Write-ups

September 22, 2016

1 Abstract

In this article we review the paper "Statistical analysis of numerical preclinical radiobiological data". The work is submitted as a term project for the Graduate Level Course on Statistical Modelling 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.

2 Introduction

The paper begins by voicing a growing concern towards "Scientific fraud and Plagiarism" in the scientific community and is successful in presenting a strong message. To add more.....

Before proceeding further, we would like to comment that the organization of the paper could have been much better with the use of Sections and Subsections, and re-arranging some of the sections a little bit.

But the focus of this review is not on the readability and organizational strucutre, and we pay more attention to reproducing the results and discussing some more ways of identifying anomaly. In particular, in Section - **??** we perform certain tests that go very well with the spirit of the paper - promoting simple statistical tools for detecting anomaly.

2.1 Problem Set Up

The authors in the paper analyze anomalous patterns in radiobiological data from a lab, in particular they were able to detect suspicious patterns in the data reported by one of the 10 researchers (whom we shall refer to as RTS as per their notation). They do three different tests to validate their suspicion and also validate their tests and assumptions by looking at the data obtained from three other sources. To dive further, we discuss a bit more in detail about the nature of the data. Each researcher had to report three different measurements for two different types of numbers - Colony Count and Coulter Count. Each of these numbers represents an observation of number of cells surving some experiment, and probably three measurements are done in order to be more accurate about the observations. The concern of the authors is that, in case of fraudolent data,

it is easy to fabricate a triplet such that you get desired mean for that particular set of observations. One can, in fact, do that by setting the mean and then using two roughly equal constants, calculate the other two values as this initial value plus or minus the selected constants. Such a fabrication can be flagged easily by looking at the triplets and counting how many of them contain the mean as one of the three values. Having made these observations, the authors mainly focus "on developing a method to calculate bounds and estimates for the probability that a given set of n such triplicates contains k or more triples which contain their own mean" and mention that such probability bounds should be helpful across various other areas. Under these models they show that RTS's data is pretty surprising and that the chances of seeing such a data are astronomically low. Besides this specific set up (which requires some assumptions) they also look at some more general tests that have been used in the past to detect anomalous patterns. Namely they test for - (1) Distribution of the least significant digit, and (2) Chances of observing equal pairs of terminal digits. Ideally, for (1), we expect to see a uniform distribution over $\{0, 1, \ldots, 9\}$ unless the distribution that underlies the data suggests otherwise. Similarly, for (2), ideal chances of having an equal pair of terminal digits is 1 in 10.

Next we discuss three major concerns and justify why they were important to us:

- the paper begins with the RTS being labeled as anomalous and then a probability model is developed to determine the chances of seeing the mean in a triplet. The authors mentioned briefly that "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". The authors didn't comment how were they able to identify the particular researcher. Whether they partitioned the data into an observation set and then ran tests on the validation set is also unclear and the tables tend to hint otherwise. We would like to point out that a standard practice is usually to classify the data into training and test set. Our concern relies on the well known fact, in statistics, that "the data that raised the suspicion if used to validate it, will most likely give a very biased result";

- the authors also ran tests for the last digit and equality of the pair of terminal digits on the datasets, which can be seen as a validation of their suspicion. However all the results that are produced are of the form "RTS vs The Rest". It would have been more convincing if the authors presented some justification or some experiment results which justified such a treatment. The ideal scenario would have been presentation of results in a "Take - One - Out" fashion, where every individual would have been compared to the rest of them pooled together. This is the core principle behind the two sample permutation tests, where we test the strong null hypothesis that each researcher's data is just a random sample from the population of all the data put together. We will dive into this in detail in Section 4;

- there was no discussion about the number of data points across researchers. For some reason, the data collected by the RTS had more than twice the data put together by twelve other researchers. Such an overwhelming fraction of samples belonging to one researcher has some implications which we explore in Section 4.

In the next section, we touch upon the reproducibility of results. In Section 3, we discuss our tests and their implications and then make some final remarks in the Conclusion section.

3 Reproducibility of Results

In this section, we replicated the statistical experiments that were conducted by the researchers. There were several mismatches in our first implementation because of subjectivity at certain places. However, with 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 [https://github.com/ianno/stat215a_project1]

3.1 Mid-Ratio Analysis

To begin with, the authors first consider the histogram of 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 abnormaly 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.

3.2 Probability Model

In this section, we followed the equations provided by the authros in Appendix A to caluclate the probability - lambda table. We could replicate Table 1 from the paper exactly the same. However as we tried to increase lambda value to a fairly large one (>1000), the result that we could get is always 0 even if we chose to use the logrithm to avoid over floating of the computer. And thus, we could not verify lambda and corresponding values from around 1000 to 2500, which was said to be done in the literature. We did observe the probability started to decrease after lambda reached 4, and after a scanning of the whole probability values, 0.42, which is stated by the authors, should be a maximum threehold value.

However, we also came up some arguments about the Poisson assumption for the micro organism counts. As the samples were taken by the researchers, and if they did a manual count of the total organisms of the sample, there are a lot of subjectivity and biases in each of the experiment. For example, the location of the sight view to count the numbers really matter, as if the researcher accidentally chose a nutrient-rich location, the count will be way higher compared to if the sight view is taken at a relatively nutrient-low location. In addition, certain microorganisms have such characeristics that they would gather together, and thus when we do not have very high accuracy equipment, we would mistakeingly count several cells as one cell, or less cells based on our arbitary experiences. All of these situations will bring a lot of biases into the final cell count. Based on that uncertainty, it might not be precise and accurate to treat the triplet counts as a Poisson triplets, as there might be dependence between the samples.

" In this section, we tried to replicate what they did in the paper. A slight difference of the final probabilities were observed in our calculations compared to their results, where the turning point of lambda value is around 11, which means the probability keeps increasing to its maximum value

till lambda equals to 11, and then it decreases as lambda increases. However, we got a similar result for the threhold value for probability. In the paper, they got 0.42, and we got 0.433. As their details calculations are not included in this paper, so we completed the statistical calculations based on their assumptions, such as the triple is generated by independent, identical Poisson variables with known parameter lambda includes its own (rounded) mean value. We applied their assumptions of iid variables for the counts of micro organisms, and got our results. However, their assumptions may not always hold at different situations, thus needs further consideration. "

In this section, the researchers from the paper used the lambda values to calculate the corresponding p-value. They applied a heuristic method to estimate the actual probability that a given collection of n triples includes k mean containing tripes to legitimate experimental data, and such that they are able to confirm the validity of their models, which is the Poission model. As the true lambda value of the Poisson variables that generated the triples in the datasets are unknown, they took advantage of the lambda MidProb table to estimate the true value, based on the fact that mean of any actural triple is a resonable estimate of the lambda parameter of the variables. In addition, a Poisson binomial distribution is assigned to Poisson binominal random variables, which is the case in their paper. From the characteristics of a poisson binomial variables, the mean is the sum of all p values, and the standard deviation is the sum of all p*(1-p) values, which could be used for further corresponding hypothesis tests. And thus they used this idea to test the RTS collection of 1343 colony trples, and came to a conclusion that the probability is an extremely small number, which contradicted same test results for other investigators.

Several perspectives should be considered when conducting these statistical tests. We need to check the underlying assumptions such as a Poisson bernouli variables. One thing to note is that those data has underlying microbial phenomenna behind them, so when we just treat them as a number, then we would possibly lose a lot of intrinisic characteristics of the data. From this logic, the assumption that the triplets follows a poisson distribution will also be argued, as the sample would be taken from different growth stage of the organisms, biases would be introduced to devalidate the Possion assumption, which lead to the faliure of a Poisson binomial variable assumption.

In addition, the idea of using the existing questionable data to fit a parameter lambda should be considered further. If there are already frauds in the existing dataset, it may not be wise to use these data to fit our parameters. From the same logic, they calculated the mean of the data, and come up with the corresponding lambda value, which could be checked to get the p-values. Thus, it implies an idea of using questionable data to fit parameter, and then use this fitted parameter to check the questionable data, which is not very scientific.

A detailed regeneration of the method that was applied are listed here:

Raaz : Verify the Appendix and comment on the model.

3.3 Digits Analysis

To find additional confirmations on the suspect of fabricated data, the authors of the paper perform two additional tests, namely *terminal digit analysis* and *pair of equal terminal digits analysis*. Both such analyses are based on the intuition that the least significat digit of a sample is, in general, not very informative, i.e. it behaves as a uniformly distributed random variable. The authors do provide a reference to a work from J. E. Mosimann, 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 beneficial. The authors

include here additional data, provided by three external sources (two for Coulter counts and one for Colony counts). 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 add confusion. As for the mid-ratio case, we also claim that treating all the other lab investigator as a single pool is also not sufficient, since uniformity of the pool doesn't necessarily mean of the single contributors. For instance, analyzing Table 2 in the paper, it seems that the colony counts of the other investigators are even *too good*, having a p-value greater than 0.99. We will elaborate more in next sections.

3.3.1 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 fraudolent nature of RTS' samples.

Name	0	1	2	3	4	5	6	7	8	9	total	chi-square	P
RTS_Coulter	475	613	736	416	335	732	363	425	372	718	5185	467.33	5.64306e-95
RTS_Colony	564	324	463	313	290	478	336	408	383	526	4085	200.978	2.06634e-38
Rest_Coulter	261	311	295	259	318	290	298	283	331	296	2942	16.0068	0.0667397
Rest_Colony	191	181	195	179	184	175	178	185	185	181	1834	1.80328	0.99421
Out1_Coulter	28	34	29	25	27	36	44	33	26	33	315	9.70968	0.374496
Out2_Coulter	34	38	45	35	32	42	31	35	35	33	360	4.94444	0.839124
Out3_Colony	21	9	15	16	19	19	9	19	11	12	150	12.1333	0.205897

Reproducibility of Terminal Digit Analysis

As shown in the above picture, our results are very similar to the ones in the paper, although not identical.

3.3.2 Equal digits analysis

This test follows from the assumptions made from the previous one, where here 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 only not very significant 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 have been pre-processed. This led us to obtain similar, but not identical, results:

Name	equal digits	total	ratio	chi-square	P
RTS_Coulter	644	5184	0.124228	33.8121	6.0701e-09
RTS_Colony	135	1660	0.0813253	6.4324	0.0112057
Rest_Coulter	286	2887	0.0990648	0.0280568	0.866975
Rest_Colony	53	507	0.104536	0.115933	0.733489
Out1_Coulter	32	306	0.104575	0.0711692	0.789642
Out2_Coulter	30	360	0.0833333	1.11111	0.291841
Out3_Colony	1	7	0.142857	0.142857	0.705457

Reproducibility of Equal Digits Analysis

3.3.3 Using lambda to obtain p-values

In this section, the researchers from the paper used the lambda values to calculate the corresponding p-value. They applied a heuristic method to estimate the actual probability that a given collection of n triples includes k mean containing tripes to legitimate experimental data, and such that they are able to confirm the validity of their models, which is the Poission model. As the true lambda value of the Poisson variables that generated the triples in the datasets are unknown,

they took advantage of the lambda MidProb table to estimate the true value, based on the fact that mean of any actural triple is a resonable estimate of the lambda parameter of the variables. In addition, a Poisson binomial distribution is assigned to Poisson binominal random variables, which is the case in their paper. From the characteristics of a poisson binomial variables, the mean is the sum of all p values, and the standard deviation is the sum of all p*(1-p) values, which could be used for further corresponding hypothesis tests. And thus they used this idea to test the RTS collection of 1343 colony trples, and came to a conclusion that the probability is an extremely small number, which contradicted same test results for other investigators.

3.3.4 Discussion of Assumptions

Several perspectives should be considered when conducting these statistical tests. We need to check the underlying assumptions such as a Poisson bernouli variables. One thing to note is that those data has underlying microbial phenomenna behind them, so when we just treat them as a number, then we would possibly lose a lot of intrinisic characteristics of the data. From this logic, the assumption that the triplets follows a poisson distribution will also be argued, as the sample would be taken from different growth stage of the organisms, biases would be introduced to devalidate the Possion assumption, which lead to the faliure of a Poisson binomial variable assumption. In addition, the idea of using the existing questionable data to fit a parameter lambda should be considered further. If there are already frauds in the existing dataset, it may not be wise to use these data to fit our parameters. From the same logic, they calculated the mean of the data, and come up with the corresponding lambda value, which could be checked to get the p-values. Thus, it implies an idea of using questionable data to fit parameter, and then use this fitted parameter to check the questionable data, which is not very scientific.

As the data is acquired by different researchers in the biological lab, a lot of biases would be introduced, such as the skillness of the researcher to take samples, how fluent they are at this specific task, different growth situations for micro organisms, how accurate their instruments are. If we have those biases present in our data, our assumptions, such as i.i.d variables would not be valiaded, and thus our corresponding calculations will not be accurate enough. Thus based on this logic, their reasoning about the difference between the RTS data and outside lab researcher data analysis should probably be on held.

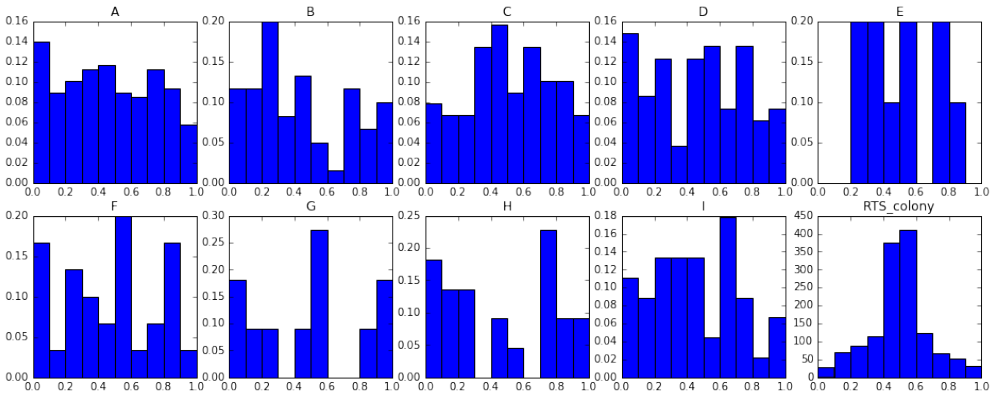
4 Our Analysis

The authors single 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 nautral question is how do we single out the anomalous resaercher if we don't know apriori who he/she is? A simple answer would be to plot histogram of the mid-ratios for the data collected by all researchers invidually, 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.

- First, histogram for researchers with labels "B, C, E, H, I" don't seem to be close to uniform as well. In particular, "B" and "C" have very different histogram when contrasted with 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 reseachers with rest of them combined which includes RTS now. In the new "rest" histgorams, RTS has a dominating effect 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 view point which besides free from such issues, as per our belief can be used in a more general and broader framework.



Individual Histograms for the Colony Data

4.1 Quick Primer to Permutation Tests

As discussed in the introduction, we felt that the justification for singling out the particular RTS was incomplete. So, we took a step back, and did some tests do identify anomalous patterns across different researchers. In order to do so, we adapted the philosophy behind permutation tests.

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.

*** Antonio, Nigel : Feel Free to Add MORE. It would be nice to have a Pseduo Code type representation. ***

4.2 Permutation Tests for Mid-Ratio

Because we agree with the remark of the authors that it is easy to tweek the data to get a desirable triplet, we decide to set the difference in standard devation 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 tweeking will lead to 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 get 0 p-value for A, B, D, and RTS; and < 0.01 p-value for all others except E, which indicates that almost all datasets are surprising with respect to this test-statistic. We would 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%.

Test Stat ->		Std Dev	Density	CDF
Name(i)	No.			
A	254	0.0000	0.0000	0.0000
B	58	0.0000	0.0000	0.0000
C	88	0.0080	0.0120	0.0040
D	80	0.0000	0.0000	0.0000
E	10	0.8940	0.9100	0.8810
F	29	0.0220	0.0190	0.0130
G	10	0.0190	0.0260	0.0150
H	21	0.0030	0.0040	0.0030
I	45	0.0080	0.0070	0.0110
RTS	1360	0.0000	0.0000	0.0000

Results for Permutation Tests

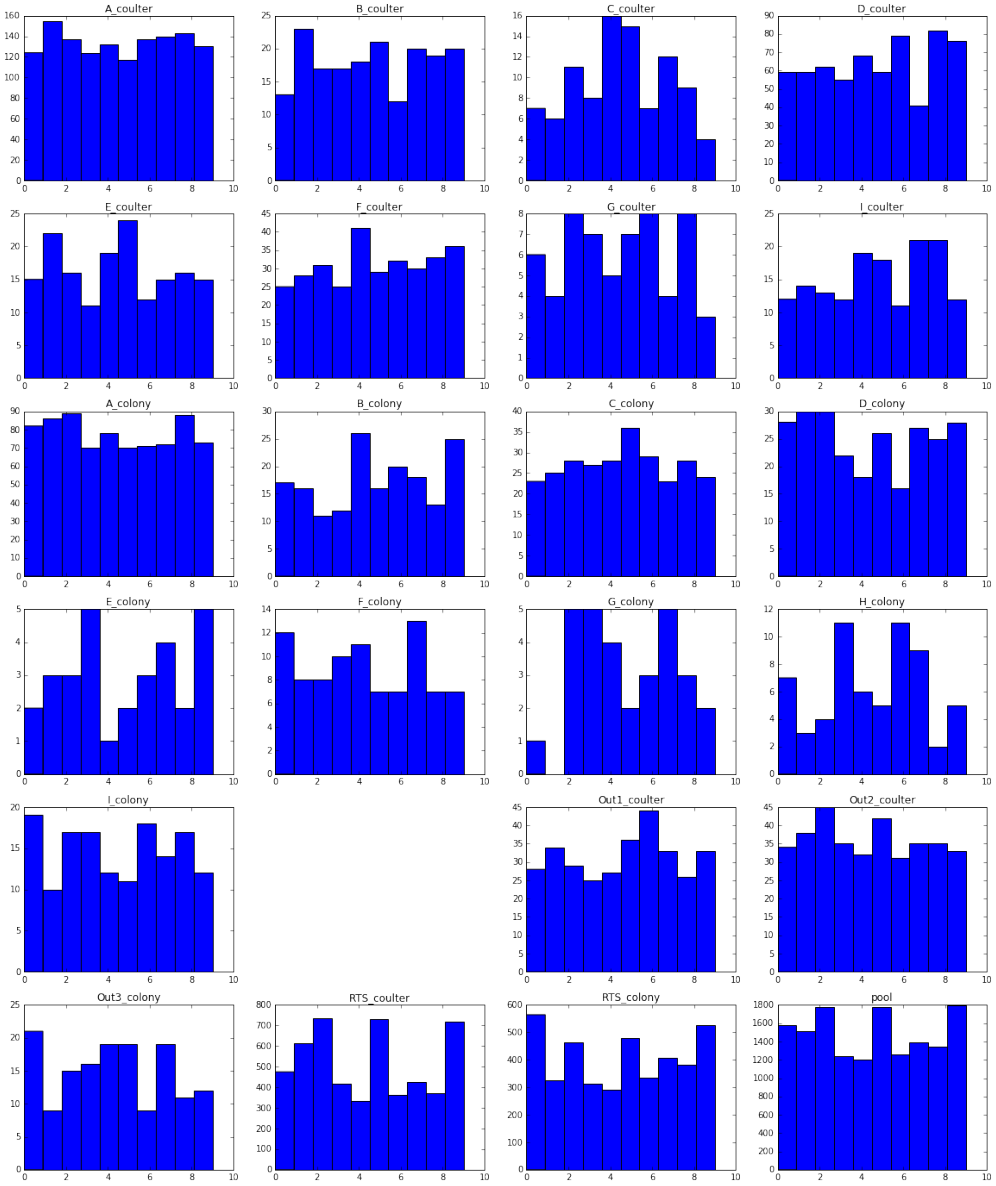
4.3 Additional Tests for Digit Analysis

One of the main concerns we have had is the decision of the authors to consider the other investigators as a single pool, instead of performing additional tests on each of them to show that even taken as single contributors their data still shows the expected behavior. 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; - permutation tests for terminal digit analysis considering RTS and the other investigators.

4.3.1 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 plotted the histograms of the individual contributions, both for Coulter and Colony counts:

Looking at the plots, it is clear that it is very hard to judge the quality of the individual contributions, since in none of those cases a uniform distribution is noticeable. The only conclusion we



Terminal digit distributions

can easily state is that the amount of data collected by RTS is incredibly higher than data collected by others.

To better understand the quality of the data collected by the single invesitgators in the lab, we performed the chi-square test for goodness of fit also for each of the individual contributors. The following table summarized our results:

Name	0	1	2	3	4	5	6	7	8	9	total	chi-square	P
A_Coulter	124	155	137	124	132	117	137	140	143	130	1339	8.21805	0.512331
B_Coulter	13	23	17	17	18	21	12	20	19	20	180	5.88889	0.750985
C_Coulter	7	6	11	8	16	15	7	12	9	4	95	15.6667	0.0741773
D_Coulter	59	59	62	55	68	59	79	41	82	76	640	21.8438	0.00938759
E_Coulter	15	22	16	11	19	24	12	15	16	15	165	9.5625	0.387049
F_Coulter	25	28	31	25	41	29	32	30	33	36	310	6.96774	0.640478
G_Coulter	6	4	8	7	5	7	8	4	8	3	60	5.33333	0.804337
I_Coulter	12	14	13	12	19	18	11	21	21	12	153	9.66667	0.378138
A_Colony	82	86	89	70	78	70	71	72	88	73	779	7.1039	0.626303
B_Colony	17	16	11	12	26	16	20	18	13	25	174	13.7647	0.130945
C_Colony	23	25	28	27	28	36	29	23	28	24	271	4.92593	0.840717
D_Colony	28	30	30	22	18	26	16	27	25	28	250	8.48	0.486588
E_Colony	2	3	3	5	1	2	3	4	2	5	30	5.33333	0.804337
F_Colony	12	8	8	10	11	7	7	13	7	7	90	5.33333	0.804337
G_Colony	1	0	5	5	4	2	3	5	3	2	30	9.33333	0.407091
H_Colony	7	3	4	11	6	5	11	9	2	5	63	15.1667	0.0864585
I_Colony	19	10	17	17	12	11	18	14	17	12	147	7.21429	0.61482

Terminal Digit Analysis

Reading the table, one can notice that Investigators C and D (for Coulter counts) and H (for Colony counts) are also quite low if compared to the others.

4.3.2 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, usign a similar approach than the Terminal Digit Analysis.

Results are summarized in the following table:

Name	equal digits	total	ratio	chi-square	P
A_Coulter	132	1318	0.100152	0.000337211	0.985349
B_Coulter	16	180	0.0888889	0.246914	0.619257
C_Coulter	8	95	0.0842105	0.263158	0.607959
D_Coulter	62	638	0.0971787	0.0564263	0.812236
E_Coulter	13	134	0.0970149	0.013267	0.908301
F_Coulter	40	309	0.12945	2.97771	0.0844189
G_Coulter	4	60	0.0666667	0.740741	0.389424
I_Coulter	11	153	0.0718954	1.34277	0.246545
A_Colony	28	263	0.106464	0.122095	0.726773
B_Colony	4	48	0.0833333	0.148148	0.700311
C_Colony	1	28	0.0357143	1.28571	0.256839
D_Colony	7	41	0.170732	2.27913	0.131125
E_Colony	1	16	0.0625	0.25	0.617075
F_Colony	2	31	0.0645161	0.433692	0.510183
H_Colony	4	33	0.121212	0.164983	0.684609
I_Colony	6	47	0.12766	0.399527	0.527334

Equal Digit Analysis

In this case, data from other investigators is more consistent than the Colony counts, probably because of the higher number of samples available.

4.3.3 Permutation Test for Terminal Digit Analysis

The same considerations we elaborated to justify the permutation tests statistics for the mid-ratio scenario also hold for the Terminal Digit Analysis.

The following plots illustrate the permutation tests results using the ℓ_1 distance between density functions and between CDF's, both for Coulter and Colony Counts:

Coulter Counts			
Test Stat ->		Density	CDF
Name(i)	No.		
A	1215	0.0000	0.1250
B	180	0.4260	0.7460
C	75	0.0250	0.1040
D	633	0.0000	0.0240
E	165	0.4750	0.6660
F	306	0.0070	0.1570
G	60	0.5000	0.8010
I	153	0.0240	0.1110
RTS	5185	0.0000	0.0000
Colony Counts			
Test Stat ->		Density	CDF
Name(i)	No.		
A	765	0.0000	0.1310
B	174	0.0380	0.2340
C	267	0.0500	0.1320
D	240	0.6960	0.5390
E	30	0.6550	0.7260
F	87	0.3890	0.6340
G	30	0.1340	0.3410
H	63	0.0130	0.3280
I	135	0.3240	0.8400
RTS	4085	0.0000	0.0270

Permutation Tests for Terminal Digit Analysis, Coulter and Colony 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.

5 Conclusion

Data fraud is an extremely critical issue in science, enginering 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 scentific 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 do not leave any space to ambiguity. Reproducibily 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; - proposing and implementing additional tests to clarify doubts and suggesting additional directions 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 pre-processed). Moreover, we suggested setting up additional tests, including permutation tests, to clearly understand how every single investigator's data, besides the RTS, compares to the general data pool. At the end of our review, we do believe that RTS has fabricated data, but we suggest the authors to collect additional material and investigate more, since our tests suggest that other investigators's data present anomalies as well.

In [ ]: