Write-ups

September 21, 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 Environmental Sciences 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 ind etail 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 represent 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 it is easy to fabricate a triplet such that you get desired mean for that particular set of observations, and one can 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. 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 require 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 last digit when it is insignificant and (2) Chances of observing equal pairs of terminal digits when they are not significant. 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 are 1 in 10.

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

- The paper begins with the RTS guy being labelled 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 pretty unclear and the tables tend to hint otherwise. We would like to point 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 "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 Secion 4.

- There was no discussion about the number of data points across researchers. For some reason, the data collected by the researcher in question, namely 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. 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 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

*** Nigel : Please add details about your trials, and your concern about the choice of poisson. ***

I AM PASTING YOUR LINES FROM THE SECTIONS BELOW:

3.2.1 Introduction

" 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.

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 validated, 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.

" ### Calculations

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.

*** Antonio : Feel free to add your comments about the model. ***

3.3 Digits Analysis

*** Antonio : Please elaborate your tests. ***

3.4 Normal Estimation of p-values

Nigel : Please complete this section, mentioning your results.”

From the literature, the researchers obtained reasonable approximations of the upper tail probabilities of a Poisson binomial random variable using normal probabilities. This idea is obtained from Central Limit Theorem, and thus when we did the replication, we directly applied this idea to calculate the corresponding upper tail probability from Z scores. However, as the normal distribution could not catch all characteristics of a Poisson bbinominal distribution, a lot of considerations were taken into account by the researchers, such as implementing a second-order correction.

Thus, using the probability values obtained from previous calculation, they were able to calculate the corresponding mean and standard deviation for the Poission binominal variables, which were further assigned to be the mean and standard deviation for the normal distribution assimilation. The researchers also noted that the normal distribution probabilities are not exact values for the Poisson binominal probabilities, thus this biases introduced by using normal distribution to simulate the original distribution should be considered if we are going to deal with further analysis.

However, as we mentioned in previous sections about their underlying assumptions for the past procedures to calculate the lambda-probability table based on poisson process assumption, and also use the questionable existing data to fit lambda values to get p-values. A lot of uncertainties were introduced by using the uncertain, inaccurate results to calculate the upper tail probability even if the assumption for this section looks reasonable based on Central Limit Theorem. Thus, our results are not very close to what they got in their literature.

A detailed calculation is as follows:

4 Our Analysis

4.1 Ideal Scenario

An ideal scenario would have been to decide on tests before looking at the data, i.e., decision of tests to be conducted is taken independent of the data, to avoid any bias. We do not know if the authors decided to test for the mid-ratio before or after looking at the data, but assuming that is a natural test for data that comes in triplet, next we have to decide on some test statistic.

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 visually 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.

4.2 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 partitoning 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. ***

4.3 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 stastic 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 signifiance level of even 1%.

4.4 Permutation Tests for Digit Analysis

> **Antonio : Please add stuff here. You can directly discuss the results and add philosophical stuff in the sections above.**

4.4.1 Terminal Digit Analysis

4.4.2 Equal Digit Analysis

5 Conclusion

> *** Antonio : Please work on this section. ***

Data fraud is an extremely important topic in science, enginering and many other subjects. Methods to detect the manually manipulated data are needed to identy the existance of data fraud, data fabrication and falsification. In our review, we conducted permutation test, blah blah tests to deterine if there is any underlying data fabrication and falisicaiton of this paper.

From permutation test, we found…

From … test, we found that…

Examples of citations: CITE or CITE.