Summary of Analysis Performed in Paper

Asmit Bhowmick[1*], Sukanya Sasmal[1], Caroline Sofiatti[2]
[1]*Department of Chemical and Biomolecular Engineering,*
[2]*Department of Physics*
University of California, Berkeley, CA 94720, United States

Introduction:

In this paper, the researchers have tried to use standard statistical analysis tools to analyze scientific generated in a research project and to identify whether any of the 10 researchers have engaged in any data fraud. The authors have specifically looked at three aspects of the data - i) the probability of obtaining mean containing triplets ii) terminal digit analysis and iii) equal digit analysis. Based on their observations, they have concluded that the data generated by one of the investigators (RTS) seemed to have been manipulated. In our review, we have tried to reproduce some of the main results presented in the paper and have discussed the validity of the key assumptions made in the study.

Methodology

The authors looked at data sets containing triplets from 10 researchers, with each of the triplets reporting on Coulter count of a cell culture grown under certain conditions. It is to be noted here that the authors first visually went through the raw data and suspected researcher RTS of fraud. All the analysis done in the paper, were done to verify whether researcher RTS has fabricated the data, rather than finding out which of the 10 researchers (if any) fabricated the data.

i) Probability of obtaining mean containing triplets
In this analysis, the authors hypothesized that if the dataset has been fabricated, then it is more likely that the triplets will contain their own mean. As a first pass test, the authors studied the mid-ratio probability, defined as (middle-low value) divided by (high- low value) for the triplets. We were able to reproduce the data given in the paper (figure 1), and indeed for researcher RTS, the data was more biased towards a mid-ratio of 0.5, indicating that the many of the triplets contained their own mean. We also analysed the data separately for each of the researchers. Ideally, one would expect the mid-ratio distribution for each of the researchers to be uniform. But the data isn't uniform for other researchers, in particular researcher I, E, H, F, and G. One can ignore researcher E, H, F and G, since they have performed lesser number of experiments as compared to other researchers, but researcher I doesn't pass this test. Although it is clear that his/her data doesn't contain abnormal number of mean containing triplets, it is not uniform either. One can argue, that since the distribution is not uniform, there is a likelihood that the data is manipulated in some other way.

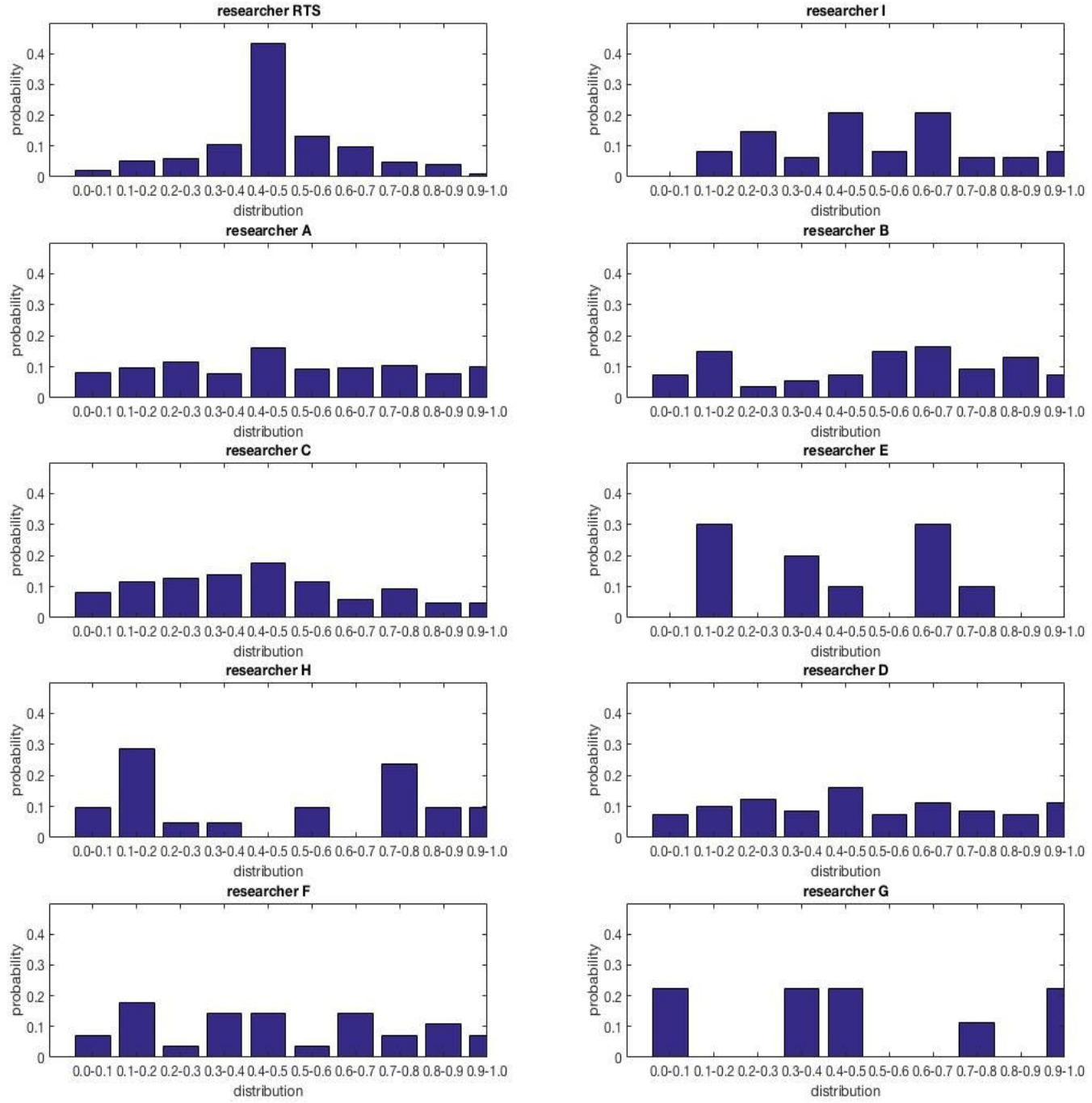

Figure 1: Distribution of mid-ratios (middle -low) / (high-low) for colony triplets for different researchers.

Why Poisson distribution?

Since, each data for the triplets come from three separate test-tubes, the assumption that the triplets are independent is valid. Also, cell growth is a first-order process for the initial phase (3

days in this case). So modelling the triplets as independent Poisson random distribution is a reasonable assumption.
The authors next tested three hypotheses for the occurrence of mean containing triplets in a given set of independent Poisson random variables. We looked into the first two in this review.

a) Hypothesis Testing 1:
The authors modelled the probability of occurrence of k- mean containing triplets for sample size n as independent Bernoulli's trials, which is reasonable for a non-parametric test. We can feed in the data points in the link below by Wolfram Alpha to check if the probabilities match.
https://www.wolframalpha.com/input/?i=probability+of+690+successes+in+1343+trials+with+p%3D0.42
The parameters used and which were reported in the paper are
N = 1343, k = 690, p=0.42 for RTS
N= 572, k = 109, p = 0.42 for others
We were able to reproduce the reported probabilities for this test.

b) Hypothesis Testing 2:
In this section, the authors have estimated the probability of occurrence of mean-containing triplets for different λ values. The probability increases till λ=4, after which it starts to decrease. We were able to get the same trend using a bootstrap calculation as given in Table 1.

Table 1: Replicating partial MidProb table (Table 1 in paper). P is the probability that a triple generated from a Poisson random distribution with parameter λ will contain the mean. The probability values are based on bootstrap calculations with 200,000 trials. The values in the brackets are from the paper.

λ		λ		λ		λ		λ	
1	0.287 (0.267)	6	0.388 (0.372)	11	0.324 (0.317)	16	0.285 (0.281)	21	0.257 (0.254)
2	0.343 (0.387)	7	0.372 (0.359)	12	0.316 (0.309)	17	0.281 (0.275)	22	0.254 (0.250)
3	0.434 (0.403)	8	0.362 (0.348)	13	0.308 (0.301)	18	0.274 (0.269)	23	0.250 (0.246)
4	0.419 (0.397)	9	0.347 (0.337)	14	0.299 (0.294)	19	0.267 (0.264)	24	0.246 (0.242)
5	0.404 (0.385)	10	0.336 (0.327)	15	0.292 (0.287)	20	0.264 (0.255)	25	0.240 (0.238)

ii) Terminal digit analysis

The authors assumed that the terminal digits from the Coulter and Colony data would be uniformly distributed since they are the least significant. First we attempt to reproduce the data reported in Table 3 of the paper. For some of the data provided, we could not exactly reproduce the numbers reported in the paper. However, the overall results and trends are consistent with what has been reported.
The jupyter notebook TermDigits_Table3.ipynb was used to generate data in table 3.

Table 2: Replicating table 3 in paper on terminal digit analysis of coulter and colony count. Red indicates the lab data for which we couldn't relicate the data exactly

		Digits												
Type	PI	0	1	2	3	4	5	6	7	8	9	Total	Chi2	P
Coulter	RTS	475	613	736	416	335	732	363	425	372	718	5185	466.7	0
Coulter	Others	261	311	295	259	318	290	298	283	331	296	2942	16.0	0.067
Coulter	Outside lab-1	28	33	28	25	27	36	43	33	26	33	312	8.8	0.45
Coulter	Outside lab-2	34	38	45	35	32	42	31	35	35	33	360	4.9	0.84
Colony	RTS	564	324	463	313	290	478	336	408	383	526	4085	200.7	0
Colony	Others	191	181	195	179	184	175	178	185	185	181	1861	1.79	0.994
Colony	Outside lab-3	21	9	15	16	19	19	9	19	11	12	150	12.1	0.21

Since the trends and data are reproducible (except for a few data points that may have something to do with how the raw data was processed), we wanted to see if another test gives us the same conclusion.

We next tried the chi2 test by normalizing the data and redoing the analysis. The differences are not as glaring anymore between the RTS and others (Table 2). We also did the KS-test on the normalized data and found that the p-values are practically 0 for all the data sets, implying we cannot reject the null hypothesis that the data came from a uniform distribution. This is applicable to the RTS data as well. The analysis reveals how sensitive the conclusions can be depending on how the data is processed.

One of the reasons that the chi2 values obtained by the authors for RTS is so high is because the data was not normalized and RTS has conducted 2-35X more experiments than the other labs. Hence, any slight discrepancy in the values were amplified for RTS by virtue of having conducted more experiments. When the data was normalized, this effect of sample size was removed.

Table 3: Chi2 values for terminal digit analysis, this time done by normalizing the counts within each data set. The observed differences between RTS and other researchers in Table 2 disappear dramatically.

Type	PI	Chi2	Chi2 from normalized data	P	P from normalized data
Coulter	RTS	466.7	0.09	0	~1
Coulter	Others	16.0	0.005	0.067	~1
Coulter	Outside lab-1	8.8	0.028	0.45	~1
Coulter	Outside lab-2	4.9	0.013	0.84	~1
Colony	RTS	200.7	0.049	0	~1
Colony	Others	1.79	0.0009	0.994	1.0
Colony	Outside lab-3	12.1	0.081	0.21	~1

Table 4: Chi2 and P-values for coulter counts of other researchers in the lab. RTS data is provided for reference. Note that researcher D could be labeled as suspicious depending on the level of significance chosen (which is usually 0.01-0.05 in most scientific works)

Researcher	Chi2	P-value
RTS	466.7	0
A	8.1	0.52
B	5.9	0.75
C	14.6	0.1
D	21.8	0.009
E	9.1	0.42
F	7.0	0.64
G	5.3	0.8
I	9.4	0.4

We also did a chi-square test on each individual scientists in the same lab that conducted experiments. The data used was the Coulter counts.In the excel file provided by Pitt and Hill,

they are names A through I with H omitted. Table 4 summarizes our finding and Jupyter notebook Breakup_byScientists_TermDigits.ipynb has been provided on github that was used to generate this table. For comparison, statistics for RTS obtained on his/her coulter counts is also provided in table 4. Interestingly, researcher D can also be labeled suspicious by this statistical test based on what is our threshold level for P-value.

iii) Equal digit analysis

Lastly, the authors hypothesized that if the data was manipulated then the last two terminal digits for the subset of experimental data with values larger than 99, will be expected to be equal more frequently. Under normal conditions, the probability of last two digits being equal is 10%. The authors again used the principle of Bernoulli's trial to show that the investigator RTS has more than expected equal last two terminal digits. We were able to reproduce the result using Wolframalpha.
https://www.wolframalpha.com/input/?i=probability+of+636+successes+in+5155+trials+with+p%3D0.1
We would like to point out here that the authors did not investigate each of the other researcher separately in this study, and RTS might not stand out if the study was performed for each researcher as in the case of terminal digit analysis.

Conclusion:

This review had 2 main aims - 1) Reproduce the data in the paper by Pitt and Hill and 2) discuss and test some of their assumptions/methods by carrying out further tests. We were able to reproduce their data fairly well except for a few places where the type of data processing left out some points that was not mentioned clearly in the paper. Our scripts on github clearly show what we did and it would be great if we can look at the authors' R scripts to understand their methods.
With regards to the 2nd aim, one of the key takeaways from our additional tests was the data from all other researchers should not have been combined. This kind of mixing of data masks flaws that they were not looking for. In Figure 1 and Table 4, splitting the analysis by researchers revealed trends that belie their assumption of uniformity in mid ratio values and terminal digit distribution. It appears that they went into the investigation with the notion that only one type of data manipulation was done and that too by one researcher. Given the limited amount of data from other researchers in the lab, the deviation from the assumed behavior of the data should elicit prudence in future statistical scrutiny.
Beyond these concerns however, the overall analysis did a good job to statistically differentiate RTS from other researchers using multiple tests. Most of the other assumptions (like poisson process for colony counts) were reasonable. We also observed that normalizing the data can affect conclusions as seen in Table 3.

Data available:

Link to paper reviewed:
https://www.scienceopen.com/document?vid=8aa0f248-2bad-44c6-adfd-42816c14c272

Link to data:
https://osf.io/hg3rc/
Analysis:
Python scripts used in this analysis can be found at:
https://github.com/sofiatti/stat_analysis_radiobiological_data

Contributions:
All the authors contributed equally.

Acknowledgements:
The reviewers thank the authors Joel H. Pitt and Helene Z. Hill for making the data available online for this review. This review was vetted by Philip B. Stark. However, the work was conducted entirely by the authors, and the opinions expressed in this review are those of the authors.

** Corresponding author for this review:*
asmit@berkeley.edu