This review is a term

project for the Graduate Level Course -- Statistical Modeling and

Practices at University of California Berkeley. The authors are graduate

students from department of Elecrical engineering and Computer Science and Civil &Environmental Engineering

and have restricted their attention to the methods and analysis conducted 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, which can be seen as an alternative to

making certain assumptions and finding surprises in the data. No attempt

has been made to investigate from the biological aspects and validity of

certain assumptions associated with 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.

We review the paper in the spirit of promoting reproducibility of research and attempting 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, which lead 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 comments on the methods used. In Section~\ref{reproducibility-of-results} we replicate the authors' work and results to some extent.

In Section~\ref{our-analysis}, TODO rewrite(what is this) :we discuss and point out some gaps as a reader and attempt to take a step back and conduct additional tests using statistical methods

We conclude with some remarks in Section~\ref{conclusion}.

The paper begins by voicing a growing concern towards the topic ``Scientific fraud

and Plagiarism'' in the scientific community and is successful in

advocating a strong message. The authors present some statistical figures and illustrate the existence of easy statistical tools to detect fabricated data and ignorance related to the tools.

The authors examine the 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 other 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 neighboring the desired mean, which would result in triples containing 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 from each other. (Different triples may have different Poisson rates.)

For this Poisson model, chances that the RTS's data would contain triples that include their (rounded) mean should be 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 of 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, which raised the suspicions about the RTS, was used to test whether the RTS's data were anomalous based on the feature. If so, the nominal p-values are likely to be misleadingly small.

\item

Most of the tests compare the RTS data to what would be expected for a model of the observations, then the authors validate the test by comparing data pooled from other researchers to the model. Pooling the data in this way may hide anomalies in other researchers' data. Permutation tests allow the data for the RTS to be compared to the data from other researchers (and to compare each researcher's data with those from the group) without positing a model for how the data was generated. On the other hand, the majority of the data available are from the RTS, so as 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 (this Is too complicated, better if we spilt this long sentence to 2-3 shorter ones). See section~\ref{our-analysis} of this review for more discussion.

\end{itemize}

This section discusses our attempts to replicate the analyses in the paper.

After several trials and errors 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 \hyperlink{https://github.com/ianno/stat215a\_project1}{github}[github.com/ianno/stat215a\\_project1]. We first discuss specifics about the replication and then comment about the tests and methods involved.

\subsection{Mid-Ratio Analysis}\label{mid-ratio-analysis}

The authors first consider the mid-ratio, which is defined as (abc) for a, and show that the histogram of RTS concentrates

abnormally within the \(0.4-0.6\) range, compared to anyone else’s 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 quite different. Then, we tweaked the histogram to include the

right edge of the bins , which looked more 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 (they talked about the gaps, max-min > 2).

Similarly, we had 595 triplets to plot the histogram for the rest of the

researchers (from the same lab). However, our plots can be categorized

very similar to theirs after the bin adjustment, and we characterized

these differences too minor for investment of more time.

In this section, we followed the equations provided by the authors in

Appendix A to calculate the probability - $\lambda$ table. Here, first they model each triplet of observations as a three identically independent distributed (i.i.d.) Poisson random variables with mean $\lambda$ (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 $\lambda$.. We could

replicate Table 1 from the paper and the trends in the values as a function of $\lambda$. However for large $\lambda$ 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 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(p = $0.42$), 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.

T the researchers used their probability model calculations to compute the chance of observing the data (this needs correction). 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 (This affects the calculations for later coulter table 2). For completeness we mention the interpolated probabilities for Coulter Data used for computing statistics as in Table 2 of the original paper:

\begin{figure}[H]

\centering

\includegraphics[width=0.9\linewidth]{images/Lambda\_Coulter.png}

\caption{Approximate $p$-values for Coulter Data}

\end{figure}

\begin{figure}[H]

\centering

\includegraphics[width=0.9\linewidth]{images/HT\_Stat\_values.png}

\caption{Approximate Replication of Table 2}

\end{figure}

\subsection{Digits Analysis}\label{digits-analysis}

To find additional confirmations on the suspect of fabricated data, the

authors perform two additional tests, namely \textit{terminal

digit analysis} and \textit{pair of equal terminal digits analysis}. Both

such analyses are based on existing work that the least significant digit of a sample is, in general, not very informative, i.e.~it is reasonable to expect it to be a uniformly distributed random variable.

\subsubsection{Terminal digit analysis}\label{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 which is possibly due to the minor difference in the number of data points as illustrated earlier.

\subsubsection{Equal digits analysis}\label{equal-digits-analysis}

This test follows the assumptions made from the previous one (what are the assumptions), and the claim is that as for 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 (\textgreater{}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.

\subsubsection{Discussion of

Assumptions}\label{discussion-of-assumptions}

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

\begin{itemize}

\item 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 degree. One can possibly think of various reasons and situations where doing so is hard to qualify. But, beyond our intuition we didn't investigate the validity in detail. ( I listed several senerios but got deleted tho)

\item Though one can argue that the parameters fitted to the 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. (This contradicts what I originally meant, so no change from me)

\item The authors provide a reference for the uniformity of last insignificant digit to a work \cite{mosimann2002terminal}, 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 additional data, provided by three

external sources (two for Coulter counts and one for Colony counts) which suffered from relatively small total 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

set up, these additional samples would not help them in making a stronger

case, but instead can be misleading and definitely contributing to our confusion.

\item 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.

\end{itemize}

\section{Our Analysis}\label{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 a 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 the 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 us very

interesting results and also raises an important issue with this

approach.

\begin{figure}[H]

\centering

\includegraphics[width=0.8\linewidth]{images/new\_mid\_ratio.png}

\caption{Individual Histograms for the Colony Data}

\end{figure}

\begin{itemize}

\item

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 a histogram for

uniform distribution. They have distinct peaks but around 0.2 and 0.4

respectively.

\item

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 the rest of the data.

\end{itemize}

The previous two remarks point out the limitations on the visual

comparison of histogram and assumptions of ``uniform distribution'' for

mid ratios. Next we try to present a different perspective 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.

\subsection{Quick Primer to Permutation

Tests}\label{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 before.

Given a treatment and control group of size \(T\) and \(C\)

respectively, we want to test the hypothesis that 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 the 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 the outcome was from what

we originally had.

The conclusion that one draws, when the p-values are very low is that \textit{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. ( I DON’T quite understand this)

\subsection{Permutation Tests for

Mid-Ratio}\label{permutation-tests-for-mid-ratio}

As the fact that it is easy to tweak

the data to get a desirable triplet, we decide to set the difference within the

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 \textit{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 statistics 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 calculate the corresponding 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 one if one looks at the location of the test-statistic in the

tails of the distribution.

Next we look at \(\ell\_1\) distance between the density, followed by

\(\ell\_1\) distance between the CDF of two samples for each researcher,

and obtain very similar results as in the previous case, which is the case that several

researchers will be rejected by the test at significance level of even

\(1 \%\). We present all these $p$-values in Figure~\ref{mid\_ratio\_perm}.

\begin{figure}[htbp]

\centering

\includegraphics[width=0.8\linewidth]{images/mid\_ratio\_perm.png}

\caption{Results for Permutation Tests for Mid Ratios}

\label{mid\_ratio\_perm}

\end{figure}

\subsubsection{Limitations of Permutation Test} % (fold)

\label{ssub:limitations\_of\_permutation\_test}

% subsubsection limitations\_of\_permutation\_test (end)

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 from the control

group. And here the control group has a dominant effect due to the data

provided by RTS, hence a heuristic conclusion is that the data from

other lab mates is very different from 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 and rerun the

permutation tests. We would like to note that this has a bias because we

ignore almost 2/3rd of the data, but by doing so, we do obtain some answers

that were expected before running these tests, which were consistent with the authors' expectations.

\begin{figure}[htbp]

\centering

\includegraphics[width=0.8\linewidth]{images/mid\_ratio\_perm\_no\_rts.png}

\caption{Results for Permutation Tests without RTS for Mid Ratios}

\end{figure}

Owing to the high $p$-values, now we would 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.

\subsection{Additional Tests for Digit

Analysis}\label{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

\begin{itemize}

\item chi-square test for goodness of fit

for each of the lab members and outside labs for terminal digit analysis,

and equal digits analysis and,

\item permutation tests for terminal digit analysis considering RTS and the other investigators.

\end{itemize}

\subsubsection{Chi-square test Tests for Terminal Digit

Analysis}\label{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:

\begin{figure}[H]

\centering

\includegraphics[width=0.7\linewidth]{images/raaz\_term\_chi\_summary.png}

\caption{Chi Square Tests for Terminal Digits in Coulter and Colony

Counts}

\end{figure}

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

\subsubsection{Chi-square test Tests for Equal Digits

Analysis}\label{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.

\begin{figure}[H]

\centering

\includegraphics[width=0.9\linewidth]{images/raaz\_eq\_chi\_elaborate.png}

\caption{Chi Square Tests for Equal Terminal Pair in Coulter and Colony

Counts}

\end{figure}

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. (ok)

\subsubsection{Permutation Test for Terminal Digit

Analysis}\label{permutation-test-for-terminal-digit-analysis}

The following tables explains the permutation test results using the

same test statistics as for mid-ratios:

\begin{figure}[H]

\centering

\includegraphics[width=0.7\linewidth]{images/raaz\_eq\_perm\_summary.png}

\caption{Permutation Tests for Terminal Digit Analysis, Coulter counts}

\end{figure}

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.

\section{Conclusion}\label{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

pre-processed). 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 validate data without introducing any assumptions, thereby shifting the focus from debate on assumptions to actual present anomalies and to a better understanding of each individual

investigator's data (besides the RTS) as to how 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.

\section\*{Acknowledgments} % (fold)

\label{sec: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.

% section acknowledgments (end)

\bibliographystyle{apalike}

\bibliography{biblio}

\end{document}