Review of Statistical Analysis of Numerical Preclinical Radio-biological Data

Raaz Dwivedi, Antonio Iannopollo and Jiancong Chen

September 26, 2016

Abstract

This review is a term project for the Graduate Level Course on Statistical Modeling and Practices at University of California Berkeley. The authors are graduate students from department of EECS and Civil&Environmental Engineering and have restricted their attention to the methods and analysis done in the paper. The review is an attempt to reproduce the tests and results presented in the paper, and discuss some other non-parametric tests and results eg. Permutation tests, that can be seen as an alternative to making certain assumptions and finding surprises in the data. No attempt has been made to look into the biological aspects and validity of certain assumptions related to them. We did not search the literature for other methods for fraud detection. We do believe that permutation tests have promise, as demonstrated by calculations we present.

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.

We offer a minor suggestion: we would have found the paper easier to read if the sections and subsections had been numbered; reorganization of some of the material would have helped, too. Next, we discuss the problem set up considered by the authors, and make some remarks on the methods used. In Section 2 we replicate authors' work and results to some extent. In Section 3, TODO rewrite:we discuss some gaps as a reader and attempt to take a step back and do some tests, and point the gaps in the work and how to address them using statistical methods. We conclude with some remarks 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 sending a strong message. The authors present some statistical figures and point the existence of easy statistical tools to detect fabricated data and ignorance about the tools.

The authors examine patterns in radio-biological data. 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 triplicate 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 equal distances above and below that value. This would result in triples that contain the (rounded) mean as one of the values.

The methodological contribution of the paper is "bounds and estimates for the probability that a given set of n such triplicates contains k or more triples which contain their own mean" when each of the n triples is independent and identically distributed (IID) Poisson, and triples are independent of each other. (Different triples may have different Poisson rates.) For this Poisson model, the chance that the RTS's data would contain so many triples that include their 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 compare the RTS data to what would be expected for a model of the observations, then validate the test by comparing data pooled for the other researchers to the model. Pooling the data in this way may hide anomalies in the other researchers’ data. Permutation tests allow the data for the RTS to be compared to the data for the other researchers (and to compare each researcher’s data with that of the group) without positing a model for how the data were generated. On the other hand, the bulk of the data available are for the RTS, so 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–primarily shows that that researcher’s data is not like that of the RTS, not that they are suspicious. See section 3 of this review for more discussion.

2 Reproducibility of Results

This section discusses our attempts to replicate the analyses in the paper. After some trial and error and fine tuning we were able to replicate most of their results, obtaining similar results in the other cases. All our results and code are available at 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 triplet $(a, b, c), a < b < c$ as $\frac{b-a}{c-a}$, and show that the histogram of RTS concentrates abnormally around $0.4 - 0.6$ range, compared to everyone else put together. We tried to reproduce the histogram in python using the numpy’s histogram plots (and in an early test also using Matlab) and it looked very different. Then, we tweaked the histogram to include the right edge of the bins and it looked very similar to the Figure(1) of the paper. But the histogram still had differences, for instance, the authors get very close to 50% chance of obtaining a mid-ratio of 0.4-0.5, while we get close to 44% chance. Also, we used 1361 values for computing the histogram after removing the triplets with missing values (in fact, 1360 because one triplet had all equal values) while the authors used 1343/1361 and provided no justification for the same. Similarly, we had 595 triplets to plot the histogram for the rest of the researchers (of the same lab). However, our plots can be categorized very similar to theirs after the bin adjustment, and we categorized these differences too minor for investment of more time.

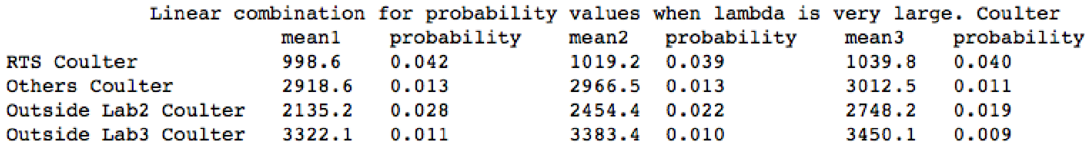
2.2 Probability Model

In this section, we followed the equations provided by the authors in Appendix A to calculate the probability - λ table. Here, first they model each triplet of observations as a three identically independent distributed (i.i.d.) Poisson random variables with mean λ (which could differ from triplet to triplet). Next they model the occurrence of mean (rounded off) in such a triplet as a Bernoulli random variable whose success probability is tabulated in Table 1 as a function of λ. They provided analytical expressions in the Appendix (which looked fine at a glance) to compute this table. We could replicate Table 1 from the paper and the trends in the values as a function of λ. However for large λ for couple of implementations we got 0 value, in place of very small values for the probabilities, and we didn’t improve our implementation.

The authors used the Table 1 in two ways to choose the probability for the Bernoulli random variables. First, they used the maximum value from Table 1 as a uniform parameter for all triplets, essentially treating all triplet as i.i.d. Bernoulli(0.42), and in the second set of results, they used the Maximum Likelihood estimate (sample mean in this case) for each triplet to find the probability of success value in the table thereby treating each triplet having a different probability of success.

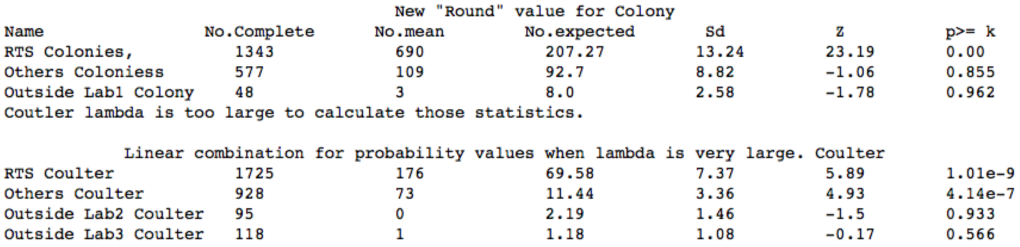
2.2.1 Using λ to obtain p-values

In this section, the researchers used their probability model calculations to compute the chance of observing the data. While replicating, it worked fine for us with the colony data as the mean of the counts < 100, and we were able to replicate their computations to minor errors. However, when we conducted the same experiments for Coulter data, due to the limitation of our implementations, we could barely come up with a reasonable probability value as the mean value of counts were a lot larger, and thus we could not replicate the values for the Coulter. We tried a regression based on the statement from the literature that when $\lambda = 100$ we use probability < 0.14, and for $\lambda = 2000$ we use probability $= 0.032$. However the take away message is hardly unaffected, and these section were not the focus of our review. For completeness we mention the interpolated probabilities for Coulter Data used for computing statistics as in Table 2 of the original paper:



```
                  Linear combination for probability values when lambda is very large. Coulter
                             mean1    probability    mean2   probability     mean3    probability
RTS Coulter                  998.6    0.042          1019.2  0.039           1039.8   0.040
Others Coulter               2918.6   0.013          2966.5  0.013           3012.5   0.011
Outside Lab2 Coulter         2135.2   0.028          2454.4  0.022           2748.2   0.019
Outside Lab3 Coulter         3322.1   0.011          3383.4  0.010           3450.1   0.009
```

Figure 1: Approximate p-values for Coulter Data



```
                                        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 2: Approximate Replication of Table 2

2.3 Digits Analysis

To find additional confirmations on the suspect of fabricated data, the authors perform two additional tests, namely *terminal digit analysis* and *pair of equal terminal digits analysis*. Both such analyses are based on existing work (and intuition) that the least significant digit of a sample is, in general, not very informative, i.e. it is reasonable to expect it to be uniformly distributed random variable.

2.3.1 Terminal digit analysis

The assumption behind this test is that for experiments including counts, the last digit of a sample represented by a big number (> 100) can be expected to be uniformly distributed. On the other hand, fabricated data often fail to show such peculiar property. The authors use the chi-square test for goodness of fit to demonstrate the fraudulent nature of RTS' samples. Our results are very similar to the ones in the paper, although not identical possibly due to the minor difference in number of data points as pointed earlier.

2.3.2 Equal digits analysis

This test follows from the assumptions made from the previous one, and the claim is that in case of genuine data, one should see an equal pair of terminal digits only in 1/10 of the samples. In this case the authors consider only big numbers (>100), to ensure the analysis of insignificant digits. In this scenario, however, the authors fail to state what kind of test they have performed (we assume again chi-square test for goodness) and how the data was pre-processed. This led us to obtain similar, but not identical results.

2.3.3 Discussion of Assumptions

We discuss the assumptions and tests in bullet points, for brevity.

- We felt that the justification for the Poisson assumption for the triplet data was given less importance. And the applicability of the model to the data was also not underlined to a desirable extent. One can possibly think of various reasons and situations where doing so is hard to justify. But, beyond our intuition we didn't investigate the validity in detail.

- Though one can argue that the parameters fitted to suspected data should not be used to test the validity of the data, we agree with the authors that such a practice only lowers the chances of the suspicion, and gives the person in question a benefit of doubt.

- The authors provide a reference for the uniformity of last insignificant digit to a work [Mosimann et al., 2002], but fail in explaining why such framework can safely be applied in this context. For instance, there might be some characteristics of the underlying biological process which prevent the last digits to be uniformly distributed. An attempt to clarify and justify this choice in the current setting would have been beneficial. The authors include here additional data, provided by three external sources (two for Coulter counts and one for Colony counts) which suffered from relatively very low number of data points. Although the authors comment on the number of these additional samples in the Discussion section, we still believe that, in the current setting, these additional samples do not help them in making a stronger case, but instead can be misleading and definitely added to our confusion.

- We reiterate that treating all the other lab investigators as a single pool and singling out RTS is not sufficient, since uniformity of the pool doesn't necessarily imply a similar property for each contributors. This is the starting point of our next section.

3 Our Analysis

The authors begin by singling out that the histogram of RTS looks anomalous compared to the rest of them put together. They assume that one is likely to observe uniform distribution for mid-ratio, and this fact is validated by the histogram of the 9 researchers put together which looks close to uniform. The first question that came to our mind which motivated this section was - how do we single out the anomalous researcher if we don't know a priori who he/she is? If we decide on the histogram as the first test, then a simple way would be to plot histogram of the mid-ratios for the data collected by all researchers individually, and look for anomalous patterns across all these plots. For sake of similarity to the authors' set up, one will detect anomaly by contrasting each researcher's histogram with the histogram of all others put together. Such an experiment gives very interesting results and also raises an important issue with this approach.

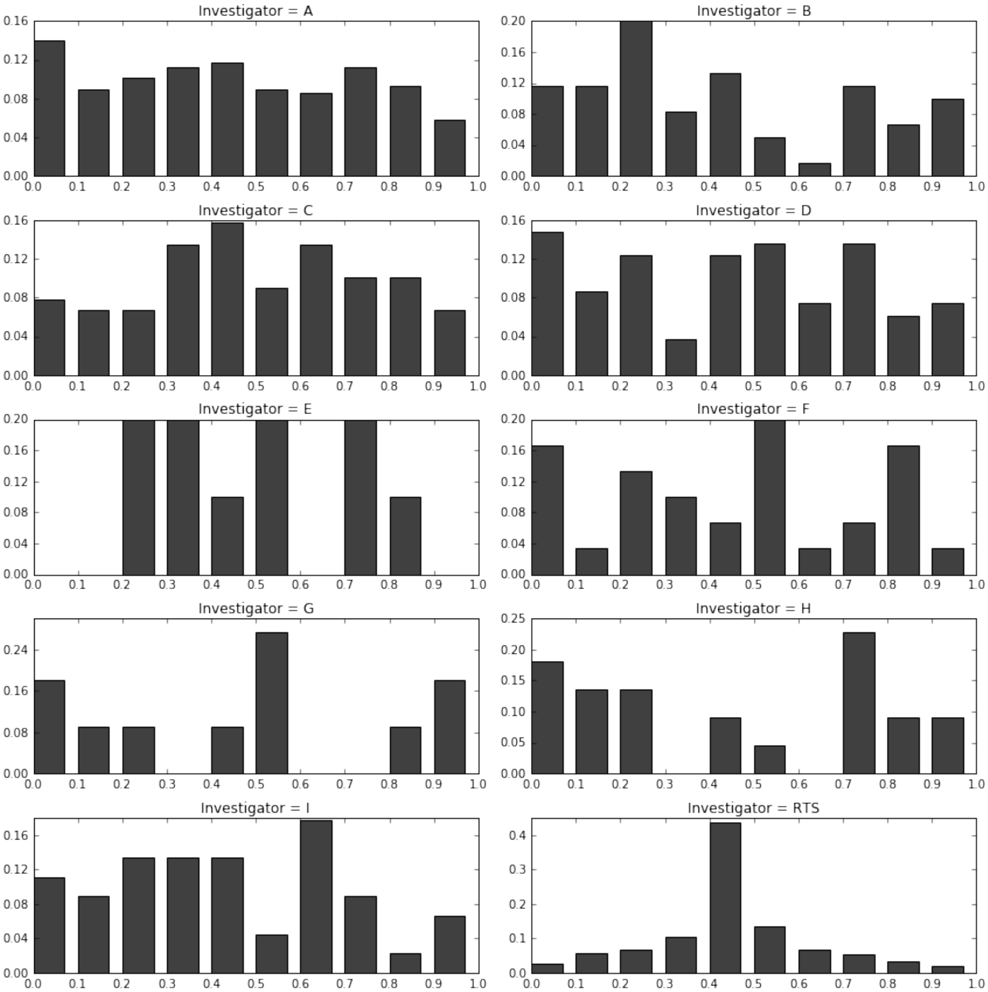


Figure 3: Individual Histograms for the Colony Data

- First, the histogram for researchers with labels "B, C, E, F, G, H, I" do not seem to be close to uniform as well. In particular, "B" and "C" have a very different histogram when contrasted with the histogram for uniform distribution. They have distinct peaks but around 0.2 and 0.4 respectively.

- Second, when we try to contrast the individual histogram of researchers with rest of them combined which includes RTS, the new "rest" histograms are dominated by RTS's data because of the comparatively huge fraction of data collected by RTS, and so most of the other researchers look anomalous when contrasted with it.

The previous two remarks point out the limitations on the visual comparison of histogram and assumption of "uniform distribution" for mid ratios. Next we try to present a different viewpoint which has two advantages - it is free of such assumptions, and thus extends to far more general cases where even slight intuition about the data is missing.

3.1 Quick Primer to Permutation Tests

As discussed above, we felt that the justification for singling out the particular RTS was incomplete. So, we took a step back, and did permutation tests to identify anomalous patterns across different researchers. We briefly discuss the test set up and the philosophy of the test.

Given a treatment and control group of size T and C respectively, we want to test the hypothesis if the treatment has an effect on the population. In permutation test, the data pooled together is considered as the population (here it will have size $N = T + C$). Next, one decides on a test statistic that is consistent with our hypothesis and is expected to contrast the two set of samples if the treatment has any effect. The distribution of test statistic has an exact theoretical representation but is often computationally intractable. An empirical approximation can be made by randomly partitioning the data into groups of T and C several times, and computing the test statistic contrasting the two datasets. With the distribution in hand, we can now test how surprising was the outcome that we originally had.

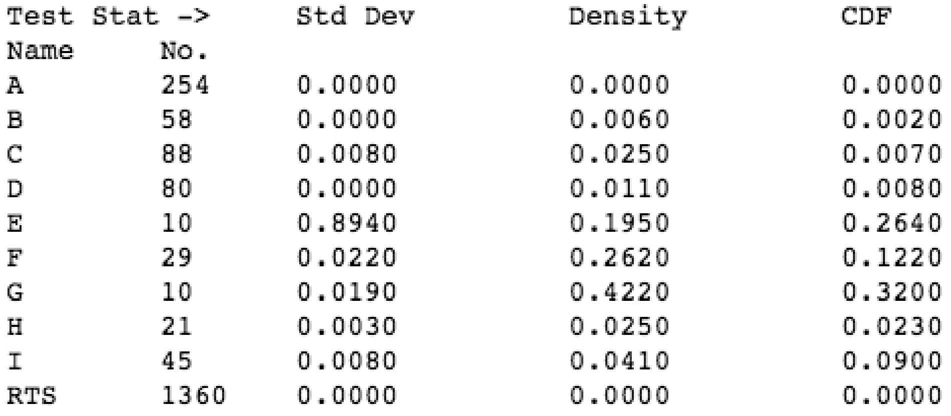
The conclusion that one draws when the p-values are very low is that *the two groups are different to each other* than expected had we randomly partitioned the pooled dataset, i.e., the labels of the data matters.

3.2 Permutation Tests for Mid-Ratio

Because we agree with the remark of the authors that it is easy to tweak the data to get a desirable triplet, we decide to set the difference in standard deviation of mid-ratios of two datasets. The choice of standard deviation as the first statistic in place of mean makes sense because uniformity as well as convenient tweaking will lead to the same expectation of 0.5; and 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. That is, for a given researcher, eg. A with dataset D_A with size n_A, we look at test statistic computed for a random partition of the entire data (size N) into two groups n_A and $N - n_A$ and compute the test statistic. We repeat this experiment 1000 times to plot the empirical distribution and then compute the p-values. We obtained 0 p-value 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 0 p-value means that there is less than 1 in 1000 chance of observing the event, because of 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.

Next we look at ℓ_1 distance between the density, followed by ℓ_1 distance between the CDF of two samples for each researcher, and obtain very similar results as in the previous case, that is several researchers will be rejected by the test at significance level of even 1%. We present all these p-values in Figure 4.

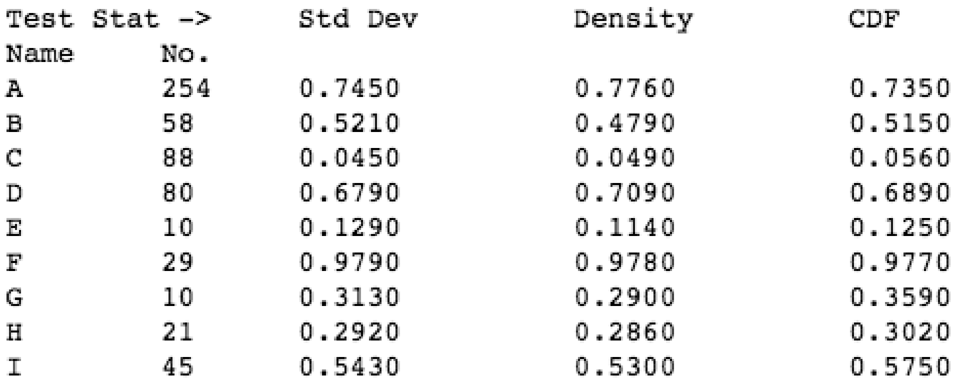


Test Stat ->		Std Dev	Density	CDF
Name	No.			
A	254	0.0000	0.0000	0.0000
B	58	0.0000	0.0060	0.0020
C	88	0.0080	0.0250	0.0070
D	80	0.0000	0.0110	0.0080
E	10	0.8940	0.1950	0.2640
F	29	0.0220	0.2620	0.1220
G	10	0.0190	0.4220	0.3200
H	21	0.0030	0.0250	0.0230
I	45	0.0080	0.0410	0.0900
RTS	1360	0.0000	0.0000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

Figure 4: Results for Permutation Tests for Mid Ratios

3.2.1 Limitations of Permutation Test

A concern in such a test is the effect of the huge fraction of the data contributed by RTS. The p-values indicate the chance of the difference between the two groups - treatment and control, so a low p-value means that the treatment group is likely to be different than the control group. And here the control group has a dominant effect from the data provided by RTS, hence a heuristic conclusion is that the data of the other lab mates is very different than the data of RTS. To be more concrete about drawing conclusions about the surprises in data about other researchers, we exclude the data provided by RTS to run the permutation tests. We will like to note that this has a bias because we ignore almost 2/3rd of the data, but doing so does give some answers that we were expecting before running these tests, which were consistent with the authors' expectations.



```
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 5: Results for Permutation Tests without RTS for Mid Ratios

Owing to the high p-values, now we may say that the data provided by each individual researchers looks like a random partitioning when compared to the entire data pooled together excluding RTS, which gives some statistical evidence to RTS being the odd one out.

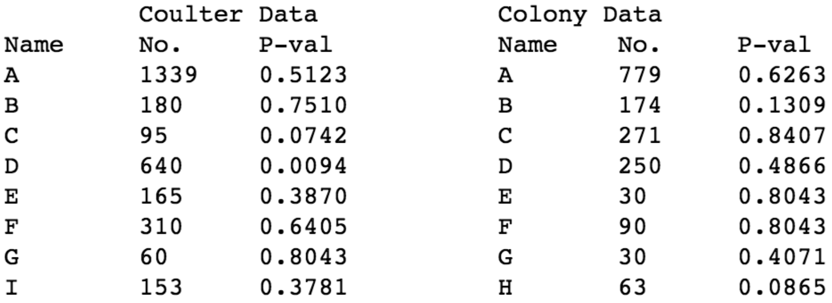
3.3 Additional Tests for Digit Analysis

For the terminal digit and equal digits tests, we extended the tests provided by the authors by considering the individual contribution of the single members of the lab and performing

- chi-square test for goodness of fit for each of the lab members and outside labs for terminal digit analysis,
- chi-square test for goodness of fit for each of the lab members and outside labs for equal digits analysis and,
- permutation tests for terminal digit analysis considering RTS and the other investigators.

3.3.1 Chi-square test Tests for Terminal Digit Analysis

To understand how single investigators contributions are distributed with respect to RTS and the outside labs, we decided to analyze data from all the other investigators taken one by one. To do so, we performed the chi-square test for goodness of fit for each of them. The following tables summarized our results:



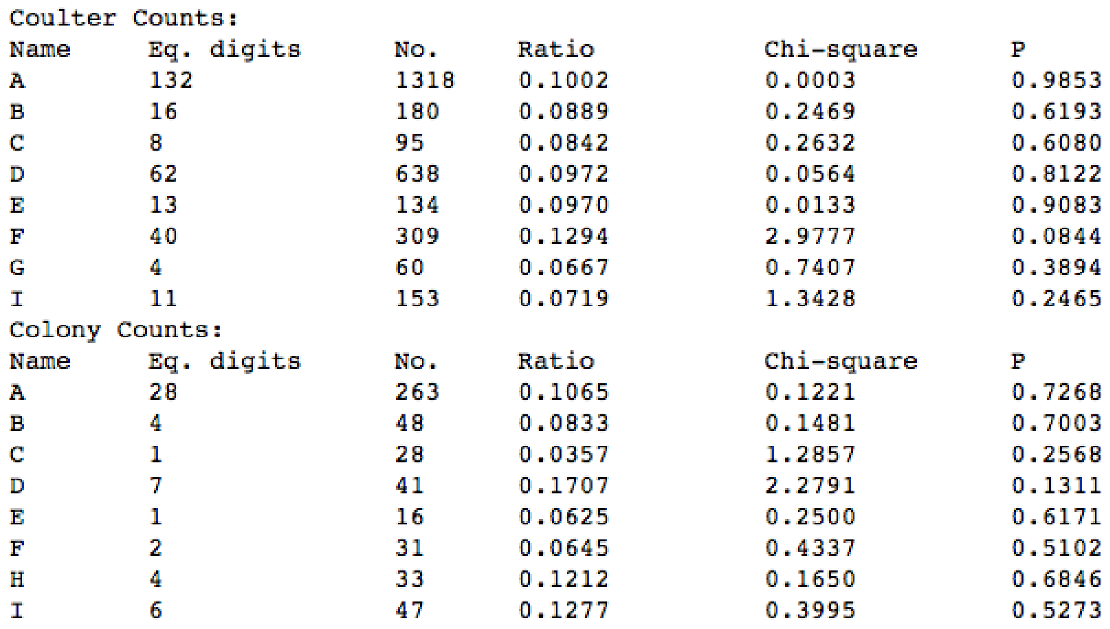
```
          Coulter Data              Colony Data
Name      No.       P-val           Name      No.       P-val
A         1339      0.5123          A         779       0.6263
B         180       0.7510          B         174       0.1309
C         95        0.0742          C         271       0.8407
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 6: Chi Square Tests for Terminal Digits in Coulter and Colony Counts

Reading the tables, one can notice that p value for D, for Coulter Data is $< 1\%$.

3.3.2 Chi-square test Tests for Equal Digits Analysis

Also for the Equal Digits Analysis we performed the chi-square test for goodness of fit using the data of the individual investigators in the lab, in a similar fashion as before.



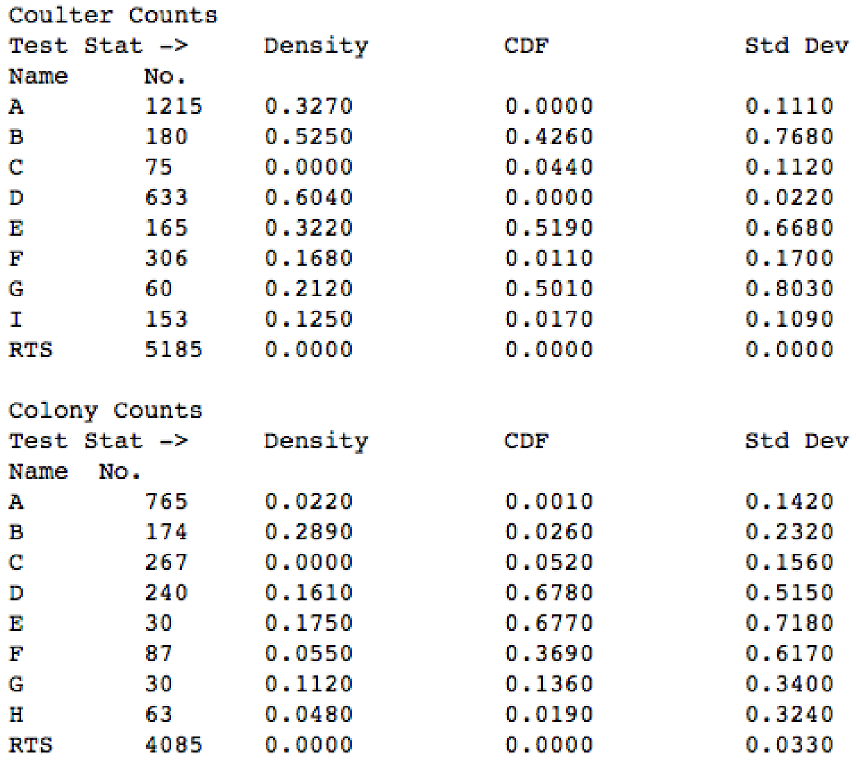
```
Coulter Counts:
Name     Eq. digits      No.     Ratio           Chi-square      P
A        132             1318    0.1002          0.0003          0.9853
B        16              180     0.0889          0.2469          0.6193
C        8               95      0.0842          0.2632          0.6080
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 7: Chi Square Tests for Equal Terminal Pair in Coulter and Colony Counts

Here none of the p-values look abnormally low. One can argue that for A it is very high, but going by the practice of deciding thresholds before seeing the results none of the results are surprising.

3.3.3 Permutation Test for Terminal Digit Analysis

The following tables illustrate the permutation test results using the same test statistics as for mid-ratios:



```
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 8: Permutation Tests for Terminal Digit Analysis, Coulter counts

In all the above cases, it is possible to see how RTS data is consistently suspicious, which is a confirmation of the authors' suspects. And as pointed before, the huge fraction of data contributed by RTS contributes towards the low p-values for other individual researchers as well. We tried permutation tests after excluding RTS data and get better p-values as before, for brevity we do not mention the values here.

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, and 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 proposing and implementing additional tests to clarify doubts and suggesting additional possibilities to the authors.

We found 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

[Mosimann et al., 2002] Mosimann, J., Dahlberg, J., Davidian, N., and Krueger, J. (2002). Terminal digits and the examination of questioned data. Accountability in Research: Policies and Quality Assurance, 9(2):75–92.