Review of Statistical Analysis of Numerical Preclinical Radio-biological Data

Raaz Dwivedi^{+}, Antonio Iannopollo^{+} and Jiancong Chen*
$^{+}$ Dept of EECS, * Dept of Civil & Environmental Engineering
University of California, Berkeley

September 27, 2016

Abstract

This review reproduces tests and results presented by Pitt and Hill and discusses some other non-parametric techniques, such as Permutation Tests, which allow to analyze data with less restrictive assumptions. The focus of the review is on the statistical methodology rather than the underlying biological aspects and assumptions of the original work, which are not discussed. Although not expert in statistical methods for fraud detection, we do believe that permutation tests are promising in this context, as demonstrated by the results presented here. This review has been developed as a term project for a Graduate Level Course on Statistical Models at University of California Berkeley, by graduate students from EECS and Civil & Environmental Engineering departments.

1 Introduction

We review the paper in the spirit of promoting reproducibility of research and attempt to replicate the authors' work. We also discuss other methods to identify anomalies, and present results based on our analysis using Permutation Tests. Permutation tests are consistent with the aim of the paper–providing simple tools to detect anomalies–and validate the results in the paper, leading to the same conclusions.

Before diving into technical details, we make a minor observation: the organization of the paper was not properly introduced. The use of distinct sections for - (1) the discussion on data and experiments; (2) their model and related calculations; (3) the application of common tests from the literature and (4) conclusions - would have been helpful. The review is organized as follows. In section 2 we replicate authors' work and results. In section 3, we analyze weaknesses of the approach followed in the paper and propose additional techniques to consolidate the results. We finally draw our conclusions in section 4.

1.1 Problem Set Up

The paper begins by voicing a growing concern towards "Scientific fraud and Plagiarism" in the scientific community and is successful in conveying a strong message. The authors present some statistical figures and point out the existence of easy statistical tools to detect fabricated data and ignorance about such tools.

The authors examine datsets from radio-biological experiments. They find that data reported by one of 10 researchers, the "RTS", is suspicious. They perform three different tests to validate their suspicion and also validate their tests and assumptions by looking at the data obtained from three other sources.

Each researcher made two types of triple measurements - colony counts and Coulter counts. The authors suspect that the RTS fabricated data triples to get the mean s/he desired in each triple by setting one observation equal to the desired mean and the other two roughly equidistant above and below that value. This would result in triples that contain the (rounded) mean as one of their values.

The methodological contribution of the paper is "bounds and estimates for the probability that a given set of n such triples contains k or more triples which contain their own mean" when each of the n triples is independent and identically distributed (i.i.d.) Poisson, and triples are independent of each other. (Different triples may have different Poisson means.) For this Poisson model, the chance that the RTS's data would contain so many triples that include their rounded mean is astronomically low. They also apply more common tests for anomalous data, based on statistics such as the frequency of the terminal digit and the frequency with which the last two digits are equal.

However, some of the questions that were slightly untouched upon are discussed below:

- The authors write, "Having observed what appeared to us to be an unusual frequency of triples in RTS data containing a value close to their mean, we used R to calculate the mid-ratios for all of the colony data triples that were available to us." This suggests that the same data–and the same feature of the data–that raised their suspicions about the RTS was the data used to test whether the RTS's data were anomalous on the basis of that feature. If so, then the nominal p-values are likely to be misleadingly small.
- Most of the tests assume a model for the observations and compare the RTS's data to that model. The authors validate the assumptions of the model by comparing it with the data pooled for the other researchers. Pooling the data in this way may hide anomalies in the other researchers' data. Permutation tests allow the data from each researcher to be compared to the data from the other researchers without positing a generative model for the data. On the other hand, the bulk of the data available is from the RTS. To reject the hypothesis that another researcher's data looks like a random sample from the pooled data, if it includes the RTS's data, does not imply s/he is suspicious. Instead, it shows that his/her data is not like that of the RTS. See section 3 of this review for more discussion.

2 Reproducibility of Results

This section discusses our efforts to replicate the analyses in the paper. After fine tuning, we were able to replicate most of their results, obtaining similar results in the other cases. Our work is available on github [github.com/ianno/stat215a_project1]. We first discuss specifics about the replication and then comment about the tests and methods involved.

2.1 Mid-Ratio Analysis

The authors first consider the mid-ratio, which is defined for a triple $(a, b, c), a < b < c$ as $\frac{b-a}{c-a}$, and show that the histogram of RTS's data concentrates abnormally around the $0.4 - 0.6$ range, compared to the data taken by all the other lab members. After tweaking the default histogram function on numpy, we were able to obtain plots similar to the ones reported in Figure (1) of the paper. Two noticeable differences were - (1) we obtain 44% chance of seeing mid-ratio in $(0.4, 0.5]$ interval for RTS, compared to 50% chance reported in the paper and (2) we used 1360/1361 and 595/595 triples to compute histogram for RTS and the rest respectively, compared to the use of 1343/1361 and 572/595 triples by the authors. We believe the authors did not provide enough information about the methods used to filter data for this section. However, such minor differences did not demand further investigation.

2.2 Probability Model

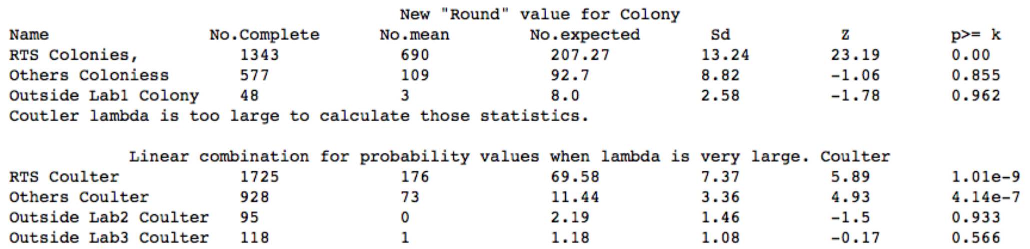
The authors develop a model to bound the probability of observing k out of n triples contain their mean. Each entry in a triple is assumed to be an independent sample from a Poisson distribution with mean λ. (Different triples may have different means.) The event of observing the rounded mean in such a triple is a Bernoulli random variable (BRV) whose success probability depends on λ. The authors derive analytical expressions for these success probabilities in Appendix A. Numerical values of these probabilities, for $\lambda = \{1, \ldots, 25\}$, are presented in Table 1. We could replicate this table exactly. For large λ (> 2000), for which the authors provide only few representative probability values, our implementation suffered from numerical issues.

Using Table 1, the authors determine the success probability for the BRV in two different ways and use it to compute the chance of observing the data. For hypothesis test I (non-parametric) they used the maximum value from Table 1 as an upper bound for all triples, essentially treating all BRVs as i.i.d. Bernoulli(0.42). Replicating this was straightforward.

For hypothesis test II and III, the authors use maximum likelihood estimate of λ for each triple to look-up Table 1, essentially treating each BRV to have a different success probability. We discuss this in the next section.

2.2.1 Hypothesis Test II and III

We could replicate the chance values up to minor errors for the colony data. Limitations of our implementation gave wrong results for Coulter data. For sanity checks of the results, we used linearly interpolated values from the paper (for intermediate λ) and obtained values similar to those in the paper for these tests. Figure 1 is the approximate replication of Table 2 from the paper.



```
                              New "Round" value for Colony
Name                  No.Complete      No.mean         No.expected       Sd            Z           p>= k
RTS Colonies,            1343           690              207.27         13.24         23.19        0.00
Others Coloniess         577            109              92.7           8.82          -1.06        0.855
Outside Lab1 Colony      48             3                8.0            2.58          -1.78        0.962
Coutler lambda is too large to calculate those statistics.

             Linear combination for probability values when lambda is very large. Coulter
RTS Coulter              1725           176              69.58          7.37          5.89         1.01e-9
Others Coulter           928            73               11.44          3.36          4.93         4.14e-7
Outside Lab2 Coulter     95             0                2.19           1.46          -1.5         0.933
Outside Lab3 Coulter     118            1                1.18           1.08          -0.17        0.566
```

Figure 1: Approximate Replication of Table 2

2.3 Digits Analysis

Next, the authors perform some common tests for fraud detection - *terminal digit analysis* and *pair of equal terminal digits analysis*. These tests assume that in general insignificant digits of a sample are not very informative.

2.3.1 Terminal digit analysis

The first test assumes that the last digit in samples of large numbers (> 100) should empirically show uniform distribution. Also, some previous works have shown that fabricated data often fail to show such peculiar property. The authors use the chi-square test for goodness of fit, and get good fits for the rest data and low p-values for the RTS data. Our results are very similar to theirs, although not identical.

2.3.2 Equal digits analysis

This test assumes that for large numbers empirical frequency of observing a pair of equal terminal digits is close to 1/10. Again, the authors consider only big numbers (> 100), to ensure the analysis of insignificant digits. The authors didn't mention the name of tests for this analysis and but assuming chi-square test for goodness of fit we obtain similar, but not identical results.

2.3.3 Discussion of Assumptions

Next, we discuss the assumptions made by the authors:

- The authors didn't justify the assumption of Poisson distribution for the underlying data. Beyond our intuition we didn't investigate the validity in detail.

- The authors suspected RTS data, but used the suspected data to fit a model and quantify their suspicion. While sometimes this may raise flags, here we agree with the authors that doing so increases the odds in favors the person in question and hence gives us desirable conservative results.

- The authors don't discuss why simply filtering the data for counts > 100 justifies the use of tests for suspicious patterns in insignificant digits. The authors include here additional data, provided by three external sources (two for Coulter counts and one for Colony counts) but all three had relatively small number of data points. Despite authors attempts to account for this, we believe that in the current setting, these additional samples do not more compelling evidence. Instead, they added to our confusion. TODO : What confusion?

- We reiterate that pooling the data may hide anomalies in the other researchers' data.

3 Our Analysis

As a preliminary test for identifying suspicious datasets, we plot histograms of mid-ratios for the data provided by individual researchers. We also contrast each histogram from the histogram of the pooled data. We only plot the former here. Two important observations can be made:

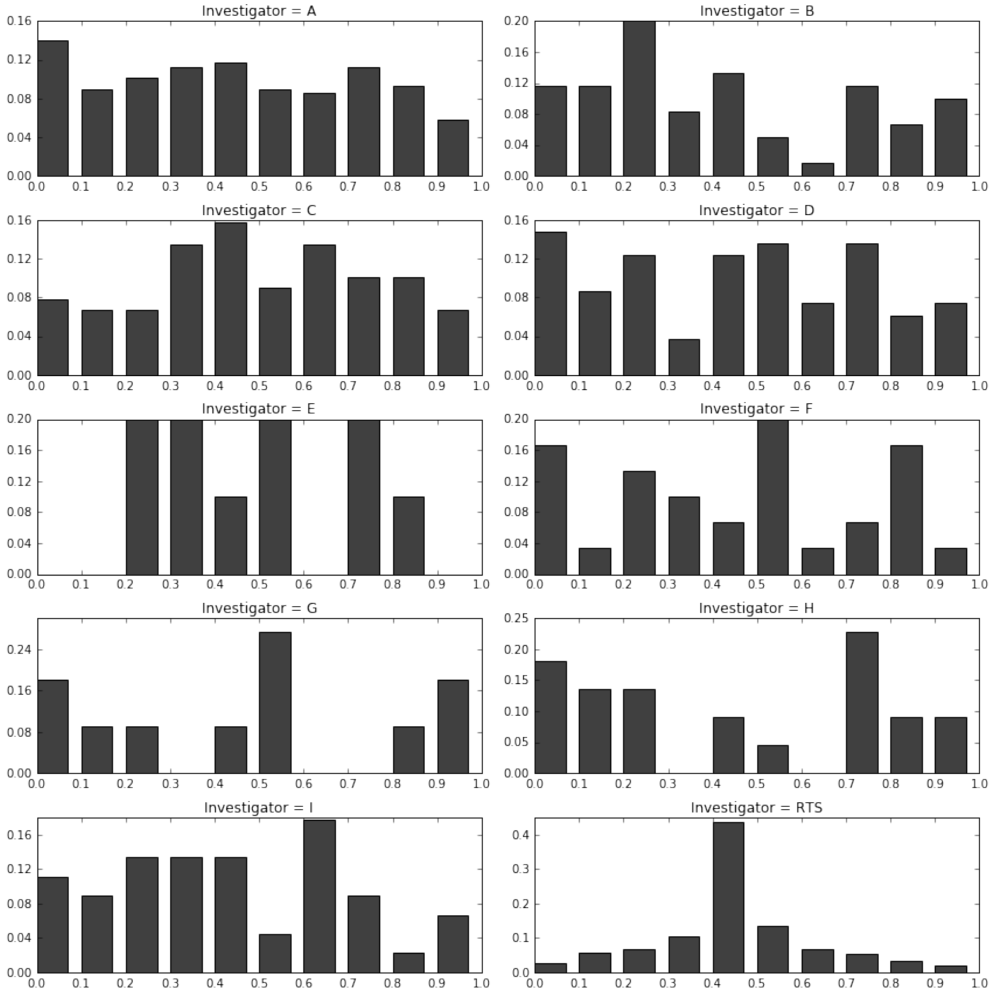


Figure 2: Individual Histograms for the Colony Data

- First, the histogram for researchers with labels "B, C, E, F, G, H, I" do not appear as uniform distribution.

- Second, when the pool includes RTS, the histogram of the pool looks similar to the histogram of RTS due to its dominating presence in the pool. Hence, most of the individual researchers look anomalous when contrasted with the pool.

These observations point the limitations of assumption of uniform distribution for mid-ratios and use of visual dissimilarity between histograms of RTS to the pool. Next, we discuss permutation tests which are free from many such limitations, and come in handy when little information is available about the distributions of the data.

3.1 Permutation Tests

The problem of determining whether a treatment has an effect is widespread in various real world problems. To evaluate whether a treatment has an effect, it is crucial to compare the outcome when treatment is applied (the outcome for the treatment group) with the outcome when treatment is withheld (the outcome for the control group), in situations that are as alike as possible but for the treatment. This is called the method of comparison. We will describe this method for a specific set up which is relevant for the problem in discussion.

Suppose that we are given two labeled sets of observations - T of them labeled 'treatment' and C of them labeled 'control'. We assume that the prior of these sets has received a treatment and we wish to test the hypothesis whether this treatment affected the outcomes. In a two-sample permutation test, the data is pooled together to form a population of size $N = T + C$. Next, we decide a test-statistic that can capture the effect of the treatment (if any) between the two groups, and compute it for the given labeled sets. As an example, we can consider the absolute difference between the sample means of the two datasets as the test statistic. Under the null hypothesis that the treatment has no effect, one can analytically derive the distribution of this test statistic. However, often it is easier to empirically approximate rather than numerically compute the distribution. To do so, one repeats the following procedure several times: partition the data into groups of size T and C and compute the test statistic contrasting the two datasets. We use the empirical histogram obtained from these repeated experiments, as a proxy for the true distribution of the test statistic. Next, just like typical hypothesis testing, we compute the chances (p-value) of observing the test statistic that we computed in the beginning.

When the p-value is below a preset significance level, we infer that the treatment has an effect at that level of significance. That is, we conclude that *the two groups are different to each other* than if we randomly partitioned the pooled dataset.

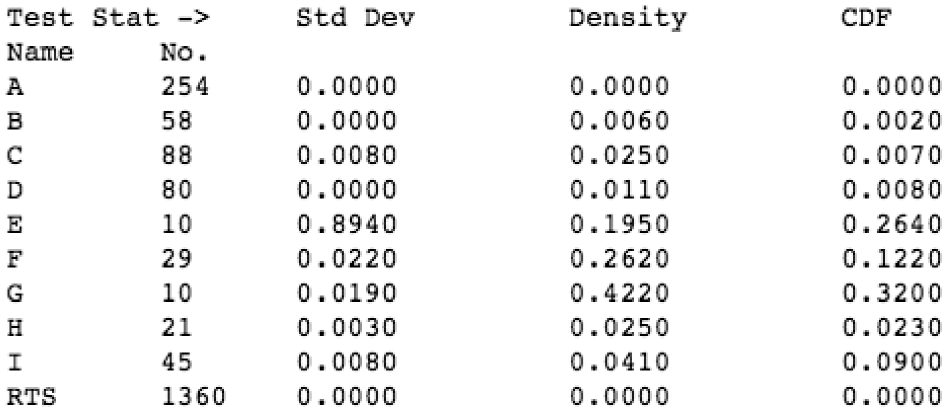
3.1.1 Results for Mid-Ratio

We set the test statistic to be the difference in standard deviation of mid-ratios of two datasets. We choose standard deviation because our null and alternative hypothesis for mid-ratio (uniform distribution versus concentration around 0.5) lead to same mean (0.5). We expect standard deviation to capture the *unintentional reduction in spread caused in data due to intentional adjustments.*

We consider each researcher equivalent to a treatment. We carry these tests in the following fashion. Suppose we want to test if data provided by investigator A were different from the others. Let n_A denote the number of points in A's dataset, and let $N - n_A$ denote the size of the dataset for the rest. These two datasets form the treatment and control group in our permutation tests for contrasting the effect of label 'A' on the data. We do the procedure explained in the previous section 1000 times. We obtained a p-value of 0.00 for A, B, D, and RTS; and < 0.01 p-value for all others except E,F,G which indicates that almost all datasets are surprising with respect to this test-statistic. We would like to note that here a p-value of 0.00 in fact denotes a p-value < 0.001, because of the finite resolution owing to 1000 tests. We would also like to mention that RTS is still the most surprising if one looks at the location of the test-statistic in the tails of the distribution.

We also look at ℓ_1 distance between the density[1], and the ℓ_1 distance between the cumulative distribution function (CDF) as the test statistic. Again, we reject several researchers of the lab at a significance level of 1%. We present all the p-values in Figure 3.

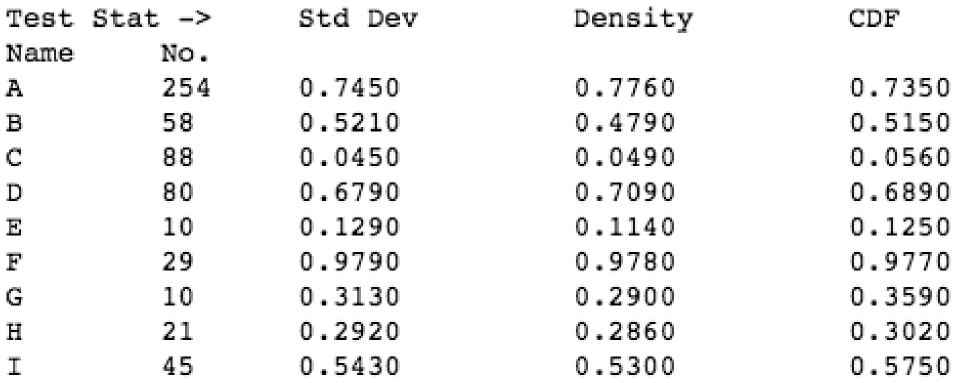
[1] abuse of terminology, used in place of normalized histograms



```
Test Stat ->     Std Dev           Density           CDF
Name      No.
A         254    0.0000            0.0000            0.0000
B         58     0.0000            0.0060            0.0020
C         88     0.0080            0.0250            0.0070
D         80     0.0000            0.0110            0.0080
E         10     0.8940            0.1950            0.2640
F         29     0.0220            0.2620            0.1220
G         10     0.0190            0.4220            0.3200
H         21     0.0030            0.0250            0.0230
I         45     0.0080            0.0410            0.0900
RTS       1360   0.0000            0.0000            0.0000
```

Figure 3: Results for Permutation Tests for Mid-Ratio

Remark We would like to mention that when RTS is included in the control group, it constitutes the bulk of the group. As a result, rejecting the null hypothesis is almost equivalent to rejecting the hypothesis that the data of the researcher is same as RTS data. This doesn't help us to make meaningful inferences if we already believed or later found that RTS data was suspicious. To correct for this, we do another set of permutation tests after excluding the RTS data. Now we don't find strong evidence to reject the null hypothesis, and conclude that none of the researchers have any effect at a significance level of 1%. However, these set of tests suffer from a bias because of our manual throwing away of 2/3rd data.



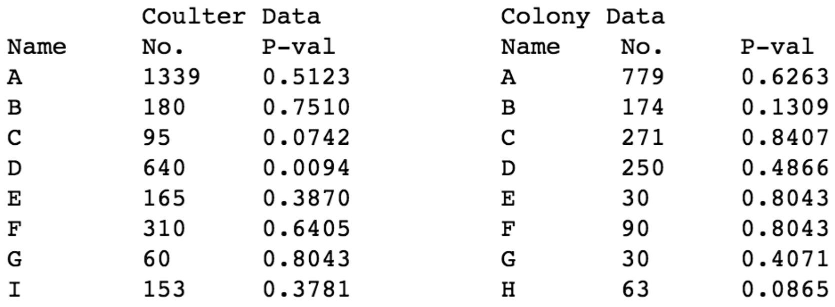
```
Test Stat ->     Std Dev           Density           CDF
Name      No.
A         254    0.7450            0.7760            0.7350
B         58     0.5210            0.4790            0.5150
C         88     0.0450            0.0490            0.0560
D         80     0.6790            0.7090            0.6890
E         10     0.1290            0.1140            0.1250
F         29     0.9790            0.9780            0.9770
G         10     0.3130            0.2900            0.3590
H         21     0.2920            0.2860            0.3020
I         45     0.5430            0.5300            0.5750
```

Figure 4: Results for Permutation Tests without RTS for Mid Ratios

Putting together the pieces, we can conclude that we do have statistical evidence to claim that RTS has suspicious data.

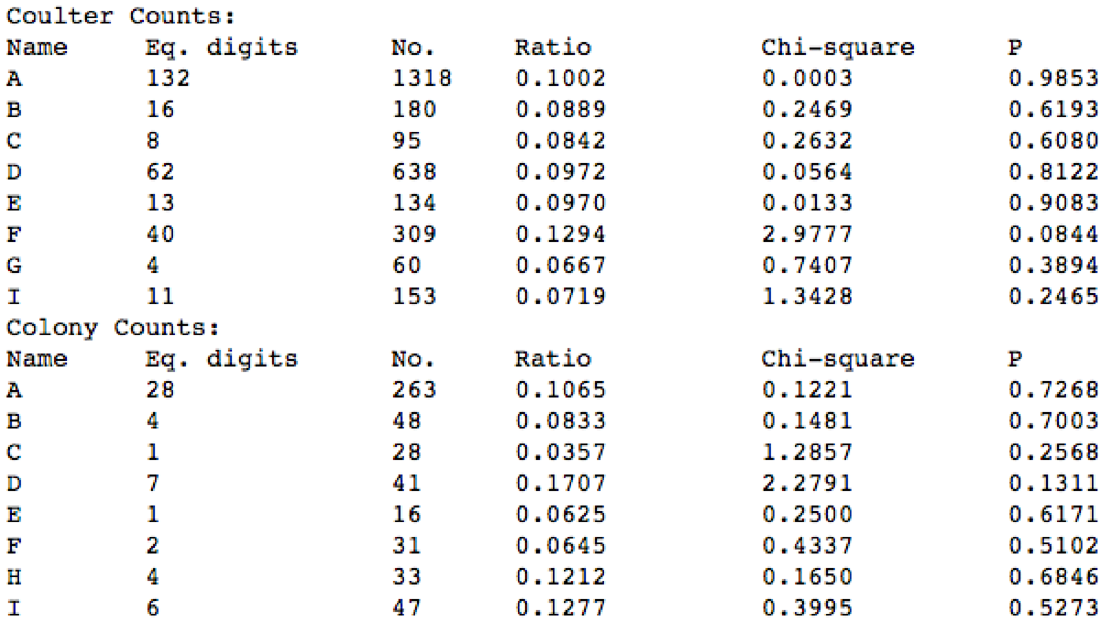
3.2 Additional Tests for Digit Analysis

For the terminal digit and equal digits tests, we extended the tests done by the authors to individual members of the lab and performed - chi-square test for goodness of fit for terminal digit; chi-square test for goodness of fit for equal digits and permutation tests for terminal digit. For permutation tests, we used the test statistics listed in the previous section. Results are tabulated in the figures below:



```
              Coulter Data                 Colony Data
Name          No.       P-val              Name     No.       P-val
A             1339      0.5123             A        779       0.6263
B             180       0.7510             B        174       0.1309
C             95        0.0742             C        271       0.8407
D             640       0.0094             D        250       0.4866
E             165       0.3870             E        30        0.8043
F             310       0.6405             F        90        0.8043
G             60        0.8043             G        30        0.4071
I             153       0.3781             H        63        0.0865
```

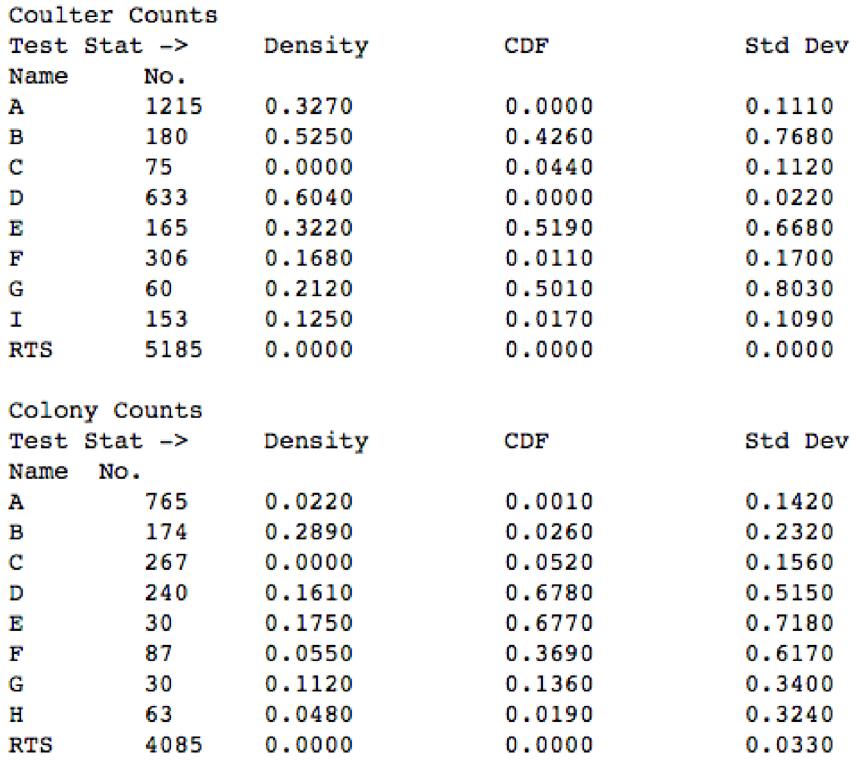
Figure 5: Chi Square Tests for Terminal Digits in Coulter and Colony Counts



```
Coulter Counts:
Name      Eq. digits       No.      Ratio           Chi-square      P
A         132              1318     0.1002          0.0003          0.9853
B         16               180      0.0889          0.2469          0.6193
C         8                95       0.0842          0.2632          0.6080
D         62               638      0.0972          0.0564          0.8122
E         13               134      0.0970          0.0133          0.9083
F         40               309      0.1294          2.9777          0.0844
G         4                60       0.0667          0.7407          0.3894
I         11               153      0.0719          1.3428          0.2465
Colony Counts:
Name      Eq. digits       No.      Ratio           Chi-square      P
A         28               263      0.1065          0.1221          0.7268
B         4                48       0.0833          0.1481          0.7003
C         1                28       0.0357          1.2857          0.2568
D         7                41       0.1707          2.2791          0.1311
E         1                16       0.0625          0.2500          0.6171
F         2                31       0.0645          0.4337          0.5102
H         4                33       0.1212          0.1650          0.6846
I         6                47       0.1277          0.3995          0.5273
```

Figure 6: Chi Square Tests for Equal Terminal Pair in Coulter and Colony Counts

Figure 5 shows that p-value for D, for Coulter Data is less than 1%.



```
Coulter Counts
Test Stat ->     Density          CDF              Std Dev
Name      No.
A         1215   0.3270           0.0000           0.1110
B         180    0.5250           0.4260           0.7680
C         75     0.0000           0.0440           0.1120
D         633    0.6040           0.0000           0.0220
E         165    0.3220           0.5190           0.6680
F         306    0.1680           0.0110           0.1700
G         60     0.2120           0.5010           0.8030
I         153    0.1250           0.0170           0.1090
RTS       5185   0.0000           0.0000           0.0000

Colony Counts
Test Stat ->     Density          CDF              Std Dev
Name   No.
A         765    0.0220           0.0010           0.1420
B         174    0.2890           0.0260           0.2320
C         267    0.0000           0.0520           0.1560
D         240    0.1610           0.6780           0.5150
E         30     0.1750           0.6770           0.7180
F         87     0.0550           0.3690           0.6170
G         30     0.1120           0.1360           0.3400
H         63     0.0480           0.0190           0.3240
RTS       4085   0.0000           0.0000           0.0330
```

Figure 7: Permutation Tests for Terminal Digit Analysis, Coulter counts

Figure 7 once again confirms that RTS data is suspicious. As before, the huge fraction of data by RTS contributes towards the low p-values for some of the other researchers. In permutation tests after excluding RTS, none of the researchers look suspicious. We skip the table of the p-values for this case.

4 Conclusion

Data fraud is an extremely critical issue in science, engineering and many other fields. Methods to detect manipulated data are needed to identify fraudulent research behaviors. Detecting frauds, however, is a delicate matter. Challenging the credibility of a researcher or of a scientific work, in fact, can have heavy consequences for all the parties involved in the process. Methodologies and techniques used in this kind of work need to be clear and widely accepted. They need to produce results which leave minimal (ideally no) space to ambiguity. Independently, reproducibility of results is a fundamental element to rule out any doubts that could arise at any time.

In our review, we carefully analyzed the authors' results and conclusions by: reproducing all the results that have been discussed in the paper and using additional tests to avoid several assumptions.

TODO : We out that authors' results are correct, although it has not been possible to reproduce exactly all the experiments due to lack of some key pieces of information (for instance how data has been preprocessed). Moreover, we encourage the use of stronger tools like permutation tests and our demonstration can be considered as a promotion of the same. Such tests help the analysis to get *rid of assumptions*, thereby shifting the focus from debate on assumptions to actual anomalies present and to better understanding of individual investigator's data (besides the RTS) as to how do they compare to the general data pool.

At the end of our review, we do believe that there is a significant evidence that RTS has suspicious data, but we suggest the authors to collect additional material and investigate more, since some of our tests suggest that other investigator's data have anomalies as well if we do not discount the huge fraction of data given by RTS.

Acknowledgments

We would like to thank the authors H. Pitt and H. Hill for publishing in an open journal, and making the data available for everyone. Also, we would like to thank Prof Philip Stark for his valuable and critical guidelines and timely feedback. We would also like to thank Yuansi Chen for valuable tips with python. As a final note, we would like to claim complete responsibility for all the opinions expressed in this paper.

References