Testing for anomalies in count data
Review of Pitt and Hill, ScienceOpen Research 2016

Boying Gong[1], Aaron J. Stern[2,*], Yu Wang[1], and Yun Zhou[1]

[1]Department of Statistics, University of California, Berkeley, Berkeley, CA 94703
[2]Computational Biology Graduate Group, University of California, Berkeley, Berkeley, CA 94703

October 20, 2016

1 Summary

Pitt and Hill have presented an exciting and lucid analysis that introduces several new methods for detecting fraud. More precisely, the authors analyze count data recorded by an individual referred to as "research teaching specialist" (RTS). They analyze the RTS data using a set of statistical tests designed to detect anomalous patterns in count data. The p-values for these tests are compared to those for data from groups of other investigators using the same protocols. For all the tests they performed, the authors reject the hypothesis that the RTS reported the data accurately. Conversely, they find no significant anomalies in the comparison groups.

2 Reproducibility

We strove to reproduce the main results (i.e., hypothesis tests) presented in this paper. Code for our review is available on Github at

`https://github.com/35ajstern/reproduce_sor_2016/`

in the `report/` folder, written as a Jupyter notebook; main results of the reproduction are also presented at the end of the review as data tables with red marks that indicate our own results (Tab. 1-4).

We were able to reproduce most of the authors' results, with a couple of minor discrepancies that may have arisen from filtering that was unspecified in the paper. The Jupyter notebook also contains our own novel analysis of the paper's data, which we discuss throughout this review.

3 Study design and alternative analyses

3.1 Mean-containing and mid-ratio tests

3.1.1 Poisson assumption

A central contribution of this paper is its presentation of novel mid-ratio/mean-containing tests for count data. These methods assume that count data $\{X_j^i\}_{j=1}^3$ within a triple is $X_j^i \overset{iid}{\sim} \text{Pois}(\lambda_i)\ \forall i \in \{1, \ldots, N\}$, where N is the number of triples in the population. Does this probability model sufficiently describe the dynamics of a cell population? The fate of a cell is likely dependent on the fates of its neighbors; this relationship is not captured in the Poisson model. We are concerned that the Poisson assumption might be unrealistic, and wish the authors had discussed in more detail what behavior could be expected from their tests when this assumption breaks down.

[*]All authors contribute equally to the work. Correspondence should be addressed to AJS: `ajstern@berkeley.edu`

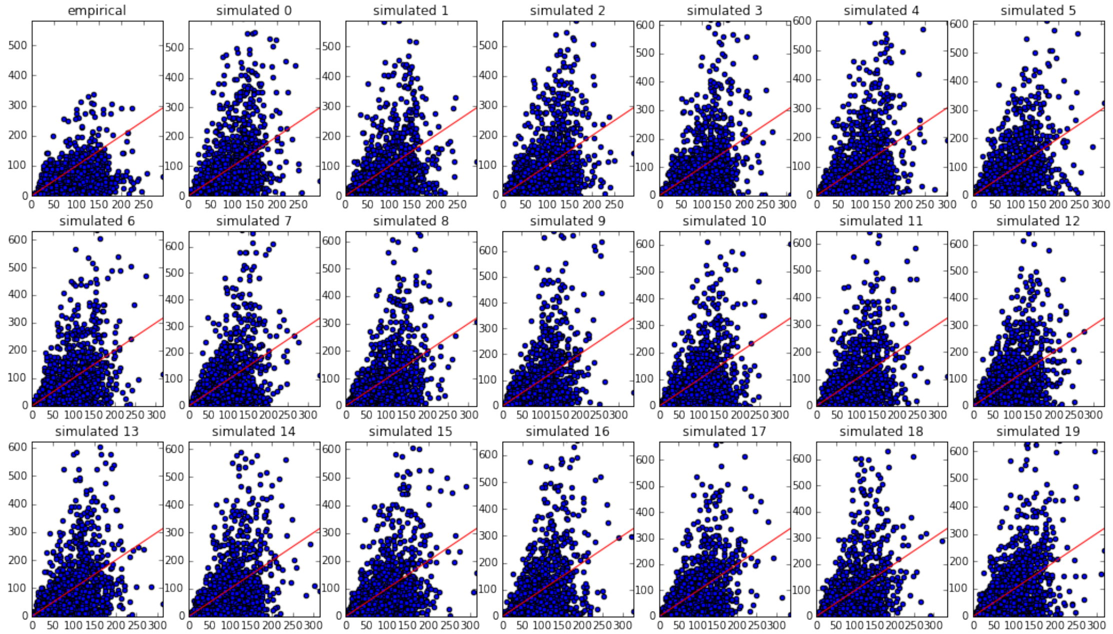

Figure 1: Empirical compared to simulated distributions of $\hat{\sigma}^2$ v.s. $\bar{X}$. The Poisson parameter of each randomly simulated Poisson triple is the sample mean of a corresponding real triple in the RTS data. The red line represents the expected $\hat{\sigma}^2$.

We independently examined the claim that counts within a triple are distributed Poisson by comparing the real data with simulated Poisson triples. Independently and identically distributed Poisson variables should on average have sample mean equal to the unbiased estimate of variance

$$\mathbb{E}[\bar{X}] = \mathbb{E}\big[\hat{\sigma}^2(X)\big] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

where $n = 3$ for triples.

To test whether the experimental data adheres to this canonical relationship, we performed linear regression on $\bar{X}$ and $\hat{\sigma}^2$ for all the triples from a putative control group (colony triples from other investigators in the RTS lab). If the Poisson distribution assumption were true, then the slope of this regression would be approximately equal to 1 (Fig. 1). However, the regression coefficient for the real data is 0.73, which means the sample variance of the real data is substantially smaller compared to Poisson distribution. This suggests that the colony count data from other investigators do not follow a Poisson distribution. We performed the same test on Coulter machine-counted data from the group of other investigators, and found a regression coefficient that seems implausible under Poisson assumption (Fig. 2). In this case, however, the data is over-dispersed with a regression coefficient of 1.37.

Figure 1 in the paper shows that distribution of the mid-ratio from RTS colony triples is very different from that of the other investigators. While the authors just use this as supporting evidence rather than a concrete argument, it is worth pointing out that there is no reason these distributions should look similar. Assuming each triple X^i comes from a Poisson distribution with rate parameter λ_i for $i = 1, \dots, N$, where N is the number of triples, then the empirical distribution of mid-ratios will depend on the composition of the set $\{\lambda_i\}_{i=1}^{N}$, which certainly varies across experiments. Since colony counts from other investigators might have totally different rate parameters to those of RTS's experiment, there is no reason to expect the corresponding mid-ratios to be similar. To illustrate this issue, consider the difference in the empirical distribution of simulated mid-ratios for triples with $\lambda = 1$ vs $\lambda = 100$ (Fig. 3).

3.1.2 Stratification of mean-containing/mid-ratio tests

In their study, the authors present Figure 1 to suggest that RTS data has an anomalous proportion of mean-containing triples compared to the agglomerated group of other investigators from his lab. We stratified the histogram for these 9 investigators to see if there were anomalous patterns lingering within

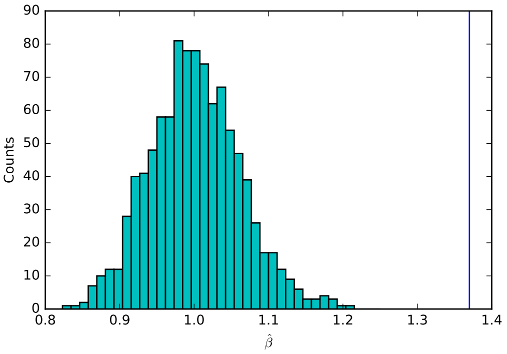

Figure 2: A simulated distribution of $\hat{\beta} = \frac{\hat{\sigma}^2}{\bar{X}}$ for $N = 1000$ random Poisson triples. Each triple was parameterized uniformly at random by $\hat{\lambda}_{ML}$ of a triple, which is sampled randomly from the real data. The blue line on the right represents the actual value of $\hat{\beta}$. Outliers of real data ($\hat{\sigma}^2 > 3\bar{X}$) were excluded.

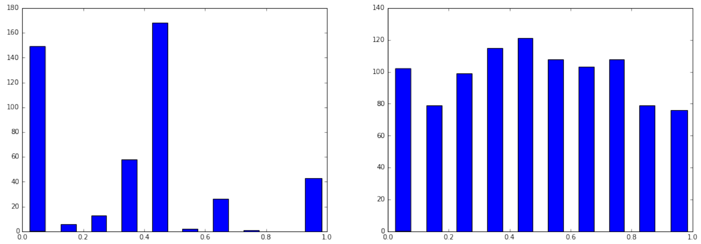

Figure 3: Left: mid-ratios for $N = 1000$ simulated Poisson triples with $\lambda = 1$. Right: mid-ratios for $N = 1000$ simulated Poisson triples with $\lambda = 100$.

the lumped group (Fig. 4). While sample size is too small to declare significance, there are numerous investigators within the lumped group who individually appear to record a high proportion of triples with mid-ratio concentrated about 1/2.

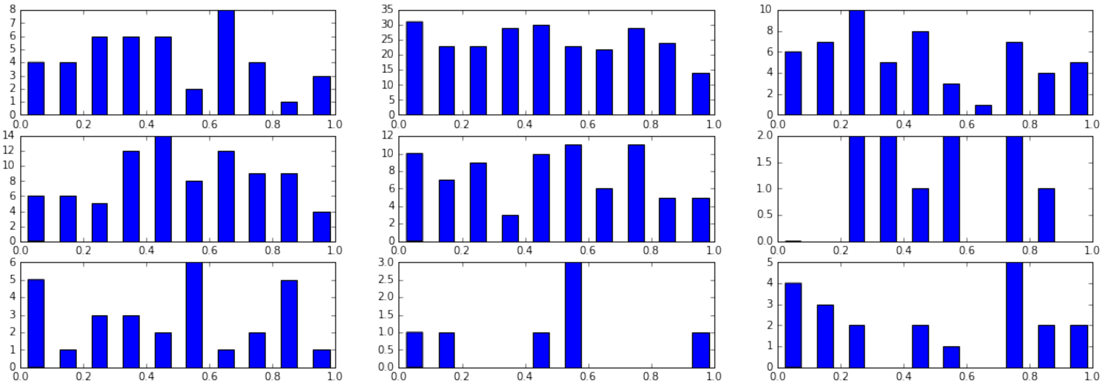

Figure 4: Mid ratio histogram for other 9 investigators stratified by individual.

3.1.3 Hypothesis testing

While the authors propose a plausible mechanism for how their novel mid-ratio/mean-containing tests might detect fraud, we wonder if they came to settle on performing this test only after they "peeked" at the data. Designing a statistical test in full knowledge of the data to be tested can often produce smaller p-values. We recommend that if data is used to guide test design, then some of the data should be apportioned into a disjoint testing set; there it remains unobserved until application of the hypothesis test to the test set (and not the previously observed set) to find significance.

Furthermore, the authors do not divulge the hypothesis tests that were considered or performed before the ones which they present in the paper. Our concern here is that the disclosed hypothesis tests may have happened to reject the null, while many more undisclosed tests may have not. Providing this information is invaluable for managing the false discovery rate.

Hypothesis test I is based on a (numerically) conservative bound, while test II treats estimated values of λ as if they have no uncertainty, which might result in an unconservative test (the true p-value could be rather larger than the nominal p-value). For the data in the paper, the conservative test yields an extremely low p-value; we suppose the authors presented the other test because it might be useful in other situations. However, we prefer the cruder test I to test II; because $\hat{\lambda}_{ML}$ (the maximum likelihood estimate (MLE) of λ) is crude estimate for the rate parameter of a triple, test I seems to be conservative. For example, take the case that a triple occurs far from its expectation–say, (72,102,104)–when its "true" $\lambda = 70$. In that way, test I could be robust to the under- or over-dispersion of count data that we pointed out previously. Conversely, we view this as a problem in hypothesis test II, where the sample mean is used to stratify the triples by their "true" λ. The authors mention the sample mean is the MLE; however, they do not discuss its large mean squared error (MSE):

$$\text{MSE}_\lambda(\hat{\lambda}_{ML}) = \mathbb{E}_\lambda[(\bar{X} - \lambda)^2] = 1/9(\sum_{i=1}^{3} \text{Var}_\lambda(X_i)) = \lambda/3$$

To this end, we would have liked to see an exploration of the sensitivity of the true level of the tests to the uncertainty in the sample mean as an estimate of λ.

Hypothesis test III applies the Lindeberg-Feller Central Limit Theorem (L-FCLT) to approximate the distribution of occurrences of mean containing triples. We do agree with the authors that the Bernoulli events "triple X^i is mean-containing" satisfy the Lindeberg Condition as the number of triples grows large. However, the authors use the mean of each triple–a highly unstable estimate, as we just discussed–and a comparably small sample size (i.e., the number of triples). Therefore, we are concerned that the authors take for granted that the L-FCLT would be suitable to approximate the number of mean-containing triples when the total number of triples is only on the order of 10^3. Furthermore, the unbiasedness of the estimate $\hat{p}_i = f(\hat{\lambda}_{ML})$ is not guaranteed, and therefore it is not guaranteed that the L-FCLT holds for $\{p_i\}_{i=1}^{N}$.

3.2 Tests of digit uniformity

We find the subsequent tests deployed by the authors to be more compelling than the aforementioned mean-containing/mid-ratio tests. That the authors cite usage of terminal digit analysis in previous studies of fraud suggests to us that these tests were more likely to have been selected agnostic of the data. As a result, we have fewer concerns about "peeking" at the data and ensuing selective inference on these two tests.

The authors' chi-squared test on the occurrence of terminal digits banks on the assumption that the distribution of terminal digits is uniform when a single count is iid Pois(λ). We checked this claim and found that when we simulated Poisson random variables with $\lambda < 30$, it is not a reasonable assumption (Fig. 5). That being said, the majority of the colony count data in the study takes on values larger than 30, so that assumption works well if the observed rates indicate the underlying theoretical rates (we also suspect that terminal digits under an over-dispersed distribution converge even more rapidly to uniformity). However, the authors do not appear to have filtered out data with small empirical rates; in fact, our reanalysis suggests they did not discard single-digit numbers in the terminal digit analysis. Nonetheless, our reanalysis was largely concordant with the authors', with slight differences that do not affect significance.

That said, we have a serious concern about the usage of this test to compare the χ^2 of an individual to the χ^2 of a lumped group; for example, consider a group consisting of two individuals–one of whom only records even numbers and the other only odds. If their counts "cancel out" sufficiently, their group may have an insignificant χ^2 value (perhaps even equal to 0). Separately, these two individuals would no doubt have significant χ^2 statistics. This pathology arises from testing individuals against groups. While this example directly concerns terminal digit uniformity, it can also apply to the authors' equal digit analysis and mean-containing/mid-ratio tests. In all of these tests, opposite biases can cancel each other out when lumped into a single group. In the following section, we examine how the authors' results change when data is stratified individual-by-individual.

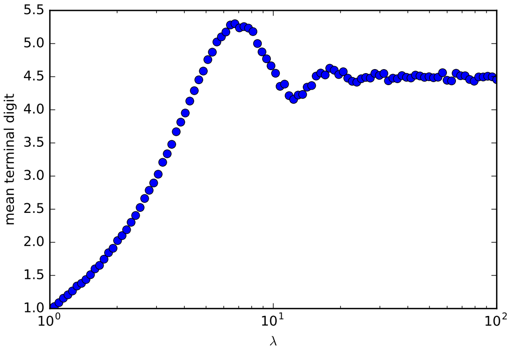

Figure 5: The mean terminal digit of a Poisson random variable does not converge to 4.5 (necessary for uniformity) until $\lambda > 30$. We simulated $N = 10^4$ variables for each value of λ.

3.2.1 Stratification of digit uniformity tests

To examine how lumping of individuals affects significance, we stratified data from other investigators in RTS's lab individual-by-individual based on codes in the authors' spreadsheets. We performed a terminal digit and equal digit analysis on these groups and found several individuals who produced unlikely data: Investigators D and F had statistically significant terminal ($p < 0.01$) and equal digit data ($p < 0.05$), respectively (Tab. 5,6).

3.2.2 Arbitrary digit pairs

We were confused that the authors looked for an enrichment of equal digits in the data. People committing fraud may avoid fabricating equal digits by the token that they are 9 times less likely than non-matching digits under uniformity (this reasoning is congruent to the authors' motivation for mid-ratio tests, which look for an enrichment of likely triples). We performed a test equivalent to the equal digit analysis on 10 non-equal digit pairs – $\{01, 12, \cdots, 90\}$ – and looked at how anomalous individuals appeared under this test versus equal digits. We found that this choice of digit pairs produced a test that suggested 4 of the 9 other investigators (as well as RTS) had unlikely data (Table 7).

3.3 Permutation testing: terminal and equal digits

Touching back on our criticism of how grouping affects calculation of χ^2, we reiterate our concern that the individual RTS was tested against groups of other investigators. It is not clear why RTS was singled out; other researchers might also have fabricated data. To control for the effects of this way of testing the data individual-to-group, we implemented two non-parametric permutation tests.

To test the abnormality of RTS's data, we took data from RTS and other investigators, combined it into one group, and repeated permuted their labels ("RTS" or "Other Investigators"). These new permuted populations were used to calculate the chi-squared and total variation distance between terminal digit frequencies of each pair of permuted groups (Fig. 6). Indeed, the p-value of the actual RTS data's distance (both TVD and χ^2) are extremely small. This results reinforces the claim that RTS's was unlikely to have occurred by chance, even if the data are not observations of Poisson variables. We would have liked to do pairwise permutation tests stratified individual-by-individual, but no individuals besides RTS contributed sufficient data to perform such tests. Please refer to our Jupyter notebook for additional permutation tests of equal digit pairs and triple mid-ratios.

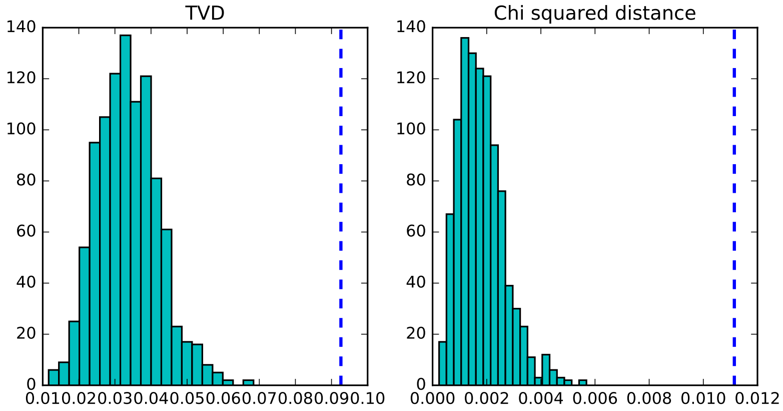

Figure 6: Left: Total variation distance (TVD) of terminal digit frequency in $N = 1000$ permutations of RTS vs others (cyan); TVD of the actual RTS data vs others (dashed bar). Right: χ^2 distance applied to the same permutation scheme.

4 Concluding remarks

The authors offer an overall persuading analysis of the data. Ultimately, we believe the authors' tests indicate that some fraction of the RTS data is fabricated. However, we are concerned that their novel hypothesis tests may have been designed deliberately to detect anomalies they observed *a priori* in the RTS data. We showed evidence contrary to some of the authors' main assumptions, including the Poisson distribution of triples. We also showed that the design of the test groups glazes over potentially suspicious individuals within the comparison groups. Lastly, we designed two new permutation tests for count data abnormality that do not rely on parametric assumptions. Test for fraud should be careful to avoid selective inference, and we find evidence of fraud that depends on parametric assumptions is less compelling than evidence based on nonparametric tests.

We would like to thank the authors of the paper we reviewed for contributing this very interesting study, for making their data public, and for choosing to publish in an open-access journal.

Also, we would like to acknowledge Philip B. Stark, who vetted this review. However, the work was conducted entirely by the authors, and the opinions expressed in this review are those of the authors.

Table 1: Table 1 in the paper

λ	P	λ	P	λ	P	λ	P	λ	P
1	0.267✓	6	0.372✓	11	0.317✓	16	0.281✓	21	0.254✓
2	0.387✓	7	0.359✓	12	0.309✓	17	0.275✓	22	0.250✓
3	0.403✓	8	0.348✓	13	0.0301✓	18	0.269✓	23	0.246✓
4	0.397✓	9	0.337✓	14	0.294✓	19	0.264✓	24	0.242✓
5	0.385✓	10	0.327✓	15	0.287✓	20	0.259✓	25	0.238✓

Table 2: Table 2 in the paper

Type	Inv.	# complete/tot.	# mean	# exp.	SD	Z	$p \geq k$
Colony	RTS	1,343/1,361✓	690✓	220.3	13.42	34.97	0
Colony	Others	572/591 (578/597)	109✓	107.8	9.23	0.08	0.466
Colony	Lab 1	49/50 ✓	3 ✓	7.9	2.58	2.11	0.991
Coulter	RTS	1,716/1,717(1726/1727)	173(176)	97.7	9.58	7.80	$6.26 \cdot 10^{13}$
Coulter	Others	929/929 ✓	36✓	39.9	6.11	0.71	0.758
Coulter	Lab 2	97/97 ✓	0✓	4.4	2.03	2.42	1.00
Coulter	Lab 3	120/120 ✓	1✓	3.75	1.90	1.71	0.990

Table 3: Table 3 from paper

Type	Investigator	χ^2	p
Colony	RTS	200.7 ✓	0
Colony	Same lab	1.65 (1.79)	0.994363
Colony	Other lab	12.1 ✓	0.205897
Coulter	RTS	456.4 (466.88)	0
Coulter	Same lab	16.0 ✓	0.0669952
Coulter	Other lab 1	9.9 (9.48)	0.394527
Coulter	Other lab 2	4.9 ✓	0.839124

Table 4: Equal digit analysis (Coulter)

Investigator	x	n	p
RTS	636 (644)	5155 (5187)	8.57787e-09
Same lab	291 (286)	2942 (3021)	0.827748
Other lab 1	32	327	0.504864
Other lab 2	30	360	0.83282

Table 5: Stratified terminal digit analysis (Coulter)

Investigator	χ^2	p	n
A	8.10232	0.523869	1401
C	14.5789	0.10317	105
B	5.88889	0.750985	180
E	9.12121	0.426161	165
D*	21.8438	0.00938759*	645
G	5.33333	0.804337	60
F	6.96774	0.640478	312
I	9.4183	0.399591	153

Table 6: Stratified equal digit analysis (Coulter)

Investigator	x	n	p
A	132	1401	0.748688
C	8	105	0.733914
B	16	180	0.634373
E	13	165	0.777841
D	62	645	0.597186
G	4	60	0.729042
F*	40	312	0.0436366*
I	11	153	0.848016

Table 7: Alternative digit pairs

Investigator	x	n	p
RTS*	560	5187	0.027532*
A*	156	1401	0.0738102*
C	11	105	0.35797
B*	23	180	0.0896297*
E	10	165	0.947312
D	72	645	0.147142
G	6	60	0.393549
F*	39	312	0.0624213*
I	16	153	0.361155