Referee Report on Pitt and Hill (2016): Statistical Analysis of Numerical Preclinical Radiobiological Data

Madeleine Sheehan
m.sheehan@berkeley.edu

Kenneth Hung
kenhung@berkeley.edu

Yiyi Chen
yiyi.chen@berkeley.edu

Yulin Liu
liuyulin101@berkeley.edu

September 26, 2016

1 Introduction

In Pitt and Hill (2016), the authors examine the radiobiological data sets from 10 individuals in the same laboratory, along with data from three outside laboratories applying similar methods. As the data from one of the 10 individuals appears anomalous, the authors employ statistical techniques to determine if the anomaly could have happened at random.

In their analysis, the authors conduct hypothesis testing on four main metrics: triplicate mean count, mid-ratio count, terminal digit distribution, and equal last two digits distribution. For each metric, they conclude that the pattern seen in the one individuals data is too abnormal to have occurred by chance.

As part of the referee process, we carry out three main activities in reviewing the paper. We first examine and verify the assumptions the authors make on the distributions of the mid-ratio and terminal digits of triplicate counts (modeled as Poisson Binomial variables). We then attempt to replicate the tests conducted in the study, and compare our results with those in the paper. Finally, we apply an alternative statistical testing method for cross validation.

Overall we agree with the authors' analyses and conclusion, that it is unlikely the anomaly in the data from the one individual could have happened by chance. We see similar results from the replicated tests. For the alternative method, we perform permutation testing on the pooled data sets and reach a similar conclusion.

It should be noted, however, that we do see a few minor discrepancies between our results and those in the paper. Additionally, there are remaining open questions about the assumptions made in the paper, which we believe would strengthen the analysis and conclusion if they are addressed.

2 Poisson Assumption

As the authors explain — the random variables reported here correspond to the final counts of cells in three Petri dishes. There are initially some cells in each dish — the authors report that the initial counts can be modeled as a Poisson distribution with an unknown parameter λ_0. Each of the three dishes is subject to the same treatment (radiation level) and the probability that a given cell survives to generate a colony is p. The resulting final counts make up the radiobiological data sets analyzed in the paper. The authors claim that the final counts of the cells in a triplicate follow a Poisson distribution with $\lambda = \lambda_0 p$.

As we do not have a radiobiology background, we do not have the domain expertise to validate the assumption that the triples actually come from a Poisson distribution with a common parameter λ. Three data points drawn from the distribution do not give us enough information to verify this assumption either.

This assumption also comes with implications. The common cell survival rate p implies the survival of an individual cell to start a colony is independent of initial number of cells in the dish and the all other cells present.

On the other hand, the assumption that the initial counts follow a Poisson distribution seems to lack justification. The initial counts for the three Petri dishes are intended to be the same. If they are modeled as random, then the final counts can no longer be assumed as independent i.i.d. random variables. In fact, if the initial counts n is fixed and sufficiently large, and we just assume that the survival probability p is sufficiently low, the final counts would have followed a Poisson distribution with parameters np anyways.

Without additional domain knowledge in radiobiology, we will henceforth restrain from commenting on the setup of the model. In the sections that follow, we will generally trust that the triplicate counts come from a Poisson distribution. However, we will also provide alternative analysis through simulations and permutation tests, bypassing the assumption of a Poisson distribution and comparing the results from the RTS investigator against his / her peers. For example, we will perform a permutation test / hypergeometric test to see if the number of mean containing triplicates for the RTS investigator is significantly large in comparison to that of the other researchers.

3 Triplicate Analysis

The authors first come up with a model to simulate the natural process of producing triplicate data for colony. They assume that each of the three values are generated as three i.i.d. Poisson variables sharing a common parameter λ, which never exceeds 1000. The authors propose a method of calculating the probability that a triplicate generated by such process includes their own rounded mean. The derivation of the formula (Appendix A) seems correct.

We implement the approach proposed by the authors, and regenerate Table 1 (we generate the probabilities for λ ranging from 0 to 1999) in the paper. The results are consistent with the authors' (except for $\lambda = 13$, but we believe it is a typo from the authors). To further verify

the results, we use a simulation-based approach to recreate Table 1, and the results are very similar with the authors for large numbers of trials.

3.1 Mean Containing Triplicate Analysis

Replication The authors claim that they detect an unusually large number of triples that include their own rounded mean. To determine if the high number of rounded mean containing triples may have occurred by chance, the authors first construct a nonparametric test to get a upper bound of p-value to reject the null.

First, a note of clarification — the authors never explicitly define the term "complete triple". Table 2 of the paper identifies that the RTS colony dataset has 1361 total triples, and 1343 complete triples. From our analysis, we recovered their definition of a complete triple to be one that has a gap ≥ 2 — there are 18 triples with gap < 2 in the data set. For Hypothesis Test I, the authors test the hypothesis that 690 of 1343 triples contain their mean. They are therefore omitting all gap < 2 triples from the test. The rationale behind this omission is not entirely clear. For the purpose of replicating the test, we dutifully exclude these triplicates in our calculation as well.

Our replication of Hypothesis Test I is consistent with the authors'. As pointed out in the paper, the method is intentionally conservative (it overestimates the p-value). The authors thus propose a heuristic method to get a more sensitive and accurate estimates of p values. They assume that the event of each triplicate containing its own rounded mean is a Bernoulli trial with a known probability of success, and such probability that k of n triplicates in a data set contain their mean is assumed to follow a Poisson-binomial distribution. While they do not know the parameter λ for the distribution, they use the mean of each triplicate as the λ. While this is the MLE of λ, assuming the actual λ for the triplicate is known contradicts the frequentist assumption under which hypothesis testing is done.

Nonetheless, we are able to use the `poibin` package in R to replicate the tests that produced Table 2 of the paper. In the body text of the paper, the authors suggest that they apply the test to the 1343 complete RTS samples. We believe that might be a typo in the text — our replication matches the authors results only if we include all 1361 samples. If we replicate Tables 2 of Pitt and Hill (2016), but use only the 1343 complete (gap ≥ 2) samples, we get the results shown in Table 1, below. Discrepancies between this result and Table 2 of the paper are in bold. We will not replicate the Coulter count mean-containing triples analysis in this section.

Type	Investigator	# exps	# comp. / total	# mean	# expected	StDev	Z	$p \geq k$
Colonies	RTS	128	1343/1361	690	**214.9**	**13.28**	**35.73**	**3.66e−15**
Colonies	RTS	59	**578/597**	109	**103.4**	**9.06**	**0.56**	**0.284**
Colonies	RTS	1	49/50	3	**7.8**	**2.55**	**−2.07**	**0.989**

Table 1: Replication of colony count from Table 2 of Pitt and Hill (2016) using only complete (gap ≥ 2) triples. Discrepancies between this and Table 2 of Pitt and Hill (2016) are in bold.

To sum up, both hypothesis testing methods are reasonable. However, Hypothesis Testing I gives us a conservative estimation of p values, and it rejects the null hypothesis that the high

number of rounded mean containing triples could have occurred by chance. We think it might not be necessary to do another experiment using Hypothesis Testing II, especially when the method has a shaky assumption of parameters.

Permutation test (hypergeometric) As an alternative test, we circumvent the assumption that the samples are generated from a Poisson process, by conducting permutation tests on the same Coulter and colony data sets to verify the conclusion from the earlier section. For this, we first pool together all the sample data for Coulter and colony counts respectively. We then draw random samples from the pool, where the size of the sample is equal to that of the RTS investigator. Two methods are employed.

Method 1. We first run a simulation of $10,000$ draws. In those $10,000$ draws, we count the number (and proportion) of draws containing equal or more mean-containing triplicates than those observed in RTS investigator data set. After running the simulation a few times, we notice that the number (and proportion) of draws containing more mean-containing triplicate is consistently 0. This preliminary test suggests that it is highly unlikely that the sample in the RTS investigator data set would have occurred by chance.

Method 2. To get a more precise bound, we proceed to calculate the probability analytically using a hypergeometric distribution. In modeling the distribution, we define the drawing of a mean-containing triplicate a success event. The population N is the total number of triplicates in the pooled samples (for Coulter and colony counts respectively), K is equal to the total number of mean-containing triplicates, and n is set to the sample size of the RTS investigator data set. We then calculate the probability of k successes in n draws, where k is equal to the count of mean-containing triplicates in the RTS investigator's data set.

Type	**Sample size (n)**	**Test statistic (k)**	**Probability**
Coulter (all triplicates)	**1727**	177	1.33e−13
Coulter (excluding consecutive triplicates)	**1726**	**176**	1.84e−13
Colonies (all triplicates)	1361	708	2.40e−43
Colonies (excluding consecutive triplicates)	1343	690	2.96e−48

Table 2: Hypergeometric Distribution Probability for Mean-Containing Triplicate Sampling. Discrepancies from Table 2 in Pitt and Hill (2016) are bolded.

The exact probabilities of obtaining the same number of mean-containing triplicates as in the RTS investigator's dataset is exceedingly small ($< 10^{-10}$) for both Coulter and colony data sets, thereby supporting our observations in *Method 1* and the earlier conclusion.

In their analysis, the authors exclude triplicates with adjacent counts, where the maximum and minimum of the triplicate count differ by at most one. Since the rationale behind this treatment is not entirely clear, we carry out permutation test on both the full pooled samples and the samples excluding triplicates with adjacent counts. The results are only marginally different. We therefore maintain the original conclusion.

3.2 Mid-Ratio Analysis

Similarly, the authors suggest that the RTS investigator observes a surprisingly high percentage of triples that contain a value close to their mean. A triple is said to contain a value close to its mean if the triples mid-ratio falls in the interval $[0.4, 0.6]$, where the mid-ratio is defined as the ratio of the difference between the mid and the smallest value in the triple to the difference between the largest and smallest value in the triple.

By simulation, we corroborate the authors' finding that, for triples generated from a Poisson distribution with parameter λ from 1 to 2000, the expected percentage of triples with mid-ratio in the interval $[0.4, 0.6]$ never exceeds 0.26. The simulated results are shown in Figure 1

As stated in the paper, for a collection of n triples, the probability of observing k or more triples with mid-ratios in the interval $[0.4, 0.6]$ cannot be greater than the probability of k successes in n Bernoulli trials with the probability of success, p.

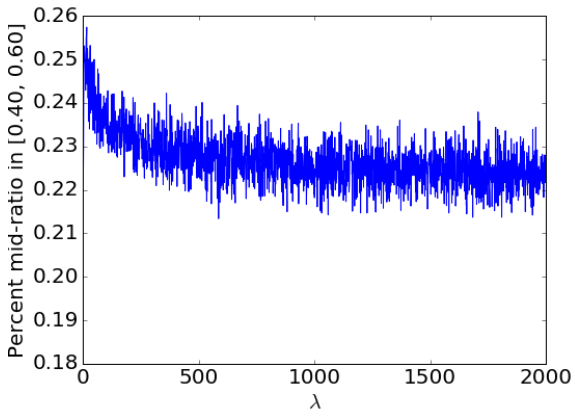

Figure 1: Simulated mid-ratio test

Carrying out this test we find that 824 of 1362 colony counts and 523 of 1729 Coulter counts produced by the RTS investigator have a mid-ratio value in the interval $[0.4, 0.6]$. If we model each triple as a Bernoulli trial with probability of success, $p = 0.26$, then the probability of observing 824 or more successes in 1362 colony count trials is 1.11e−16. The probability of observing 523 or more successes in 1729 Coulter count triplicates is 3.26e−5. These p-values are not reported in the paper. Both corroborate the finding that it is very unlikely that a Poisson process produced this many triplicates with mid-ratios in the interval $[0.4, 0.6]$.

While these results seem far too unlikely to have happened by chance, we do take issue with the fact that Pitt and Hill decided to perform this test after observing "what appeared to be an unusual frequency of triples in RTS data containing a value close to their mean". Other investigators besides RTS may well have trends that make their data look anomalous under these *post hoc* analyses. The argument would be stronger if the authors had decided what tests to perform *a priori*, before looking at the data and observing an "unusual" trend; or if they have adjusted for their *post hoc* analysis explicitly. We believe such selection bias is not as problematic for terminal digit analysis, as it seems to be a more routine tool for investigation.

Data set	0	1	2	3	4	5	6	7	8	9	Total	χ^2	p
RTS Coulter	**475**	**613**	**736**	**416**	**335**	**732**	**363**	**425**	**372**	**718**	**5185**	**466.9**	7.06e−95
Other Coulter	261	311	295	259	318	290	298	283	331	296	2942	16.0	6.70e−02
Outside Coulter 1	28	34	29	**25**	27	36	44	33	26	33	**315**	**9.476**	**3.95e−01**
Outside Coulter 2	34	38	45	35	32	42	31	35	35	33	360	4.9	8.39e−01
RTS Colony	564	324	463	313	290	478	336	408	383	526	**4085**	200.7	2.33e−38
Other Colony	**191**	**181**	**195**	**179**	**184**	**175**	**178**	**185**	**185**	**181**	**1834**	**1.79**	**9.94e−01**
Outside Colony	21	9	15	16	19	19	9	19	11	12	150	12.1	2.06e−01

Table 3: Our replication of Table 3 from Pitt and Hill (2016). Discrepancies are bolded.

4 Terminal Digit Analysis

Assumption Check We start by validating the assumption that the distribution of terminal digits of a Poisson variable is approximately uniformly distributed. We approximate the probability distribution of the unit digit of a Poisson variable for λ from 50 to 500. Each probability distribution is compared to a uniform distribution and the total variation distance (a metric comparing the distance of distributions) is computed, and shown in Figure 2. The total variation distance appears relatively small, compared to 0.1, and hence affirming the assumption.

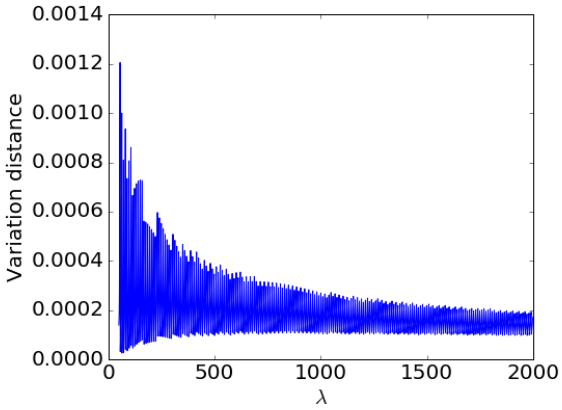

Figure 2: Terminal digit variation distance from uniform distribution

Replication Given that the terminal digits of a Poisson distribution is approximately uniform, we replicate the chi-square goodness of fit test to assess the significance of non-uniformity in each of the data sets. Our conclusions match the findings of the paper — we reject the null hypothesis of uniformity for the RTS data sets, and fail to reject the null for all other data sets. While our conclusions are the same, we did find some reporting issues in Table 3 of the paper. Our reproduction of Table 3 is shown in Table 3. The discrepancies between our terminal digit counts / goodness of fit calculations and the authors are shown in bold.

Permutation Test To bypass the assumption that the triples can be modeled as Poisson random variables, we pool all the data from all investigators and ask "how unlikely is it that

the RTS colony count terminal digits were as non-uniform as they were if they were drawn from the same distribution as all the investigators?" The procedure for performing a permutation test is outlined as follows:

1. Compute chi-squared goodness of fit test statistic for sample of interest
2. Pool all the data
3. Repeat the following for sufficiently many times
   (a) Randomly select a sample from the pooled data that is the same size as the test sample
   (b) Compute the chi-squared goodness of fit test statistic for this new sample
4. Find the percentage of random samples where the random sample test statistic is larger than the original test statistic

The returned percentage approximates how likely it is that the RTS colony counts deviate more from uniformity if the counts are drawn from the same distribution as the pooled data from all the investigators. Due to limits in computing power, the number of random samples generated is limited and so is our accuracy in approximation. Nonetheless, the returned percentage has been returned as 0 consistently, indicating that the deviation of RTS colony counts from uniformity unlikely occurs by chance, if it was drawn randomly from the pooled data.

5 Equal Digit Analysis

Assumption Check We consider the assumption, that the last two digits of three-plus digit Poisson variables being equal at probability 10%, rather imprudent.

We approximated the probability of the last two digits of a Poisson variable being equal, conditioned on the outcome bearing three digits. We plotted this in Figure 3. While in the larger regime, where $\lambda > 500$, the probability does hover around 10%; the same conclusion fails to be drawn about the smaller regime.

For the equal digit analysis proposed by Pitt and Hill (2016) to hold, one would at least hope for this probability to be significantly different from 12.3%. However, when λ gets comparable to 100 or even smaller, this probability can rise as high as 12.5%. (See Figure 3) Intuitively, if λ is significantly smaller than 100, then getting 100 as an outcome is much likelier than any outcome greater than 100, driving the probability of having the last two digits being equal much higher. Without the knowledge of the true λ and with quite a few counts falling under 100, we cannot rule out this regime and perform a proper statistical test.

6 Conclusion

For the review process, we examine the underlying assumptions put forth by the authors. We attempt to replicate all four analyses (mid-ratio and mean counts of triplicates, percentage of terminal digits and equal last two digits) outlined in the paper. In addition, we apply an

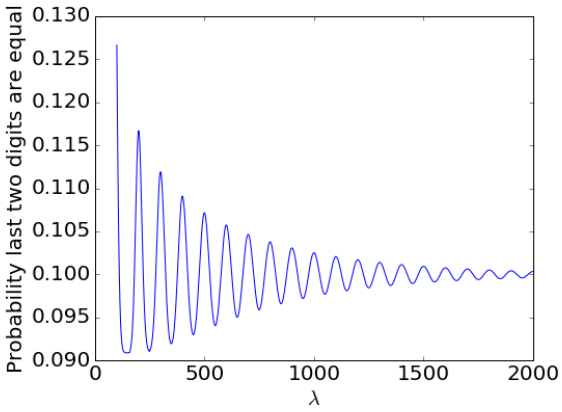

Figure 3: Probability that two terminal digits of Poisson random variable with parameter $\lambda > 100$ has equal terminal digits.

alternative testing method that does not make assumptions on the underlying distribution of the data set, in order to validate our earlier results.

Overall, our results agreed with those in the paper. Out of all the metrics tested, we feel that the test on percentage of terminal digits provides the strongest argument against the RTS investigator result. For the rest of the test metrics, the mid-ratio and mean counts rely on the strong assumption that each triplicate shares a common parameter λ, which we feel was inadequately justified. (The alternative testing method permutation test relax this assumption) The expected percentage of equal last two digits of a number produced by a Poisson process, on the other hand, has a high level of inherent variability, especially if the Poisson parameter, λ that produced the data is small. Since the counts in the observed colony data are generally small (< 200), the percentage of equal last two digits fails to suggest any significance in the result with high level of confidence.

In addition, we want to point out that the study can be strengthened with additional information on the authors' decision to look into the RTS investigators data in the first place, their rationale behind their choice of testing metrics, and the justification for the assumption on the data distribution.

Another area we did not investigate that will potentially provide additional insight is the patterns in the data of other individuals. It is possible that when examined individually, other investigators will also look anomalous for the same or other metrics. When the data are pooled together for the tests in the study, however, the individual anomalies might cancel each other out and get masked. It will therefore be useful for future studies / reviews to conduct further investigation in this area.

Reproducibility

The source code for our analysis can be found below:
`https://github.com/mads14/S215a-groupwork/`

The original full datasets for the study can be found below:
`https://osf.io/mdyw2/files/`.

Acknowledgement and Declaration

We would like to thank the authors for making their data available and for publishing in an open journal.

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.

References

Helene Z Pitt and Joel H Hill. Statistical Analysis of Numerical Preclinical Radiobiological Data. *ScienceOpen Research*, 2016.