Adjusting Misreported Count Data in Forensic Analysis

A.E. Rodriguez[1]

June 1, 2024

Abstract

In instances where count data disclosed in litigation can be characterized as susceptible to errors from deliberate alteration or non-statistical errors such as behavioral and judgment biases, the correct approach may be to adjust the series before determining related damages.

We propose to characterize the impugned series as a mixture conformed by two constituent data generating processes. This mixture can be estimated to obtain the latent, adjusted series. We show how to estimate the mixture via one of several open source R package available for the task.

To our knowledge this approach is not commonly deployed in forensic practice. We examine the feasibility and practicality of deploying these models either in support or to rebut forensic expert analysis.

Keywords: mixture modeling, anomaly detection, forensic analysis, label noise.

JEL Codes: *C52, C53, O31*

[*] Pompea College of Business, University of New Haven. Rodriguez is corresponding author: arodriguez@newhaven.edu.

> “Nothing can be known without there being an appropriate “instrument” in the makeup of the knower."
>
> *E.F. Schumacher*

Introduction

Instances of reported counts can be misreported reflecting deliberate equivocation, systemic error, arithmetic error, or judgmental bias. Data such as insurance or Medicare claims, excess deaths, fire starts, medical visits, or products consumed are countable quantities tallied via non-negative integers. Realizations of count data are typically concentrated on a few discrete, non-negative values resulting in asymmetric, positively-skewed distribution functions. Is it possible to correct or adjust counts data such as insurance or Medicaid claims in instances without resorting to costly forensic audits?

The task is to separate the latent, real claims counts from the actual, misreported ones; and, where appropriate, the rate, or incidence of the mishap (Pararai, Famoye, & Lee, 2010). Extracting the latent, real claims is obviously critical for the correct appraisal of any pecuniary damages. Damages calculations based on an inflated or deflated base series result in improper compensation amounts.

The literature on adjusting time series with misreported data in general, is considerable (Schennach, 2022). Available work can be distinguished between continuous and discrete; the latter includes count data.

This paper provides a review of the literature on algorithms for correcting the misreporting of counts with a twofold audience in mind. First, methods more suitable for forensic analysis. And second artificial intelligence and machine learning

By “suitable for forensic analysis” We refer and emphasize two things: parsimony and clarity. Parsimony and clarity speak to the complexity of the underlying parametric model and the interpretation of the estimated parameters. The two objectives are intertwined. Forensic work is conducted with the trier-of-fact in mind. A methodology that is sound and easy to explain will be more appealing than one with unappealing sophistication. We hold that model-based clustering conveys a more theoretically sound approach to forensic settings. Ad-hoc non-parametric models cannot be grounded on the factual elements of the litigated matters at hand – and thus appear less robust by comparison.

The European Union's Digital Markets Act (DMA) and especially the General Data Protection Regulation (GDPR) have institutionalized a preference for more interpretable predictions derived from classification algorithms. Note that this is not solely a EU matter. This particular trend will probably be hastened along by US State data protection legislation; Connecticut, in fact, is the latest of a spate of US State legislatures to launch a GDPR-lite version on July 1st of 2023.[2]

In this paper, we show how a hypothetical time series count representing Medicaid insurance claims, can be conceptualized as a composite series based on two constituent, latent generating processes. Thus, it is possible to estimate the constituent distributions of disclosed claims data. Once the correct data generating distributions are estimated we show how damages can be overestimated.

There are several libraries (or packages) available in R to estimate the DGPs underscoring the mixture.[3] We provide a roadmap to *mclust*, to provide an estimate the parameters of the latent series and to determine the rate of "pilfering." We show this as follows.

[2] See, The Connecticut Data Privacy Act. For related commentary on this matter see my post.

[3] See Appendix 1; the full field of available libraries in R is massive: e.g. viz. https://cran.r-project.org/web/views/Cluster.html.

In the next section we offer a succinct review of the literature on the misclassification of counts. Much of the existing work relies on Maximum Likelihood methods. And whereas the mixture model framework we propose here is amenable to treatment via maximum likelihood, the procedure itself is sufficiently complex to constitute a challenge when explaining its workings in a legal setting.

We then provide a simple hypothetical case study using synthetic data to illustrate how to use mclust for purposes of estimating the mixture model parameters. This is a roadmap to show how to adjust the reported, proffered series to account for the "adjusting." We also show how to identify the specific instances where the proffered claims data was "retouched."

A natural concern is to consider how sensitive results are to variation in the relevant parameters. Put differently, how good is the recommended approach? Accordingly, more broader simulation analysis is conducted to examine the accuracy of the results and their sensitivity to changes in the assumptions. Results are discussed. The last section offers concluding comments.

Misreported Count Data: A Review of the Literature

Concerns over misreporting of count data occur across all domains including, for example, health insurance, demographics, accident investigations, immigration, higher education, surveys and polling, epidemiology, criminology, production, auditing and assurance. In all these instances, reconstituting the latent data series is of primary interest (Li, Trivedi, & Guo, 2003) (Neubauer, Djuras, & Friedl, 2011) (Wood, Donnell, & Fariss, 2016) (Nigrini M. , 2022) (Stamey & Young, 2005).

Various factors could play a role in establishing a presumption of misreporting. Errors could result from various reasons including deliberate intent or the result of cognitive bias (Ioannidis, 2021) (Brody, DeZoort, Gupta, & Hood, 2022) (Rodriguez & Kucsma, 2023) (Harvin & Killey, 2021). The number of methods set forth to address concerns over misreporting of count data are numerous (Contzen, De Pasquale, & Mosler, 2015).

A favorite approach to a proposed estimation model is maximum likelihood. Maximum likelihood methods are commonly used to estimate any number of models across many domains (Ward & Alhlquist, 2018). ML is quite capable of estimating the constituent DGP parameters of the mixture model proposed here. In fact, under the hood, all the libraries examined for this paper utilize maximum likelihood.

Maximum likelihood estimates are consistent, unbiased, and efficient, all desirable properties. However, ML methods and the resulting estimates vary idiosyncratically across many dimensions rendering each capable of influencing estimates and thus carrying with it the potential of becoming a veto point. For instance, given that the function being optimized is non-linear, it is impossible to avoid the likelihood of arriving at a suboptimal outcome rather than the global outcome; maximum likelihood solvers are susceptible to starting values. Thus, it is not uncommon for different solvers to arrive at differing parameter estimates. The estimated parameter obtained via ML is a solution to a mathematical model – not a stochastic one. To obtain the standard deviations required for statistical testing – and a key Daubert factor - requires further numerical processing;[4]

[4] Per Daubert's factors, an expert witness' method must have an acceptable "rate of error" when considering possible random or systemic

various alternative numerical solution algorithms exist for this task in turn – again capable of resulting in differences among experts. Susceptibility of this sort hampers the robustness of any expert report relying on ML.

But-For Counts and Mixture Models

There are numerous clustering algorithms deployed in the detection of fraudulent claims. These methods group similar transactions or claims together based on their characteristics. Collectively, they are well known to effectively identify latent patterns or clusters that may be forensically flagged (Wei, Cho, Vasarhelyi, & Te-Wierik, 2024) (Huang, Zheng, Li, & Che, 2024).

Mixture modeling entails a probabilistic model deployed to detect subpopulations within a broader domain. Although finite mixture models are well known and routinely used there have been little applications in forensic economics, accounting, and financial practice.

Denote the g-components mixture model by

$$f(x;\ \Phi) = \sum_{j=1}^{g} \left(\pi_i f_j(x;\ \theta_j\right)$$

Where *$f(x;\ \Phi)$* is the probability density function of the mixture model; *$f_j(x;\ \theta)$* is the probability density function of the *j*th component of the mixture model; π_i is the proportion of the *j*th component; θ_i is the parameter of the *j*th component which can be a scale or a vectors; $\Phi = (\pi_i, \theta_i.,$ ..., $\pi_g, \theta_g)$ is a vector of all the parameters in the mixture model; and *g* is the total number of components in the

error. Methods that can show “general acceptance” or “peer review” constitute generally acceptable proxies.

mixture model. The task is to estimate the parameters of the individual distribution.

Miscounted Claims as Mixture Model

We generate a hypothetical series representing some proffered monthly insurance claims data over a period of 100 months. There is evidence that the reported data may have been systematically overcounted. Accordingly, the forensic expert is tasked with determining the underlying latent series representing the corrected claims series and the incidence or rate of manipulation of individual claims.

We hold that the claims data is the result of an alteration of an actual, "latent" series. At various points in the series the "actual" results are replaced by an inflated number resulting in a reported series with a higher average number of claims relative to the actual.

To model the incidence of "fudging" via a data generating process we need a higher, claims amount that is misreported - in lieu of the actual amount – "the fudge account."

To demonstrate this approach, we generate a simulated "reported" claims series. The incidence of "fudging" is represented by a Bernoulli distribution.

$$X \sim Bernoulli(p)$$

Where p is the probability of success (i.e., X = 1). In this instance p is the rate at which cheating occurs. The higher, claims amount that is misreported - in lieu of the actual amount - is drawn from a poisson distribution.

$$f(y) = \frac{\lambda^y}{y!} e^{-\lambda}$$

Where the mean of the bogus, reported poisson distribution is λ; λ_1 represents the mean of the true, latent series and λ_2 the mean of the higher series which is used as replacement. The resulting claims series is thus a mixture of the constituent series where the mixing proportion is unknown.

In this scenario λ_1 represents the means of the reported claims. The unknown λ_2 represents the means of the latent, unknown and corrected real claims counts.

Note that the act of "pilfering" or "fudging" creates a series of anomalies which cluster. These clusters can be identified. The identification of clusters, or clustering is routinely used in classification tasks. A clustering solution separates the data into different and distinct classes.

For this task many tools exist (Xu & Tian, 2015). Since we require not only the parameters of the underlying DGPs but also the classification of the claims associated with each DGP – we require a library that accomplishes both. Fortunately, there are many available.

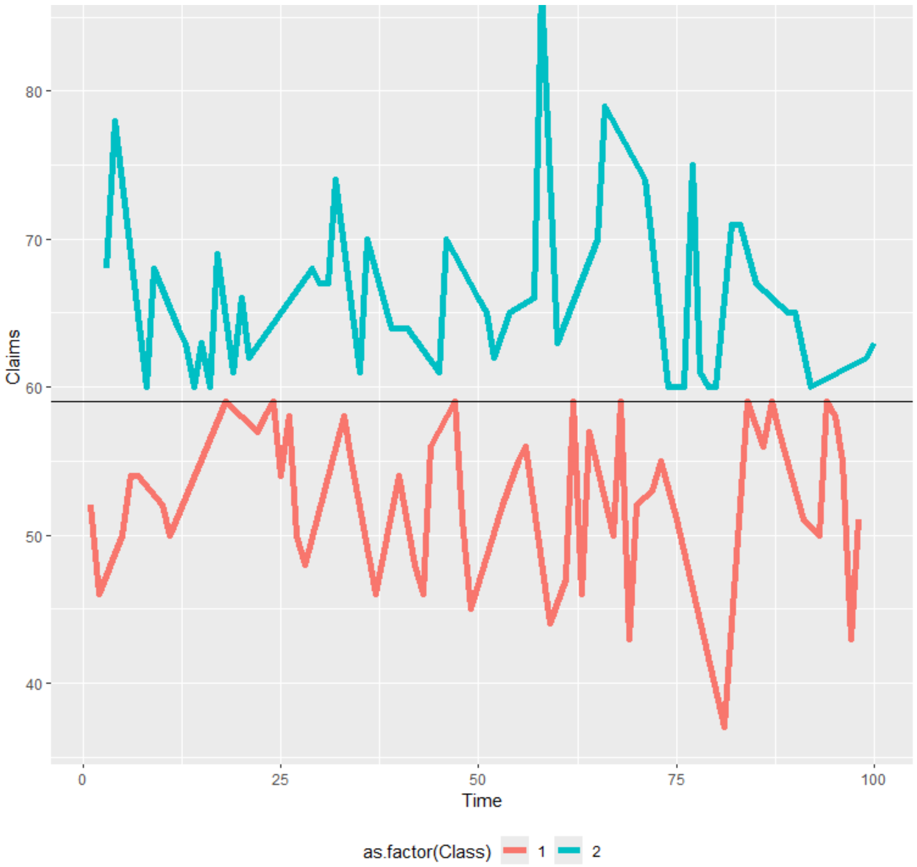
We describe and use *mclust*, an R library used solve the misreported claims problem set forth above (Scrucca, Fraley, Murphy, & Raftery, 2023) (Scrucca, Fop, Murphy, & Raftery, 2016). Among other features mclust provides the probabilities that each claim belongs to either class. This feature is key when there are possible overlaps in the cluster assignments of the underlying data.

Results

The simulated claims data is a result of cheating at a π percent rate – where the cheating is described by a Bernoulli process with parameter π. The reported claims series is as a weighted average of the true, unknown claims series adjusted by an amount Delta drawn from a kitty characterized by a poisson distribution with mean λ_2. Thus, the cheating inflates the number of claims.

The result of the fudging is a combination of two data-generating processes. Results obtained by estimating the parameters of a mixture model do show that it is possible to adjust a reported series to account for instances of deliberate or accidental misreporting. Figure 1 shows the settings of the initial conceptual framework and the settings for the simple explanations. In effect, $\pi = 25\%$, the average of the real, latent claims series $\lambda_1 = 50$, and the mean of the fudge kitty is $\lambda_2 = 60$. The resulting mixture, the reported claims series has an average of 52.8. The

Figure 1



random number seed is set at 42. This initial setting is seen in Figure 1.

Initial Results

Estimated model parameters from *mclust* are $\lambda 1$ = 49 and $\lambda 2$ = 56. The estimated levels are not too different from the simulated ones. Classification of the claims allows us to establish the rate of pilfering. The results in the table below.

Table 1
Cheat Rate

Actual or Latent	Pilfered	Est Cheat Rate
42	58	58 %

Those claims labeled as having been inflated are noted above under the column labeled “Pilfered.” The ratio of the adjusted series to the total number constitutes the rate of pilfering, in this instance, 58 percent. Note that this “rate of pilfering” is amenable to statistical testing to distinguish the events from instances where the identified errors may have occurred by chance. However, the result is grossly inaccurate when compared to the chosen simulation rate of 25 percent.

How Robust is the Model?

An examination of the sensitivity of the model is provided in Table 2 showing an instance of the true (corrected) series and the adjusted (misreported) series. Simulations were set at 100 iterations. The Cheat Rate was varied; it ranged from 0.1 through 0.9 by increments of 0.2; and the mean of the “fudging series” was varied as 52,55, 58, and 60.

Table 2
Simulation Results

Sim CheatRate	Est Claims	RMSE	Fudge Rate	Cheat Amount
0.1	46.44	9.14	0.39	2
0.3	45.96	8.87	0.55	2
0.5	46.04	9.38	0.6	2
0.7	48.83	7.62	0.39	2
0.9	48.92	8.00	0.26	2
0.1	51.48	8.65	0.76	5
0.3	48.00	8.85	0.35	5
0.5	48.92	9.01	0.49	5
0.7	50.29	8.15	0.68	5
0.9	52.29	9.40	0.3	5
0.1	45.91	8.85	0.52	8
0.3	50.56	11.10	0.05	8
0.5	50.74	9.94	0.32	8
0.7	48.05	10.31	0.72	8
0.9	52.87	9.24	0.5	8
0.1	46.38	8.91	0.59	10
0.3	46.61	10.10	0.33	10
0.5	50.35	11.11	0.33	10
0.7	52.74	9.88	0.5	10
0.9	55.98	9.49	0.48	10

The average of the estimated “cheat-rate” is 45.5 percent. The set cheat-rate of 25 percent and the average estimated cheat rate vary significantly. On the other hand, the average of the mean of the estimated series equals 49.4; this result does conform quite closely to the set mean of 50.

Limitations

The proposed DGPs were both poisson distributions. The cheat rate was established as a Bernoulli process. It is worth examining whether other DGPs fare better.

We relied on *mclust* for the task of both clustering and classification; classification led to the estimate of the cheat rate. However, there are any number of sophisticated libraries other than *mclust* available for clustering and classification.

A logical extension to the univariate work here is to repeat the analysis examining count data with possible explanatory variables.

Concluding Comments

The objective was to examine the usefulness of setting forth a mixture model as the conceptual framework underscoring the need to find hidden structure in a series of counts. Specifically, the possibility of deliberate tampering or the result of cognitive biases raises concerns over misreported outcomes in data disclosed in litigation.

If one understands that the cheating process underscoring the "misreported" data results in anomalies in the otherwise homogenous series, then it is possible to estimate the groups that emerged via clustering methods. Mixture models provide a relatively straightforward method to estimate groupings and can therefore be used to estimate the under- or over-counts as proposed here.

We empirically demonstrated how to use a model-based clustering algorithm to estimate the mixture model. Specifically, the R library *mclust* provides a relatively clear approach to estimation and cluster identification. We also examined the sensitivity of the approach to variation in cheat-rate and the “inflation amount.” Results were promising.

Importantly, a key consideration in advocating this approach was that the method should retain the focus on the constraints and concerns required of forensic expert testimony: succinctness and transparency. The method promises to be robust to opposing counsel’s imprecations.

References

Brody, R. G., DeZoort, F. T., Gupta, G., & Hood, M. B. (2022). The Effects of Cognitive Bias on Fraud Examiner Judgments and Decisions. *Journal of Forensic Accounting Research, 7*(1), 50-63. doi:https://doi.org/10.2308/JFAR-2020-030

Contzen, N., De Pasquale, S., & Mosler, H.-J. (2015, August 24). Over-Reporting in Handwashing Self-Reports: Potential Explanatory Factors and Alternative Measurements. *PLOS ONE*. doi:http://dx.doi.org/10.6084/m9.figshare.1304955

Harvin, O., & Killey, M. (2021). Do "Superstar" CEOs Impair Auditors' Judgement and Reduce Fraud Detection Opportunities? *Journal of Forensic and Investigative Accounting, 13*(3). Retrieved from http://s3.amazonaws.com/web.nacva.com/JFIA/Issues/JFIA-2021-No3-7.pdf

Huang, Z., Zheng, H., Li, C., & Che, C. (2024). Application of Machine Learning-Based K-means Clustering for Financial Fraud Detection. *Academic Journal of Science and Technology*, 33-39. doi:https://doi.org/10.54097/74414c90

Ioannidis, J. P. (2021). Over- and under-estimation of COVID019 deaths. *European Journal of Epidemiology, 36*(6), 581-588. doi:https://doi.org/10.1007%2Fs10654-021-00787-9

Li, T., Trivedi, P. K., & Guo, J. (2003). Modeling Response Bias in Count: A Structural Approach with an Application to the National Crime Victimization Survey Data. *Sociological Methods and Research, 31*(4), 514-544.

Neubauer, G., Djuras, G., & Friedl, H. (2011). Models for Underreporting: A Bernoulli Sampling Approach for Reported Counts. *Austrian Journal of Statistics, 40*(1 & 2), 85-92. doi:https://doi.org/10.17713/ajs.v40i1&2.200

Nigrini, M. (2022). Estimating the COVID-related Excess Deaths in 2020 Using Time Series Analysis. *Journal of Forensic and Investigative Accounting, 14*(2), 177-190. Retrieved from https://www.nacva.com/content.asp?contentid=1127#1

Nigrini, M. J. (2020). *Forensic Analytics*. New York: John Wiley & Sons Inc.

Pararai, M., Famoye, F., & Lee, C. (2010). Generalized Poisson-Poisson Mixture Model for Misreported Counts with an Application to Smoking Data. *Journal of Data Science, 8*(4), 607-617. doi:https://doi.org/10.6339/JDS.2010.08(4).608

Rodriguez, A. E., & Kucsma, K. (2023). Appraising Audit Error in Medicaid Audits. *International Journal of Accounting and Financial Reporting, 13*(3), 2162-3082. doi:https://doi.org/10.5296/ijafr.v13i3

Schennach, S. (2022). Measurement Systems. *Journal of Economics Literature, 60*(4), 1223-63.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, 289-317. Retrieved from https://journal.r-project.org/archive/2016/RJ-2016-021/RJ-2016-021.pdf

Scrucca, L., Fraley, C., Murphy, T. B., & Raftery, A. E. (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. Chapman and Hall/CRC.

Stamey, J. D., & Young, D. M. (2005). Maximum Likelihood Estimation for a Poisson Rate Parameter With Misclassified Counts. *Aust. N. Z. J. Stat., 47*(2), 163–172. doi:https://doi.org/10.1111/j.1467-842X.2005.00381.x

Ward, M. D., & Alhlquist, J. S. (2018). *Maximum Likelihood for Social Science*. Cambridge, UK: Cambridge University Press.

Wei, D., Cho, S., Vasarhelyi, M. A., & Te-Wierik, L. (2024). Outlier Detection in Auditing: Integrating Unsupervised Learning within a Multilevel Framework for General Ledger Analysis. *Journal of Information Systems*, 1-2. doi:https://doi.org/10.2308/ISYS-2022-026

Wood, J. S., Donnell, E. T., & Fariss, C. J. (2016, October). A Method to Account for and Estimate Underreporting in Crash Frequency Research. *Accident Analysis & Prevention, 97, Part A*(October), 57-66. doi:https://doi.org/10.1016/j.aap.2016.06.013

Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals Data Science*, 165-193. doi:https://doi.org/10.1007/s40745-015-0040-1

Appendix

Table 3
Clustering Libraries Examined

Package	Version	Non-Gaussian Components	Classification
Rmixmod	2.1.10	Yes	Yes
mixR	0.2.0	Yes	Yes
MixAll	1.5.1	Yes	Yes
mixtools	2.0.0	Yes	Yes
mclust	6.0.0	No	Yes