Write-ups

September 21, 2016

1 Introduction

This paper reviews some necessary considerations of the published paper "Statistical analysis of numerical preclinical radiobiological data". Scientific fraud and data manipulation have become more and more serious problems for scientific researchers. Statistical models to identigy potential data fabrication needs to be utilized more, which could help to determine data fabrication and falsification, suspected fraud, duplicate publication, and plagiarism [Ref_1]. Specifically for radiobiological experiment results in data sets consists of triplicate counts where the means of the triples are the key values in subsequent analyses. Potentially, researchers would be likely to manipulate the data to guide the results of such experiments, so that the average values could be within a desired region. And thus, in this manner, mid-ratio (the ratio of the difference between the middle value and the smallest value to the difference between the largest value and the smallest value was close to 0.5 or an unusually large number of triples that actually include their own (rounded) mean as one of their values) [Ref_2]. The least significant (rightmost) digits of radiobiological data have been expected to follow a uniform distribution, which could be applied to detect if there is any fraud in the data provided by other researchers, when the least significant digit does not follow that behavior. Similar expections of equal digits of non-important digit could also applied to detect fraud and manuulay manipulated data by data. By doing that, researchers made several assumptions, such as a Poisson distribution model was used to probabilistically correspond the number of cells from tripicates, which were validated by data sets from other researchers.

In this review paper, we replicated the statistical experiments conducted by these researchers to have a better understanding of their philosophy to determine data frauds. In addition, the underlying assumptions were checked by us to show if that subjectively introduced assumptions's effects on the outcome of the test statistics. A cross-validation process was also applied to check the difference between the grouped data and individual data sets. Hypothesis testings were also applied by using different models to show the effectiveness of the models that were used by researchers to detect data fraud. By way of doing this, we are able to test the effectiveness of the statistical experiments provided by these researchers, and also provide additional reasonable assumptions, which is much closer to real life situations.

2 Reproducibility of Results

In this section, we replicated the statistical experiments that were conducted by the researchers. It follows an order of: 3.1 Mid-Ratio Analysis; 3.2 Hypothesis Testing I - a nonparametric test; 3.3 Hypothesis testing II - using lambda to obtain p-values; 3.4 Hypothesis testing III - normal estimation of p-values; 3.5 Terminal digit analysis; 3.6 Equal digit analysis.

2.1 Mid-Ratio Analysis

2.2 Hypothesis Testing I - a nonparametric test

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 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.

A detailed calculation is in the Appendix.

2.3 Hypothesis testing II - 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.

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:

2.4 Hypothesis testing III - normal estimation of p-values

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:

2.5 Terminal digit analysis

2.6 Equal digit Analysis

3 Our Analysis

In this section, after we replicated the statistical experiments from the researchers, we will double check their underlying assumptions, such as poission model, normal distribution, and so on. Besides, we also used cross-validation to show if the combined data mid-ratio would give us a different results if we take each data set individually and calculate corresponding mid-ratios.

3.1 Permutation Tests

3.2 Discussions

3.3 Results interpolation

4 Conclusion

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...