Health Insurance	0.64	0.48	0	1	14,493		0.87	0.33	0	1	129,174	
Whether Rural Resident	0.60	0.49	0	1	15,537		0.54	0.50	0	1	128,010	

Notes: Self-rated health status ranges from 1 to 5 for most respondents, and a higher value indicates a worse health status. Five categories in the 2010 CFPS wave are 1-Healthy, 2-Fair, 3-Relatively Unhealthy, 4-Unhealthy, and 5-Extremely Unhealthy. Five categories in later CFPS waves are 1-Excellent, 2-Very Good, 3-Good, 4-Fair, and 5-Poor. Four categories in and before 2006 CHNS waves are 1-Excellent, 2-Good, 3-Fair, and 4-Poor. Five categories in later CHNS waves are 1-Very Good, 2-Good, 3-Fair, 4-Bad, and 5-Very Bad.

result is subject to some limitations, as both tests are highly influenced by sample size. In contrast, evidence based on V_N^* test marginally and significantly rejects the conformance hypothesis. This is in line with the understanding that V_N^* test often yields more conservative results in the context of discrete distribution. While we must interpret these test results with caution due to their sensitivity to sample size, they collectively suggest that self-reported health expenditure data in both datasets exhibit some reporting inaccuracy. Further analyses are conducted on the degree of data inaccuracy in both datasets.

Comparing the CFPS to the CHNS, all test statistics for the CHNS are much smaller, but this discrepancy can largely be attributed to the bigger sample size of the CFPS. To mitigate the impact of sample size differences, we rely on the more robust measures including *MAD*, *EXMAD*, and d^*, with a particular focus on the less sample-size-sensitive measure, *EXMAD*. The results based on *EXMAD* suggest that FSD distributions of all health expenditure variables deviate less from the Benford distribution in the CHNS. For example, the *EXMAD* for *total inpatient expenditure* is 0.0110 in the CFPS and 0.0036 in the CHNS, which is less than one-third of the CFPS. The results from all three deviation measures consistently show that the reporting accuracy of *total inpatient expenditure*, *total health expenditure*, and *out-of-pocket health expenditure* is better in the CHNS.

When assessing the extent of deviation among different variables, the results based on values of *EXMAD* show that the inpatient expenditure variables exhibit a relatively lower degree of deviation from the Benford distribution compared to the total expenditure variables. This observation is true within both datasets and for both total and out-of-pocket variables. This lends support to our surmise that the process of aggregating categorical expenses into a total cost figure may introduce an additional source of reporting inaccuracies, when contrasted with the direct reporting of inpatient expenditure.

The statistical non-conformance of health expenditure data with Benford distribution may be due to respondents rounding their hard-to-remember expenditure to heaped values. For example, a respondent is more likely to report an expense of 500 CNY rather than 480 CNY. This explanation finds support in the visual examination of the data compared to the Benford distribution. Furthermore, the more significant deviation in total expenditure than inpatient expenditure aligns with this explanation, as the process of calculating total expenditure accumulates rounding errors from each sub-category of expenses. In addition, the observed under-reported FSD of 4 indicates that 5 is more frequently reported than what would be predicted based on Benford distribution, at the cost of rounding the neighboring number.

4.2. Robustness of main results

We have demonstrated how Benford analysis identifies reporting inaccuracies in health expenditure data and compared the degree of inaccuracy using the CFPS and CHNS datasets. However, concerns may arise regarding the differences in survey scales, designs, and sample sizes, which could impact the robustness of the results and generalizability of the analysis. Although the *EXMAD* is used to mitigate the impact of sample size, it is not entirely independent of sample size. In this section, we re-conduct Benford analyses to assess the conformance of health expenditure variables to Benford distribution and compare degrees of inaccuracies of the variables between two datasets by controlling the same survey provinces and years, considering the role of recall periods, and experimenting with a set of simulation studies with the same sample size.

4.2.1. Survey area and year

The CFPS covers 31 provinces in its most recent wave, and the CHNS includes data from 12 provinces. Additionally, the CHNS data spans from 1988 to 2015, whereas the CFPS began in 2010 and continued until 2020.[27] These differences in provinces and years are associated with variations in sample demographics, local health policies, and healthcare accessibility, potentially leading to the different patterns of health expenditure and data accuracy. To assess whether the main results are influenced by survey areas and years, we conducted three sets of Benford analyses.

In the first analysis, we restrict the CFPS sample to provinces available in the CHNS, ensuring both samples represent the same survey areas. We then conduct the second analysis using the survey waves collected in the same corresponding years in both surveys to control the year effect. The survey waves we use are the 2009, 2011, and 2015 waves of the CHNS and the 2010, 2012, and 2016 waves of the CFPS. In the third analysis, we match both provinces and survey waves to control the impacts of both survey area and year on the analysis results.

[27] Another expense variable with the same recall period, available in both the CHNS and the CFPS, is the monthly housing rent. However, the reliability of this variable raises concerns in the CHNS, as the reported median housing rent is as low as 11 CNY. For this reason, housing rent is excluded from our analysis.

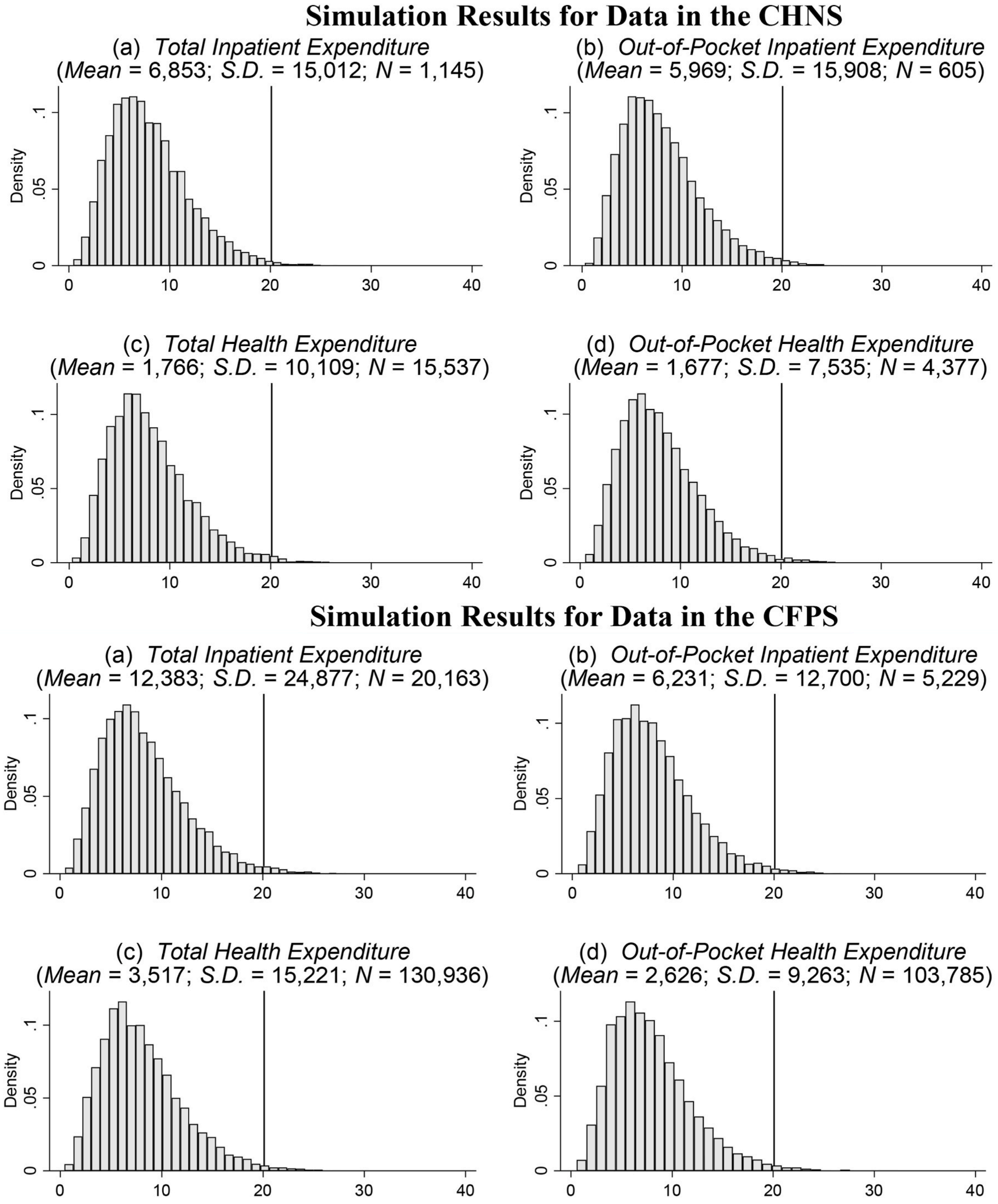

Fig. 2. Monte Carlo simulation results of χ^2 tests for hypothetical health expenditure Data

Notes: The figure graphs the simulated χ^2 values for each hypothetical health expenditure variable of the CHNS and CFPS. The horizontal axis represents the χ^2 value of each simulation, while the vertical line indicates the critical value (20.09) of the χ^2 distribution for 99% confidence interval. The 99th percentile χ^2 values of the Monte Carlo simulation for all the variables are 20.9836 (Total Inpatient Expenditure in CFPS), 20.0253 (Out-of-Pocket Inpatient Expenditure in CFPS), 20.3157 (Total Health Expenditure in CFPS), 20.3661 (Out-of-Pocket Health Expenditure in CFPS); 19.3867 (Total Inpatient Expenditure in CHNS), 20.0808 (Out-of-Pocket Inpatient Expenditure in CHNS), 20.1858 (Total Health Expenditure in CHNS), 20.6549 (Out-of-Pocket Health Expenditure in CHNS).
2-1. Simulation results for data in the CHNS
2-2. Simulation results for data in the CFPS.

The results, as presented in Table 4, suggest that statistical tests generally reject the conformance of health expenditure variables to the Benford distribution. *EXMAD* results demonstrate that the reporting accuracy of inpatient expenditure outperforms the overall health expenditure for both total and out-of-pocket measures. Results also suggest that health expenditure variables in the CHNS exhibit less deviation from the Benford distribution than those in the CFPS. This aligns with our main analysis results. Another significant takeaway is the consistent failure of all three tests to reject the *out-of-pocket inpatient expenditure* conforms to Benford distribution in the CHNS, but not in the CFPS. However, *MAD* and d^* results of the *out-of-pocket inpatient expenditure* do not favor the CHNS, and only *EXMAD* results favor the

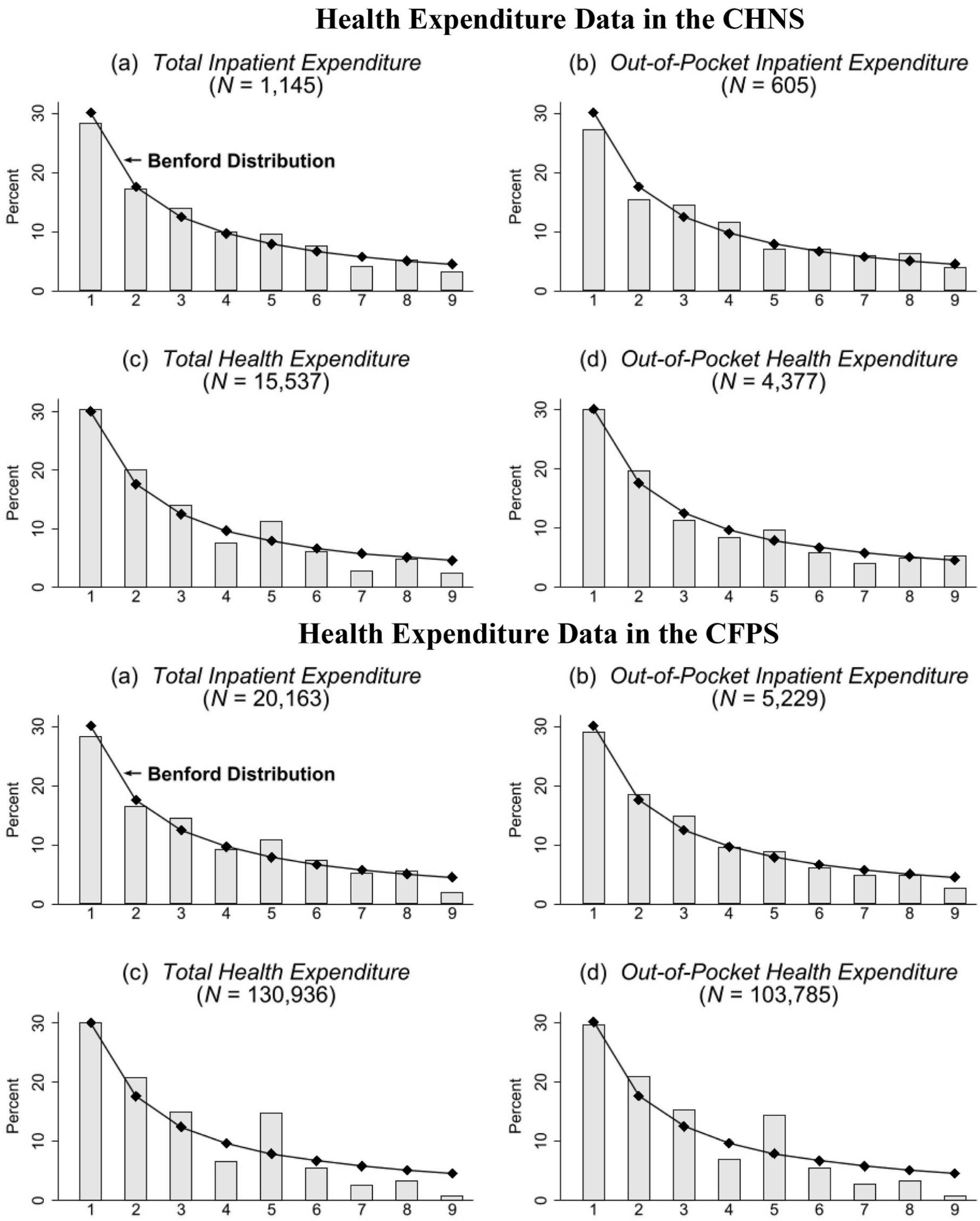

Fig. 3. Fsd distributions of health expenditure data and benford Distribution
Notes: The Benford distribution is added to all the histograms for comparison.
3-1. Health expenditure data in the CHNS
3-2. Health expenditure data in the CFPS.

CHNS. In this situation, discerning the superiority of the *out-of-pocket inpatient expenditure* between the CHNS and CFPS is challenging. This also underscores the necessity of combining statistical tests and deviation measures to ensure robust results in Benford analysis.

4.2.2. Recall period

The CFPS records the health expenditure for the past 12 months, whereas the CHNS records data for the past four weeks. The relationship between the length of the recall period and reporting accuracy has been subject to mixed findings in previous studies. For example, Clarke et al. (2008) stated that a more extended recall period increases the likelihood of recall error as a result of memory decay in reporting health care and service consumption and utilization, while Bhandari and Wagner (2006) found that data accuracy for a shorter recall period could be affected more by recall error due to the telescoping effect. To control some of the impacts of the recall period, we perform the same Benford analysis on

Table 3
Results of benford analysis on FSD.

		Statistical Tests			Deviation Measures			*Observations*
		χ^2	*Vn**	U^2	*MAD*	*EXMAD*	*d**	
CHNS	*Total Inpatient Expenditure*	19.17**	1.5342**	0.2183**	0.0105	0.0036	0.0359	1145
	OOP Inpatient Expenditure	9.57	1.2418*	0.1492	0.0129	0.0034	0.0439	605
	Total Health Expenditure	743.82***	7.1721***	5.8593***	0.0172	0.0153	0.0588	15,537
	OOP Health Expenditure	74.40***	2.0206***	0.3932***	0.0109	0.0073	0.0374	4377
CFPS	*Total Inpatient Expenditure*	649.36***	7.7234***	6.3353***	0.0140	0.0123	0.0484	20,163
	OOP Inpatient Expenditure	87.85***	3.1340***	1.2148***	0.0099	0.0066	0.0356	5229
	Total Health Expenditure	17821.54***	35.2976***	132.4932***	0.0285	0.0279	0.0977	130,936
	OOP Health Expenditure	13225.03***	32.4190***	109.2286***	0.0284	0.0276	0.0949	103,785

Notes: * indicates 90 percent, ** indicates 95 percent, and *** indicates 99 percent significantly different from the Benford distribution.

Table 4
Results of benford analysis on FSD by controlling the province and survey year.

		Statistical Tests			Deviation Measures			*Observations*
		χ^2	*Vn**	U^2	*MAD*	*EXMAD*	*d**	
Panel A. Matched Provinces								
CHNS	*Total Inpatient Expenditure*	19.17**	1.5342**	0.2183**	0.0105	0.0036	0.0359	1145
CFPS		293.89***	4.6834***	1.9337***	0.0134	0.0110	0.0454	9593
CHNS	*OOP Inpatient Expenditure*	9.57	1.2418*	0.1492	0.0129	0.0034	0.0439	605
CFPS		59.36***	2.4027***	0.6319***	0.0110	0.0063	0.0411	2455
CHNS	*Total Health Expenditure*	743.82***	7.1721***	5.8593***	0.0172	0.0153	0.0588	15,537
CFPS		7821.88***	23.8991***	57.1335***	0.0290	0.0280	0.0969	58,567
CHNS	*OOP Health Expenditure*	74.40***	2.0206***	0.3932***	0.0109	0.0073	0.0374	4377
CFPS		5850.92***	20.9476***	45.9637***	0.0281	0.0270	0.0950	45,343
Panel B. Matched Years								
CHNS	*Total Inpatient Expenditure*	14.12*	0.9853	0.1058	0.0130	0.0031	0.0428	562
CFPS		311.13***	5.7975***	3.3648***	0.0150	0.0126	0.0499	10,001
CHNS	*OOP Inpatient Expenditure*	9.02	0.9402	0.0997	0.0138	0.0026	0.0489	439
CFPS		87.85***	3.1340***	1.2148***	0.0099	0.0066	0.0356	5229
CHNS	*Total Health Expenditure*	483.74***	5.4776***	3.5737***	0.0205	0.0178	0.0705	7066
CFPS		9482.31***	26.2111***	73.8902***	0.0298	0.0289	0.1005	66,273
CHNS	*OOP Health Expenditure*	25.30***	1.2652*	0.2131**	0.0088	0.0042	0.0319	2578
CFPS		5386.85***	21.0419***	46.5206***	0.0292	0.0280	0.0966	41,160
Panel C. Matched Provinces and Years								
CHNS	*Total Inpatient Expenditure*	14.12*	0.9853	0.1058	0.0130	0.0031	0.0428	562
CFPS		153.60***	3.7586***	1.3658***	0.0152	0.0118	0.0498	4802
CHNS	*OOP Inpatient Expenditure*	9.02	0.9402	0.0997	0.0138	0.0026	0.0489	439
CFPS		59.36***	2.4027***	0.6319***	0.0110	0.0063	0.0411	2455
CHNS	*Total Health Expenditure*	483.74***	5.4776***	3.5737***	0.0205	0.0178	0.0705	7066
CFPS		4178.91***	17.2575***	30.8255***	0.0296	0.0282	0.0993	29,924
CHNS	*OOP Health Expenditure*	25.30***	1.2652*	0.2131**	0.0088	0.0042	0.0319	2578
CFPS		2416.82***	13.7074***	19.6542***	0.0291	0.0274	0.0976	17,909

Notes: Matched provinces refer to the sample, of which provinces are available in both datasets. These provinces are Liaoning, Heilongjiang, Jiangsu, Shandong, Henan, Hubei, Hunan, Guangxi, and Guizhou in the 2010 CFPS wave and Beijing, Shanghai, and Chongqing are added to later CFPS waves. Matched years refer to the sample, of which survey waves that record the health expenditure incurred in the same year of both datasets, that is, 2009, 2011 and 2015 in the CHNS and 2010, 2012 and 2016 in the CFPS. Matched provinces and years refer to the sample, of which provinces are available in both datasets and health expenditure incurred in the same years of both datasets. * indicates 90 percent, ** indicates 95 percent, and *** indicates 99 percent, significantly different from the Benford distribution.

other reported expense and income variables which have the same recall periods and are available in both datasets. The variables we examine include annual household expenses on education, events, gifts, as well as monthly individual salary and annual household income.[28] These variables, which appear to follow Benford distribution, provide valuable insights into reporting accuracy due to the distinct designs of the datasets.[29] Specifically, these variables share the identical recall periods in both datasets, mitigating concerns about the potential impact of recall periods on data accuracy. We expect that if the difference in reporting accuracy of health expenditure data is mainly due to the lengths of recall periods, other measures with the same recall periods should not clearly outperform either dataset.

The results, as depicted in Table 5, reject the hypotheses that these expense and income variables conform to the Benford distribution at the 99% confidence level. However, according to *EXMAD*, all six variables exhibit less deviation from the Benford distribution in the CHNS compared to the CFPS. *MAD* and d^* results generally show similar trends. These results suggest a preference for utilizing the CHNS dataset, particularly in health economics research that frequently relies on expenditure and income variables. The empirical observation that Benford analyses, applied to the health expenditure data and various other expense and income data, favor the CHNS, implies a possible systematic difference in the data reporting process between the CHNS and CFPS. This discrepancy may result from differences in enumerator skills, questionnaire precision, and other factors. It is challenging to rigorously test these differences within the confines of our analysis design. Nevertheless, our results help alleviate some concerns that varying recall period lengths contribute to lower reporting accuracy of

[28] Summary statistics of these variables are shown in Table A5.

[29] We perform the same bootstrap procedure for d^*, and the results are highly consistent with *MAD* and available upon request. We do not specifically on *EXMAD* in this bootstrap analysis. This is because the *EXMAD* formula consists of *MAD*, which we already addressed in this analysis, and a sample size factor that remains constant for all the variables with the same sample size.

Table 5
Benford analysis on variables with the same recall period between two datasets.

		Statistical Tests			Deviation Measures			*Observations*
		χ^2	*Vn**	U^2	*MAD*	*EXMAD*	*d**	
Panel A. Household Expense Variables								
CHNS	*Education (past year)*	250.25***	2.9179***	1.0362***	0.0193	0.0160	0.0614	5147
CFPS		8407.87***	23.8559***	67.3954***	0.0176	0.0170	0.0621	140,344
CHNS	*Events (past year)*	472.30***	5.7801***	3.0637***	0.0223	0.0194	0.0723	6312
CFPS		2062.37***	13.5372***	20.6474***	0.0274	0.0257	0.0900	18,823
CHNS	*Gifts (past year)*	6479.28***	24.1906***	60.2590***	0.0258	0.0249	0.0875	58,238
CFPS		27391.47***	49.0263***	266.7580***	0.0328	0.0323	0.1082	172,045
CHNS	*Total Expenses (past year)*	372.59***	6.5758***	8.2885***	0.0064	0.0055	0.0269	79,126
CFPS		2566.06***	22.1335***	74.6896***	0.0103	0.0098	0.0414	228,832
Panel B. Income Variables								
CHNS	*Individual Salary (past month)*	424.16***	4.9473***	2.4781***	0.0079	0.0065	0.0311	27,454
CFPS		6732.75***	36.1302***	167.7458***	0.0361	0.0350	0.1264	49,503
CHNS	*Household Income (past year)*	238.74***	6.5284***	4.8736***	0.0045	0.0039	0.0162	124,705
CFPS		4677.55***	31.8669***	118.4834***	0.0147	0.0142	0.0513	230,796

Notes: Recall periods are in parentheses. * indicates 90 percent, ** indicates 95 percent, and *** indicates 99 percent, significantly different from the Benford distribution.

health expenditure variables in the CFPS.

4.2.3. *Sample size*

There is a large difference in sample size between analyzed variables and between the CHNS and CFPS in general. For example, there are 20,163 observations with positive *total inpatient expenditure* in the CFPS, whereas there are only 1145 such observations in the CHNS. Although our study employs three deviation measures to assess the goodness-of-fit of the health expenditure data to the Benford distribution, it is important to note that these measures are not used without concern. The calculation of *EXMAD*, although designed to be less reliant on sample size, still depends on it. The adverse association between sample size and *EXMAD* amplifies as the size grows, unless the sample size is sufficiently large. To further account for the impact of sample size on our analysis results, we conduct the Benford analysis using a bootstrap procedure, and calculate *MAD* based on the bootstrapped samples with the same size in the CHNS and CFPS.[30] Similar to Michalski and Stoltz (2013), we randomly draw the same number of observations from each dataset with replacement, repeat the Benford analysis 1000 times, and calculate the mean and standard deviation of *MAD* based on the bootstrapped samples. We experiment with the sample sizes ranging from 100 to 500 with an increment of 50.

Figs. 2–4 depict the mean of *MAD* of the health expenditure variables by sample size, showing the expected decrease in *MAD* with increasing sample size. Figs. 1–4 reaffirms the observation that *total inpatient expenditure* is reported more accurately than *total health expenditure*, aligning with the main results. Across almost all sample sizes, *total inpatient expenditure*, *total health expenditure*, and *out-of-pocket health expenditure* deviate less from the Benford distribution in the CHNS. However, as shown in Figs. 2–4, the calculated *MAD* of *out-of-pocket inpatient expenditure* is slightly bigger in the CHNS. While this difference is statistically negligible with smaller sample sizes, it turns out to be significant as sample size increases. This empirical observation underscores the sensitivity of *MAD* to sample size. Bootstrap results for *MAD*, controlling for the same sample sizes for all the variables in both datasets, align with the main results. Yet, it is crucial to note that achieving identical sample sizes for variables of interest is unlikely in most empirical analyses. In our analysis sample, for example, there are 605 observations for *out-of-pocket inpatient expenditure* in the CHNS and 5229 observations in the CFPS. Despite the robustness of the *MAD* result, it does not adequately address sensitivity concerns when comparing the reporting accuracy of variables with a large discrepancy in sample sizes. In such cases, *EXMAD* stands out as a more reliable and practical deviation measure for Benford analysis.

4.3. *Indicators for data accuracy*

In this section, we explore whether potential data inaccuracy issues can be identified using readily available information within the dataset. Specifically, we investigate whether certain indicators, such as enumerators' opinions about respondents, the time interval between hospitalization and interview, and the use of proxy responses by other family members, are indicative of the reporting accuracy of health expenditure data.

4.3.1. *Enumerators' opinions*

The CFPS requested enumerators to evaluate both the credibility of respondents' answers and the perceived urgency of concluding the interview. Ratings for both questions range from 1 to 7, with 1 indicating the least credibility or urgency and 7 indicating the highest. We categorize the analysis sample into two groups based on average credibility scores: the more credible sample and the less credible sample. Similarly, we perform the same categorization based on urgency scores. Our rationale is grounded in the intuition that respondents assessed as more credible or less urgent may exhibit greater patience in answering the question, a higher degree of participation, and a more serious attitude towards the survey, ultimately leading to the improved data accuracy.

Table 6 presents a comparison of Benford analysis results between those who were judged to be above and below average in terms of level of credibility and degree of urgency in their responses. Panel A shows that test statistics for those whose answers were perceived to be above average in credibility are bigger, mainly due to the higher number of respondents falling into this category. However, intriguingly, their deviation measures, *MAD*, *EXMAD*, and *d** are also larger. If anything, the data in the group of "more credible" appears to be worse than that of "less credible". This presents a conflicting conclusion regarding the anticipated impact of numerators' assessments of respondents' credibility on data accuracy. In Panel B, three out of four variables demonstrate more accurate reporting for those who are judged to be more urgent to end the interview, as indicated by *EXMAD*. Only *total health expenditure* exhibits significantly smaller deviation measures for less urgent respondents. These results do not establish a clear link between respondents' urgency levels and data accuracy. These results suggest

[30] The CFPS provides enumerators' opinions about respondents on various other dimensions, including the levels of collaboration, intelligence, interest in the interview, concerns about the interview, and comprehension of survey questions. Analysis results of these variables are highly consistent with those of the credibility of a respondent's answers and the perceived urgency of concluding the interview presented in this section. Detailed results are available upon request.

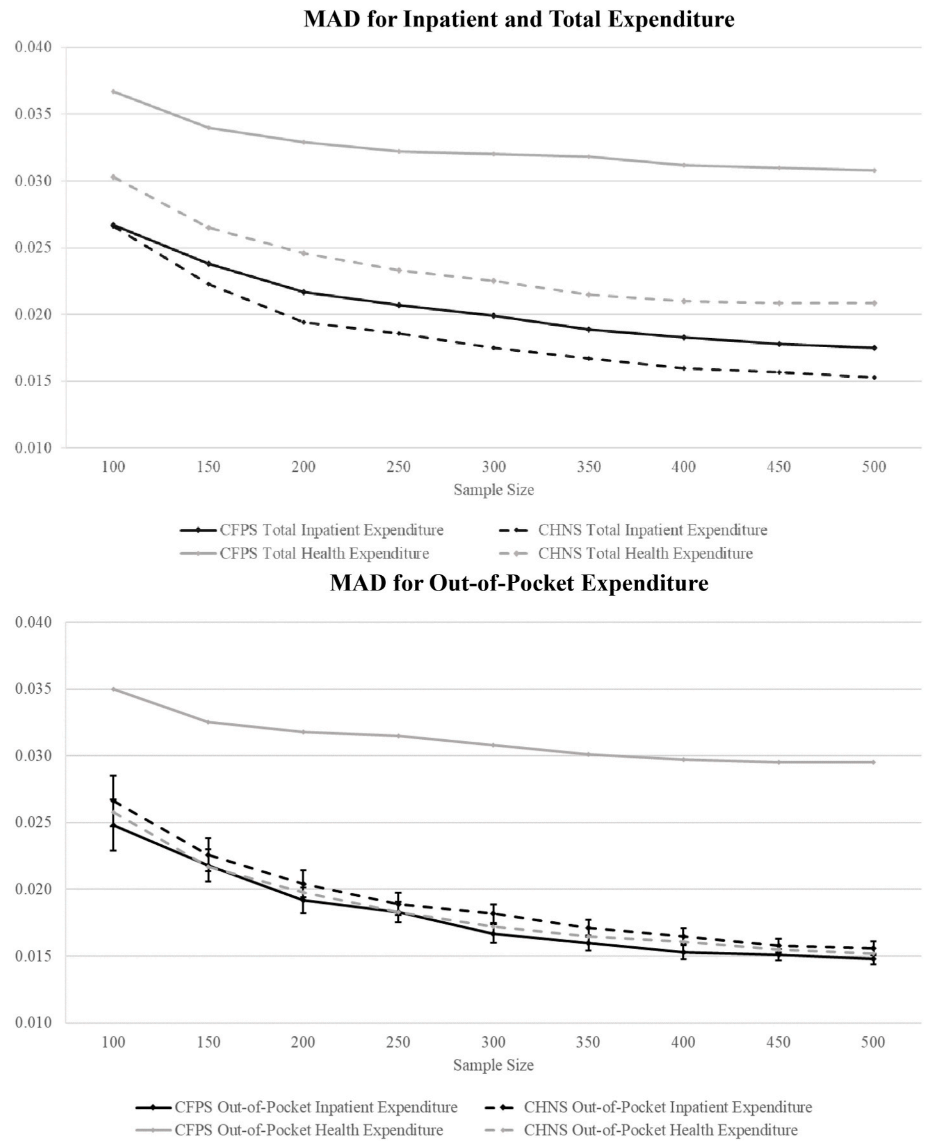

Fig. 4. Bootstrap results of MAD from benford Analysis
Notes: In order to test the statistical difference between the out-of-pocket inpatient expenditure and the out-of-pocket inpatient expenditure, 99% confidence intervals are plotted for the out-of-pocket inpatient expenditure variables.
4-1. Mad for inpatient and total Expenditure
4-2. Mad for out-of-pocket expenditure.

that enumerators' opinions may not directly influence the reporting accuracy, possibly due to the subjective nature of these assessments.[31] For example, enumerators could make themselves look good by labelling more respondents as "credible" when the answers of these respondents should not be trustworthy. Interestingly, Judge and Schechter (2009) also found that "good" crop production data marked by survey enumerators are also less in accord with Benford's law.

4.3.2. *Time intervals between hospitalization and interview*

In the 2010 and 2012 waves, the CFPS collected information on the number of times a respondent was hospitalized in the past year and the exact month of hospitalization. Examining the variation in lengths between the hospitalized month and the interview month provides us with an opportunity to assess whether a more recent hospitalization experience is associated with better accuracy in reporting health expenditure data. Each hospitalization record of a respondent is treated as a separate observation in this analysis. We categorize the sample into different analysis groups based on the time interval between hospitalization and the interview in two ways: by the average length of the interval and by every three months, ensuring comparable sample sizes in each group.

Panel A of Table 7 compares Benford analysis results for more recent incurred expenditures to the past, revealing that the more recent incurred expenditure data deviates less from the Benford distribution. Panel B breaks down the results by every three months, demonstrating that health expenditure incurred in the past recent three months conform better to Benford distribution than any other time in the past. We even fail to reject the hypothesis that *total health expenditure* conforms to Benford's law based on χ^2, V_N^*, and U^2 test results. Overall, the evidence suggests that the time interval between hospitalization and the interview might be able to serve as an informative indicator of reporting accuracy of health expenditure data.

4.3.3. *Proxy responses*

In cases where the interviewee could not be reached by any means, CFPS interviewers collected health expenditure data from other family members in the same household, known as proxy responses. These proxy

Table 6
Benford analysis results by Enumerator's opinion on respondents in CFPS.

		Statistical Tests			Deviation Measures			*Observations*
		χ^2	Vn*	U^2	*MAD*	*EXMAD*	d*	
Panel A. Credibility of Answers								
More Credible	*Total Inpatient Expenditure*	410.93***	0.0367***	4.5023***	0.0157	0.0135	0.0530	11,609
Less Credible		75.46***	0.0309**	0.7999***	0.0121	0.0075	0.0458	2533
More Credible	*OOP Inpatient Expenditure*	74.25***	2.9562***	1.0374***	0.0114	0.0077	0.0386	4044
Less Credible		22.55***	1.3394**	0.2347**	0.0093	0.0025	0.0352	1185
More Credible	*Total Health Expenditure*	10216.04***	26.9941***	76.5179***	0.0296	0.0287	0.1005	71,183
Less Credible		1749.67***	11.4560***	13.9900***	0.0284	0.0263	0.0962	13,353
More Credible	*OOP Health Expenditure*	6682.42***	23.2250***	56.3925***	0.0291	0.0281	0.0971	50,581
Less Credible		1038.09***	9.3552***	9.2617***	0.0285	0.0258	0.0950	8094
Panel B. Urgent to End the Interview								
More Urgent	*Total Inpatient Expenditure*	72.35***	2.0642***	0.6030***	0.0119	0.0066	0.0467	2028
Less Urgent		530.10***	7.3188***	5.6016***	0.0148	0.0129	0.0511	15,436
More Urgent	*OOP Inpatient Expenditure*	15.74**	0.9573	0.0776	0.0195	0.0053	0.0655	292
Less Urgent		49.48***	2.6855***	0.7559***	0.0120	0.0073	0.0421	2463
More Urgent	*Total Health Expenditure*	2030.22***	12.0163***	13.9905***	0.0303	0.0283	0.1027	13,496
Less Urgent		13713.76***	30.6769***	100.4612***	0.0281	0.0274	0.0969	102,610
More Urgent	*OOP Health Expenditure*	1483.70***	10.5626***	11.3657***	0.0291	0.0268	0.0982	10,777
Less Urgent		9813.68***	28.1082***	81.2589***	0.0282	0.0274	0.0943	78,282

Notes: * indicates 90 percent, ** indicates 95 percent, and *** indicates 99 percent, significantly different from the Benford distribution.

Table 7
Benford analysis results by the time interval between hospitalization and interview in CFPS.

		Statistical Tests			Deviation Measures			*Observations*
		χ^2	Vn*	U^2	*MAD*	*EXMAD*	d*	
Panel A. Recent vs. Past								
Recent	*Total Inpatient Expenditure*	89.44***	2.6468***	0.7583***	0.0120	0.0081	0.0408	3630
Past		81.29***	3.1158***	0.9198***	0.0132	0.0090	0.0466	3078
Recent	*Total Health Expenditure*	14.87*	1.4179**	0.2652**	0.0055	0.0014	0.0204	3315
Past		17.11**	1.7407***	0.2587**	0.0072	0.0028	0.0239	2845
Panel B. Every 3 Month								
Recent 3 Months	*Total Inpatient Expenditure*	46.23***	1.6846***	0.2828**	0.0102	0.0047	0.0388	1806
Past 3–6 Months		58.49***	2.2518***	0.5880***	0.0138	0.0083	0.0511	1824
Past 6–9 Months		48.87***	2.4390***	0.5093***	0.0154	0.0094	0.0509	1545
Past 9–12 Months		40.91***	2.2958***	0.5194***	0.0131	0.0071	0.0469	1533
Recent 3 Months	*Total Health Expenditure*	12.11	1.0765	0.1198	0.0059	0.0001	0.023	1640
Past 3–6 Months		14.23*	1.6391***	0.3062***	0.0093	0.0036	0.0344	1675
Past 6–9 Months		19.33**	1.5534**	0.2148**	0.0092	0.0029	0.0348	1427
Past 9–12 Months		12.16	1.1394	0.1785*	0.0076	0.0013	0.0302	1418

Notes: In CFPS, the 2010 and 2012 waves recorded information about the month and expenditure of each hospitalization history of respondents. Total inpatient expenditure and total health expenditure are available. Total inpatient expenditure includes all medical costs for medicine, treatment, inpatient service as well as costs of living, food, and nursing care. Total health expenditure includes both total inpatient expenditure and other expenditure on lodging, food, nursing care and other expenses. The out-of-pocket expense is not available for these detailed hospitalization records, thus, we cannot examine the out-of-pocket expenditure in this practice. * indicates 90 percent, ** indicates 95 percent, and *** indicates 99 percent, significantly different from the Benford distribution.

Table 8
Benford analysis results by proxy responses as opposed to self-reported responses in CFPS.

		Statistical Tests			Deviation Measures			*Observations*
		χ^2	Vn*	U^2	*MAD*	*EXMAD*	d*	
Panel A. Whether Proxy Response								
Self-reported	*Total Inpatient Expenditure*	484.84***	6.7777***	4.6395***	0.0145	0.0125	0.0513	13,256
Proxy		100.20***	3.3071***	1.1745***	0.0212	0.0151	0.0717	1504
Self-reported	*Total Health Expenditure*	10256.85***	26.0699***	71.7279***	0.0275	0.0266	0.0954	78,734
Proxy		2640.63***	14.4937***	21.8053***	0.0298	0.0281	0.0998	18,804
Self-reported	*OOP Health Expenditure*	9589.04***	27.6306***	78.6121***	0.0283	0.0274	0.0947	75,903
Proxy		2537.00***	14.2534***	20.9839***	0.0292	0.0275	0.0979	18,698
Panel B. Proxy for Other Adults or Children								
Adults	*Total Health Expenditure*	759.34***	8.2345***	6.8871***	0.0308	0.0276	0.1007	5529
Children		1885.59***	11.9434***	14.9801***	0.0294	0.0274	0.0997	13,275
Adults	*OOP Health Expenditure*	735.38***	8.0589***	6.5972***	0.0301	0.0270	0.0988	5514
Children		1804.84***	11.7703***	14.4361***	0.0288	0.0268	0.0977	13,184

Notes: For Panel A, the sample size for out-of-pocket inpatient expenditure answered by proxies is too small to conduct Benford analysis, and thus out-of-pocket inpatient expenditure is not analyzed. For Panel B, children's total inpatient expenditure and out-of-pocket inpatient expenditure answered by proxies are not available in the data, and thus total inpatient expenditure and out-of-pocket inpatient expenditure are not analyzed. * indicates 90 percent, ** indicates 95 percent, and *** indicates 99 percent, significantly different from the Benford distribution.

responses introduce multifaceted impacts on the data. On one hand, they provide valuable information that would otherwise be entirely missing, reducing attrition in the analysis sample. On the other hand, proxy responses may be more susceptible to significant reporting and rounding errors compared to self-reported answers, particularly in the case of retrospective data, such as detailed health expenditure, which could incur unexpectedly without the knowledge of anyone else. It is worth noting that self-reported expenditure might not always guarantee higher accuracy compared to the data reported by someone else. For example, someone who was hospitalized might not be able to remember the expenses more correctly than the family member who made the actual payment, particularly when the payment was made in cash, or the associated invoice was missing.

In Table 8, Panel A examines the relationship between proxy responses and data accuracy by comparing Benford analysis results for health expenditure data provided by respondents themselves to proxy answers from other family members. Due to the large sample size, the statistical tests reject the conformity of health expenditure data to the Benford distribution. However, all the statistical deviation measures favor self-reported expenditure data over proxy responses. Remarkably, even with a significantly larger sample size for self-reported data, *MAD*, *EXMAD*, and d^* values for self-reported data are consistently smaller, providing relatively strong evidence that self-reported expenditure is more accurate than proxy responses. In Panel B, we find some evidence that proxy responses for children tend to be more accurate than those for adults, particularly for both *total* and *out-of-pocket health expenditure*. One plausible explanation is that expenses related to children are typically paid by their parent-proxies, whereas certain expenditure incurred by adults may be only partially known to their corresponding adult-proxies.

4.3.4. Level of data: household or individual

As a longitudinal family survey, the CFPS meticulously records a comprehensive set of household information, including expenditures on healthcare. The survey involves interviewing the individual “deemed most knowledgeable about the household’s financial circumstances”, thereby capturing out-of-pocket health expenditure for the entire household over the past 12 months. Given the data availability, researchers may opt for either household or individual-level analyses depending on the specific research question. Despite that only the out-of-pocket health expenditure is available at the household level, this presents an opportunity to explore potential differences in data accuracy between household and individual levels.

Panel A of Table 9 presents the results of Benford analysis conducted on individual and household out-of-pocket health expenditure across the entire CFPS sample. Values of EXMAD as well as other deviation measures suggest that the household-level data exhibits superior performance compared to its individual-level counterpart. This finding resonates with Kaiser (2019)’s Benford analysis conclusions, which suggest that respondents tend to provide more reliable information regarding household income than individual income. In Panel B, we refine our analysis by restricting observations to instances where both individual and household health expenditure data are positive. This sample adjustment aims to mitigate any potential bias arising from the sample mismatch between individual and household data. The results remain similar with those observed in Panel A. This difference may signify the respondent’s enhanced familiarity with the household’s financial circumstances, contributing to the higher accuracy of household-level data.

5. Conclusion

This study has demonstrated how Benford’s law can be applied to assess the reporting accuracy of health expenditure data in two prominent public surveys, namely the China Health and Nutrition Survey (CHNS) and the China Family Panel Studies (CFPS). Four key health expenditure variables have been examined: *total inpatient expenditure*, *out-of-pocket inpatient expenditure*, *total health expenditure*, and *out-of-pocket health expenditure*. Results from χ^2, V_N^*, and U^2 tests indicate an overall inconsistency of these health expenditure variables with Benford’s law in both datasets. While acknowledging the sensitivity of these tests to sample size, the results still suggest the presence of reporting inaccuracies in health expenditure data that warrant further examination. Although sophisticated econometric techniques may offer solutions to mitigate measurement and reporting errors, recognizing and addressing data quality concerns at the outset of any research endeavor remains a crucial first step.

In assessing the degree of data inaccuracy, as measured by the statistical deviation from the Benford distribution, we find that expenditure data related to inpatient care is more accurately reported than that of overall health care and services. Additionally, the out-of-pocket expenditure data generally exhibits higher accuracy than the total expenditure data. Furthermore, statistical deviation measures, with a focus on *EXMAD*, demonstrate that the health expenditure data tends to be less prone to errors in the CHNS. These results hold even after accounting for differences in survey areas, years, and sample sizes between the two datasets. To eliminate the potential impact of recall periods on our results, we conduct a Benford analysis on other non-health expense and income variables with the same recall periods between two datasets, and the results also prefer the CHNS. It is noteworthy that our findings based on statistical tests and deviation measures are strongly supported by a visual comparison of FSD distributions of the variables to the Benford distribution. This visual assessment enhances the applicability of Benford’s law for future research and is highly recommended as a preliminary step before statistical tests.

Our exploration into the potential indicators of health expenditure data accuracy suggests that the enumerators’ opinions regarding respondents, including their credibility and urgency to conclude the interview, should not be relied upon as indicators of data accuracy. Counterintuitively, less credible answers deemed by survey enumerators lead to more accurate data reporting in our analysis. In contrast, the time interval between hospitalization and interviews, self-reported responses (as opposed to proxy responses), as well as household-level reported data emerge as more reliable indicators for this purpose. These findings, though specific to the CFPS and correlational rather than causal, open the possibility of identifying data reporting issues using information

Table 9
Benford analysis results by household as opposed to individual-level data in CFPS.

	Statistical Tests			Deviation Measures			*Observations*
	χ^2	*Vn**	U^2	*MAD*	*EXMAD*	*d**	
Panel A. Entire Sample							
Individual OOP Health Expenditure	13225.03***	32.4190***	109.2286***	0.0284	0.0276	0.0949	103,785
Household OOP Health Expenditure	7501.19***	23.7948***	58.2905***	0.0245	0.0237	0.0853	70,304
Panel B. Matched Individuals with Households							
Individual OOP Health Expenditure	12231.57***	31.4011***	101.3986***	0.0284	0.0276	0.0947	96,468
Household OOP Health Expenditure	4805.41***	19.1455***	37.9592***	0.0240	0.0229	0.0834	47,195

Notes: * indicates 90 percent, ** indicates 95 percent, and *** indicates 99 percent, significantly different from the Benford distribution.

within the dataset. This information is invaluable to survey designers seeking to enhance data quality and to data users aiming to access more accurate and reliable data for their research. Whenever possible, researchers should prefer the health expenditure data that adheres to Benford's law. When confronted with inaccuracies detected by Benford's Law, researchers should adopt robust strategies to mitigate their impact on subsequent analyses. This may involve rigorous data cleaning and validation processes, utilization of alternative econometric techniques resilient to data anomalies, and at least transparent reporting of encountered data quality issues. Additionally, researchers can leverage advanced statistical tools, such as Monte Carlo simulations or machine learning algorithms, to assess the sensitivity of their analyses to data inaccuracies and develop strategies for addressing them effectively.

While Benford's law offers an efficient and straightforward method for testing the reporting inaccuracy of health expenditure data and comparing the extent of data inaccuracies across datasets with different designs, certain limitations should be acknowledged. First, more evidence is needed to further validate the assumption that genuine health expenditure data conforms to Benford's Law, upon which our analysis framework is predicated. While evidence from two hospital administrative datasets and Monte Carlo simulation studies suggests the validity of this assumption, the generalizability of these results remains to be confirmed across a wider range of administrative datasets through future research. Any study utilizing Benford's law to assess data accuracy should carefully verify this assumption to avoid falsified inference. Second, it may not capture the entire impact of reporting and rounding errors of health expenditure data. For example, if two respondents A and B, have actual annual health expenditures of 5060 CNY and 4990 CNY, both are likely to report expenses of 5000 CNY during interviews. Based on the results of Benford analysis, the reported amount of respondent A contributes to the conformance to Benford's law, whereas respondent B's reported expenditure does not, despite being closer to the actual amount. Third, further investigation is needed to comprehensively explore the underlying reasons for discrepancies in reporting accuracy between datasets, such as the CHNS and the CFPS. Existing studies have identified the association between reporting accuracy and its determinants by employing regression models that project the deviation from Benford's law onto those determinants (Dang and Owens, 2020; Huang et al., 2020). However, such an approach is only applicable when there are sufficient comparison groups, as in their studies, 16,391 British charity organizations and 283 Chinese cities. However, there are only two datasets available for comparison in our analysis. Finally, the statistical tests for conformance to Benford's law are sensitive to sample size. However, we mitigate the concerns about this issue by employing deviation measures that are relatively less sensitive, although not entirely immune to large discrepancies in sample sizes. This should pose fewer concerns when comparing variables with similar sample sizes, yet real-world scenarios often involve substantial differences in sample sizes, both across variables and between datasets. Future research may investigate more robust measures, such as *EXMAD*, to minimize the impact of sample size variations on Benford analysis results. Future research should also investigate the efficacy of Benford's law in identifying data inaccuracies and quantifying the magnitude thereof, which would largely enhance the application of the Benford analysis.

Ethics statement

This study does not require ethics approval.

Funding sources

This work was supported by the National Natural Science Foundation of China [Grant No. 72203066] and the Fundamental Research Funds for the Central Universities [Program No. 2662021JGQD005and Program No. 2662021JGQD008].

CRediT authorship contribution statement

Zhuang Hao: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Xudong Zhang:** Methodology, Formal analysis. **Yuze Wang:** Supervision, Conceptualization.

Declaration of competing interest

None.

Data availability

Data will be made available on request.

Acknowledgment

The authors thank the editor, three anonymous reviewers, Canh Dang, Qu Feng, Wuyang Hu, Vidhura Tennekoon, Xiaohua Yu, Lu Zhang, and the seminar participants at Huazhong Agricultural University, University of Göttingen, and Wuhan University for their helpful comments.

Appendix A. Tables and Figures

Table A1
Survey Questions about Health Expenditure in the CHNS

Panel A. Survey Questions

Wave	Questions	Inpatient Expense	Total Expense	Out-of-Pocket
1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, 2011, 2015	During the past 4 weeks, how much money did you spend on self-treated illness or injury?		(c)	Yes
	During the past 4 weeks, how much did the outpatient treatment cost or has this outpatient treatment cost so far (including all registration fees, medicines, treatment fees, bed fees, etc.)?		(d)	Yes
	During the past 4 weeks, how much did the inpatient treatment cost or has this inpatient treatment cost so far (including all registration fees, medicines, treatment fees, bed fees, etc.)?	(a)	(e)	Yes
1989, 1991, 1993, 1997, 2000	If seeking outpatient care at a second facility, how much did the treatment cost or has the treatment cost so far?		(f)	Yes
	If seeking inpatient care at a second facility, how much did the treatment cost or has the treatment cost so far?	(b)	(g)	Yes
1989, 1991, 1993, 1997, 2000, 2004,	How much money was spent or has been spent on treating your illness or injury in addition to the costs mentioned above?		(h)	n.a.

(*continued on next page*)

Table A1 (*continued*)

Panel A. Survey Questions

Wave	Questions	Inpatient Expense	Total Expense	Out-of-Pocket
2006, 2009, 2011, 2015	In the past four weeks, how much money did you spend on preventive care if you had any?		(i)	Yes

Panel B. Constructed Variables

Wave	*Total Inpatient Expenditure*	*Total Health Expenditure*	*Out-of-Pocket Expenditure*
1989, 1991, 1993, 1997, 2000	(a)+(b)	(c)+(d)+(e)+(f)+(g)+(h)+(i)	Yes
2004, 2006, 2009, 2011, 2015	(a)	(c)+(d)+(e)+(h)+(i)	Yes

Table A2
Survey Questions about Health Expenditure in the CFPS

Panel A. Survey Questions

Wave	Questions	Inpatient Expense	Total Expense	Out-of-Pocket Share
2010 2012	In the past year, how much money in total (including pharmacy, treatment, bed, accommodation, dining, and nursing service) did you spend on all inpatient care?	(a)	(c)	Yes
2012	In the past year, how much money in total did you spend on health care? (including all expenses on inpatient care and any other expense in addition to inpatient care)		(d)	Yes
2014 2016	In the past twelve months, how much money (including pharmacy, treatment, bed, accommodation, dining, and nursing service) did you spend on inpatient care?	(b)		n.a.
2018 2020	In the past twelve months, how much money in total did you spend on health care?		(e)	Yes

Panel B. Constructed Variables

Wave	*Total Inpatient Expenditure*	*Total Health Expenditure*	*Out-of-Pocket Expenditure*
2010	(a)	n.a.	Yes
2012	(a)	(c)+(d)	Yes
2014, 2016, 2018, 2020	(b)		n.a.
2014, 2016, 2018, 2020		(e)	Yes

Table A3
Relationship between Calculated E(MAD) and Sample Size

N	E(MAD)	N	E(MAD)	N	E(MAD)
100	0.02352	300	0.01357	500	0.01051
110	0.02243	310	0.01334	550	0.01001
120	0.02142	320	0.01314	600	0.00959
130	0.02062	330	0.01292	650	0.00921
140	0.01987	340	0.01274	700	0.00888
150	0.01919	350	0.01255	750	0.00858
160	0.01858	360	0.01238	800	0.00830
170	0.01803	370	0.01221	850	0.00806
180	0.01752	380	0.01205	900	0.00783
190	0.01704	390	0.01189	950	0.00762
200	0.01661	400	0.01175	1000	0.00743
210	0.01621	410	0.01160	2000	0.00525
220	0.01585	420	0.01146	3000	0.00429
230	0.01551	430	0.01132	4000	0.00371
240	0.01514	440	0.01120	5000	0.00332
250	0.01486	450	0.01107	6000	0.00303
260	0.01457	460	0.01095	7000	0.00281
270	0.01430	470	0.01084	8000	0.00263
280	0.01404	480	0.01072	9000	0.00248
290	0.01379	490	0.01061	10,000	0.00235

Notes: The *E(MAD)* with a sample size of smaller than 500 is calculated in Python.

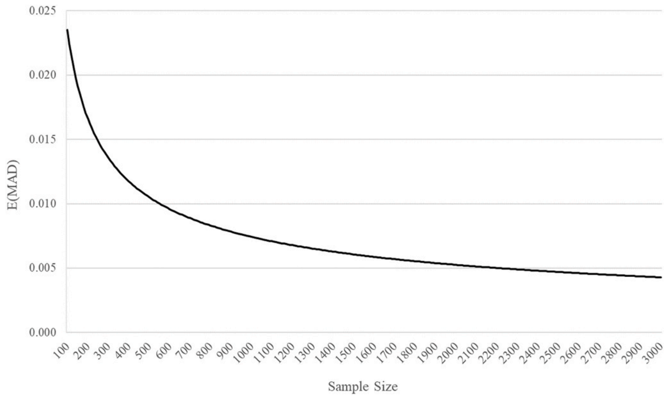

Fig. A1. Plot of E(MAD) by Sample Size

Table A4
Benford Distribution and Observed Distributions of FSD of Health Expenditure Data

		1	2	3	4	5	6	7	8	9	Observations
Panel A: Benford Distribution											
Benford Distribution		0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	
Panel B: Observed Distribution of FSD											
CHNS	*Total Inpatient Expenditure*	0.284	0.176	0.141	0.099	0.095	0.079	0.042	0.053	0.032	1145
	OOP Inpatient Expenditure	0.276	0.155	0.142	0.119	0.071	0.072	0.061	0.062	0.041	605
	Total Health Expenditure	0.304	0.202	0.140	0.077	0.113	0.061	0.030	0.049	0.025	15,537
	OOP Health Expenditure	0.301	0.198	0.114	0.085	0.098	0.058	0.041	0.051	0.054	4377
CFPS	*Total Inpatient Expenditure*	0.283	0.166	0.145	0.093	0.109	0.075	0.053	0.056	0.021	20,163
	OOP Inpatient Expenditure	0.292	0.186	0.149	0.096	0.090	0.062	0.049	0.050	0.027	5229
	Total Health Expenditure	0.302	0.209	0.150	0.066	0.148	0.055	0.027	0.034	0.009	130,936
	OOP Health Expenditure	0.296	0.210	0.153	0.070	0.145	0.055	0.028	0.034	0.009	103,785

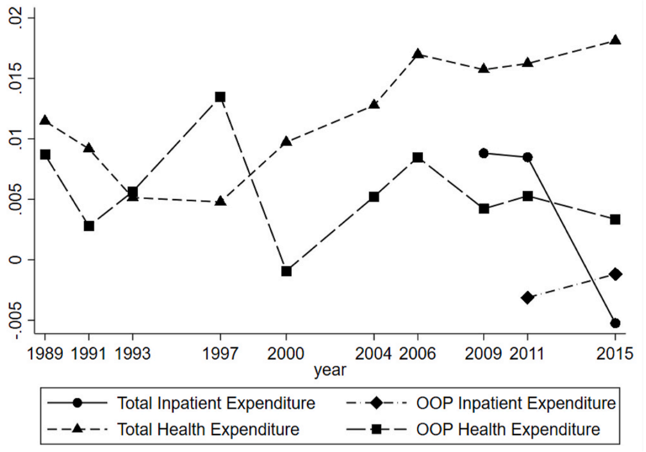

Fig. A2. EXMAD over Time in the CHNS

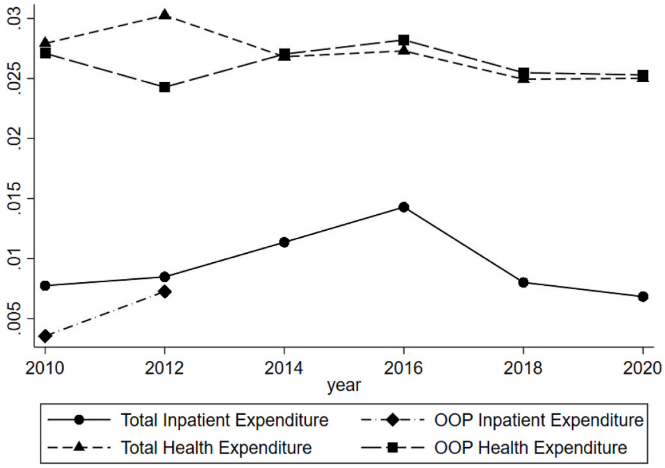

Fig. A3. EXMAD over Time in the CFPS

Table A5
Descriptive Statistics of Other Expense and Income Variables

	Mean	S.D.	Median	Skewness	Observations	Mean/Median
Panel A. CHNS (in Chinese Yuan)						
Education (past year)	3870	5491	1800	4.03	5147	2.15
Events (past year)	6667	12,176	2000	3.76	6312	3.33
Gifts (past year)	870	1761	400	15.28	58,238	2.17
Total Expenses (past year)	4701	21,169	1473	25.09	79,126	3.19
Individual Salary (past month)	1419	9393	680	91.10	27,454	2.09
Household Income (past year)	27,203	57,997	12,160	25.61	124,705	2.24
Panel B. CFPS (in Chinese Yuan)						
Education (past year)	7197	10,902	4000	9.35	140,344	1.80
Events (past year)	25,476	51,495	10,000	12.45	18,823	2.55
Gifts (past year)	4075	6296	2000	9.45	172,045	2.04
Total Expenses (past year)	62,382	97,551	40,000	20.89	228,832	1.56
Individual Salary (past month)	7210	14,235	3000	16.83	49,503	2.40
Household Income (past year)	68,528	149,562	44,000	30.2	230,796	1.56

References

Abate, G.T., De Brauw, A., Hirvonen, K., Wolle, A., 2023. Measuring consumption over the phone: evidence from a survey experiment in urban Ethiopia. J. Dev. Econ. 161, 103026 https://doi.org/10.1016/j.jdeveco.2022.103026.

Ahmad, A.S., Al-Hassan, M., Hussain, H.Y., Juber, N.F., Kiwanuka, F.N., Hag-Ali, M., Ali, R., 2024. A method of correction for heaping error in the variables using validation data. Stat. Pap. 65 (2), 687–704. https://doi.org/10.1007/s00362-023-01405-4.

Barney, B.J., Schulzke, K.S., 2016. Moderating "cry wolf" events with excess MAD in Benford's Law research and practice. Journal of Forensic Accounting Research 1 (1), A66–A90.

Benford, F., 1938. The law of anomalous numbers. In: Proceedings of the American Philosophical Society, vol. 78. JSTOR, pp. 551–572, 4.

Bhandari, A., Wagner, T., 2006. Self-reported utilization of health care services: improving measurement and accuracy. Med. Care Res. Rev. 63 (2), 217–235. https://doi.org/10.1177/1077558705285298.

Biemer, P., 2009. Chapter 12—measurement errors in sample surveys. In: Rao, C.R. (Ed.), Handbook of Statistics, vol. 29. Elsevier, pp. 281–315. https://doi.org/10.1016/S0169-7161(08)00012-6.

Blattman, C., Jamison, J., Koroknay-Palicz, T., Rodrigues, K., Sheridan, M., 2016. Measuring the measurement error: a method to qualitatively validate survey data. J. Dev. Econ. 120, 99–112. https://doi.org/10.1016/j.jdeveco.2016.01.005.

Bound, J., Brown, C., Mathiowetz, N., 2001. Chapter 59—measurement error in survey data. In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 5. Elsevier, pp. 3705–3843. https://doi.org/10.1016/S1573-4412(01)05012-7.

Boyle, J., 1994. An application of fourier series to the most significant digit problem. Am. Math. Mon. 101 (9), 879–886. https://doi.org/10.1080/00029890.1994.11997041.

Browning, M., Crossley, T.F., Weber, G., 2003. Asking consumption questions in general purpose surveys. Econ. J. 113 (491), F540–F567. https://doi.org/10.1046/j.0013-0133.2003.00168.x.

Campanelli, L., 2022. On the Euclidean distance statistic of Benford's law. Commun. Stat. Theor. Methods 1–24. https://doi.org/10.1080/03610926.2022.2082480.

Chen, X., Hong, H., Nekipelov, D., 2011. Nonlinear models of measurement errors. J. Econ. Lit. 49 (4), 901–937. https://doi.org/10.1257/jel.49.4.901.

Cho, W.K.T., Gaines, B.J., 2007. Breaking the (Benford) law: statistical fraud detection in campaign finance. Am. Statistician 61 (3), 218–223.

Clarke, P.M., Fiebig, D.G., Gerdtham, U.-G., 2008. Optimal recall length in survey design. J. Health Econ. 27 (5), 1275–1284. https://doi.org/10.1016/j.jhealeco.2008.05.012.

Clementi, F., Gallegati, M., 2005. Pareto's law of income distribution: evidence for Germany, the United Kingdom, and the United States. In: Chatterjee, A., Yarlagadda, S., Chakrabarti, B.K. (Eds.), Econophysics of Wealth Distributions: Econophys-Kolkata I. Springer, Milan, pp. 3–14. https://doi.org/10.1007/88-470-0389-X_1.

Cohen, S.B., Carlson, B.L., 1994. A comparison of household and medical provider reported expenditures in the 1987 NMES. J. Off. Stat. 10 (1), 3. Publicly Available Content Database.

Dang, C.T., Owens, T., 2020. Does transparency come at the cost of charitable services? Evidence from investigating British charities. J. Econ. Behav. Organ. 172, 314–343. https://doi.org/10.1016/j.jebo.2020.02.020.

De Groote, H., Traoré, O., 2005. The cost of accuracy in crop area estimation. Agric. Syst. 84 (1), 21–38. https://doi.org/10.1016/j.agsy.2004.06.008.

Duan, N., Manning, W.G., Morris, C.N., Newhouse, J.P., 1983. A comparison of alternative models for the demand for medical care. J. Bus. Econ. Stat. 1 (2), 115. https://doi.org/10.2307/1391852.

Durtschi, C., Hillison, W., Pacini, C., 2004. The effective use of Benford's Law to assist in detecting fraud in accounting data. J. Forensic Account. 5, 17–34.

Fang, G., Chen, Q., 2020. Several common probability distributions obey Benford's law. Phys. Stat. Mech. Appl. 540, 123129 https://doi.org/10.1016/j.physa.2019.123129.

Feng, J., Lou, P., Yu, Y., 2015. Health care expenditure over life cycle in the people's Republic of China. Asian Dev. Rev. Stud. Asian Pac. Econ. Issues 32 (1), 167–195. https://doi.org/10.1162/ADEV_a_00041.

Fu, H., Ge, R., Huang, J., Shi, X., 2022. The effect of education on health and health behaviors: evidence from the college enrollment expansion in China. China Econ. Rev. 72, 101768 https://doi.org/10.1016/j.chieco.2022.101768.

Gonzalez-Garcia, J., Pastor, G., 2009. Benford's Law and Macroeconomic Data Quality.

Goodman, W., 2016. The promises and pitfalls of Benford's law. Significance 13 (3), 38–41. https://doi.org/10.1111/j.1740-9713.2016.00919.x.

He, H., Nolen, P.J., 2019. The effect of health insurance reform: evidence from China. China Econ. Rev. 53, 168–179. https://doi.org/10.1016/j.chieco.2018.08.013.

Hernan, M.A., Cole, S.R., 2009. Invited commentary: causal diagrams and measurement bias. Am. J. Epidemiol. 170 (8), 959–962. https://doi.org/10.1093/aje/kwp293.

Hill, T.P., 1995. A statistical derivation of the significant-digit law. Stat. Sci. 10 (4), 354–363. JSTOR.

Hsieh, C.-R., Qin, X., 2018. Depression hurts, depression costs: the medical spending attributable to depression and depressive symptoms in China. Health Econ. 27 (3), 525–544. https://doi.org/10.1002/hec.3604.

Huang, F., Gan, L., 2017. The impacts of China's urban employee basic medical insurance on healthcare expenditures and health outcomes: impacts of China's urban employee basic medical insurance. Health Econ. 26 (2), 149–163. https://doi.org/10.1002/hec.3281.

Huang, W., Liu, H., 2023. Early childhood exposure to health insurance and adolescent outcomes: evidence from rural China. J. Dev. Econ. 160, 102925 https://doi.org/10.1016/j.jdeveco.2022.102925.

Huang, Y., Niu, Z., Yang, C., 2020. Testing firm-level data quality in China against Benford's Law. Econ. Lett. 192, 109182 https://doi.org/10.1016/j.econlet.2020.109182.

Judge, G., Schechter, L., 2009. Detecting problems in survey data using Benford's law. J. Hum. Resour. 44 (1).

Kaiser, M., 2019. BENFORD'S law as an indicator of survey reliability—can we trust our data? J. Econ. Surv. 33 (5), 1602–1618. https://doi.org/10.1111/joes.12338.

Kjellsson, G., Clarke, P., Gerdtham, U.-G., 2014. Forgetting to remember or remembering to forget: a study of the recall period length in health care survey questions. J. Health Econ. 35, 34–46. https://doi.org/10.1016/j.jhealeco.2014.01.007.

Lavado, R.F., Brooks, B.P., Hanlon, M., 2013. Estimating health expenditure shares from household surveys. Bull. World Health Organ. 91 (7), 519–524C. https://doi.org/10.2471/BLT.12.115535.

Lei, X., Lin, W., 2009. The New Cooperative Medical Scheme in rural China: does more coverage mean more service and better health? Health Econ. 18 (S2), S25–S46. https://doi.org/10.1002/hec.1501.

Lesperance, M., Reed, W.J., Stephens, M.A., Tsao, C., Wilton, B., 2016. Assessing conformance with Benford's law: goodness-of-fit tests and simultaneous confidence intervals. PLoS One 11 (3), e0151235. https://doi.org/10.1371/journal.pone.0151235.

Li, X., Smyth, R., Yao, Y., 2023. Extreme temperatures and out-of-pocket medical expenditure: evidence from China. China Econ. Rev. 77, 101894 https://doi.org/10.1016/j.chieco.2022.101894.

Liu, H., Zhao, Z., 2014. Does health insurance matter? Evidence from China's urban resident basic medical insurance. J. Comp. Econ. 42 (4), 1007–1020. https://doi.org/10.1016/j.jce.2014.02.003.

Liu, J., Shi, L., Khan, M., Xu, L., Wang, L., 2012. Trends of out-of-pocket expenditure for influenza in China health and nutrition survey during 1989–2006. Int. J. Publ. Health 57 (1), 193–198. https://doi.org/10.1007/s00038-011-0251-y.

Michalski, T., Stoltz, G., 2013. Do countries falsify economic data strategically? Some evidence that they might. Rev. Econ. Stat. 95 (2), 591–616. https://doi.org/10.1162/REST_a_00274.

Miller, S.J. (Ed.), 2015. Benford's Law. Princeton University Press; JSTOR. http://www.jstor.org/stable/j.ctt1dr358t.

Morrow, J., 2014. Benford's Law, Families of Distributions and a Test Basis.

Nigrini, M.J., 1996. A taxpayer compliance application of Benford's Law. J. Am. Taxat. Assoc. 18 (1), 72. ABI/INFORM Collection.

Nigrini, M.J., 2012. Benford's Law: applications for forensic accounting. In: Auditing, and Fraud Detection, vol. 586. John Wiley & Sons.

Paulin, G., Krishnamurty, P., 2018. Consumer expenditure surveys methods symposium and microdata users' workshop, july 18–21, 2017. Mon. Labor Rev. https://doi.org/10.21916/mlr.2018.15.

Qu, H., Steinberg, R., Burger, R., 2020. Abiding by the law? Using Benford's law to examine the accuracy of nonprofit financial reports. Nonprofit Voluntary Sect. Q. 49 (3), 548–570. https://doi.org/10.1177/0899764019881510.

Rosenman, R., Tennekoon, V., Hill, L.G., 2011. Measuring bias in self-reported data. Int. J. Behav. Healthc. Res. 2 (4), 320–332. https://doi.org/10.1504/IJBHR.2011.043414.

Schennach, S.M., 2016. Recent advances in the measurement error literature. Annual Review of Economics 8 (1), 341–377. https://doi.org/10.1146/annurev-economics-080315-015058.

Schräpler, J.-P., 2011. Benford's law as an instrument for fraud detection in surveys using the data of the socio-economic Panel (SOEP). Jahrb. Natl. Stat. 231 (5–6), 685–718. https://doi.org/10.1515/jbnst-2011-5-609.

Scott, P.D., Fasli, M., 2001. *Benford's Law: an Empirical Investigation And a Novel Explanation* [CSM-349. University of Essex, Colchester.

Shi, L., Smit, E., Luck, J., 2021. Panel survey estimation of the impact of urbanization in China: does level of urbanization affect healthcare expenditure, utilization or healthcare seeking behavior? Chin. Econ. 54 (3), 145–156. https://doi.org/10.1080/10971475.2020.1848472.

Si, X., Chu, F.-L., 2022. The impact of the Public Pension Program on the elderly's medical expenditures: a regression discontinuity approach. J. Appl. Econ. 25 (1), 178–196. https://doi.org/10.1080/15140326.2021.2021748.

Sun, J.Y., 2020. Welfare consequences of access to health insurance for rural households: evidence from the New Cooperative Medical Scheme in China. Health Econ. 29 (3), 337–352. https://doi.org/10.1002/hec.3985.

Villas-Boas, S.B., Fu, Q., Judge, G., 2017. Benford's law and the FSD distribution of economic behavioral micro data. Phys. Stat. Mech. Appl. 486, 711–719. https://doi.org/10.1016/j.physa.2017.05.093.

Wallace, W.A., 2002. Assessing the quality of data used for benchmarking and decision-making. J. Govern. Financ. Manag. 51 (3), 16–22. ABI/INFORM Collection.

Xie, Y., Hu, J., 2014. An Introduction to the China Family Panel Studies (CFPS).

Xu, K., Ravndal, F., Evans, D.B., Carrin, G., 2009. Assessing the reliability of household expenditure data: results of the world health survey. Health Pol. 91 (3), 297–305. https://doi.org/10.1016/j.healthpol.2009.01.002.

Yip, W., Fu, H., Chen, A.T., Zhai, T., Jian, W., Xu, R., Pan, J., Hu, M., Zhou, Z., Chen, Q., Mao, W., Sun, Q., Chen, W., 2019. 10 years of health-care reform in China: progress and gaps in universal health coverage. Lancet 394 (10204), 1192–1204. https://doi.org/10.1016/S0140-6736(19)32136-1.

Zhang, Y., Vanneste, J., Xu, J., Liu, X., 2019. Critical Illness Insurance to alleviate catastrophic health expenditures: new evidence from China. International Journal of Health Economics and Management 19 (2), 193–212. https://doi.org/10.1007/s10754-018-9252-1.

Zhao, W., 2019. Does health insurance promote people's consumption? New evidence from China. China Econ. Rev. 53, 65–86. https://doi.org/10.1016/j.chieco.2018.08.007.