**Review: Statistical Analysis of Numerical Preclinical Radiobiological Data**

Erik Bertelli, Stephanie DeGraaf, James Hicks

**Introduction**

This paper tackles the serious problem of detecting fraud by applying and developing multiple methods: digit analysis, mid-ratio review, and a triplicate probability model. In this review, we first attempt to replicate Pitt and Hill's results using the provided data, and second, offer several points of clarification and discussion about their methodological contribution.

We would like to thank the authors Pitt and Hill, for publishing their results in an open journal and making their data available online, enabling us to attempt this replication. 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.

**Summary of Main Findings**

The goal of Pitt and Hill’s work is to test the hypothesis that the data from a particular researcher (the RTS) were generated in a way so unusual as to suggest fabrication. Their first method of analysis, digit analysis, is a common method of detecting fraud in many kinds of datasets. The analysis of midratios is much more specific to the data, since the midratio is a characteristic of data with triples where the mean is of interest. Finally, the triplicate probability model makes up the bulk of Pitt and Hill’s work. This model, which incorporates particular characteristics of the data, is also designed specifically for a hypothesis which tests biological triples. All three analyses were carried out on two separate data sets: colony counts and Coulter counts, and unusual patterns in the RTS data were detected with very strong evidence.

After replicating Pitt and Hill’s major findings, we agree with the conclusion that the RTS data were not generated in the same fashion as the other investigators. While we offer small critiques of each method, the general strength of the evidence is such that we agree with the conclusions in the paper.

**Reproduction of Results**

First, we replicate the results of Pitt and Hill’s major analyses. Our code and results are available at <https://github.com/sldegraaf/preclinical-data-review>.

**Terminal Digit Analysis**

We independently replicated the terminal digit analysis for all of the data sets and found almost the exact same results. The only major differences were for the Coulter-RTS and Colonies-Others data sets, in which our counts were a bit higher than those reported in the paper. We applied the exact same code to all datasets, and the majority of them matched the results perfectly. The fact that our Coulter-RTS and Colonies-Others counts were slightly higher suggests some of the data may have been filtered out of the original author’s analysis, but we were unable to replicate that process. However, these small differences did not meaningfully change the Chi-squared test statistics or implication of the results.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Digits | | | | | | | | | |  |  |  |
| Type | Investigator | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Total | Χ2 | P |
| Coulter | RTS | 475 | 613 | 736 | 416 | 335 | 732 | 363 | 425 | 372 | 718 | 5185 | 466.87 | 7.06E-95 |
| Coulter | Others | 261 | 311 | 295 | 259 | 318 | 290 | 298 | 283 | 331 | 296 | 2942 | 15.99 | 0.07 |
| Coulter | Outside 1 | 28 | 34 | 29 | 25 | 27 | 36 | 44 | 33 | 26 | 33 | 315 | 9.48 | 0.39 |
| Coulter | Outside 2 | 34 | 38 | 45 | 35 | 32 | 42 | 31 | 35 | 35 | 33 | 360 | 4.94 | 0.84 |
| Colonies | RTS | 564 | 324 | 463 | 313 | 290 | 478 | 336 | 408 | 383 | 526 | 4085 | 200.73 | 2.33E-38 |
| Colonies | Others | 191 | 181 | 195 | 179 | 184 | 175 | 178 | 185 | 185 | 181 | 1834 | 1.79 | 0.99 |
| Colonies | Outside 3 | 21 | 9 | 15 | 16 | 19 | 19 | 9 | 19 | 11 | 12 | 150 | 12.13 | 0.21 |

We also replicated the analysis of the final two digits in each of the counts and also found that the RTS data was suspiciously non-uniform. In the paper, the authors test the occurrence of equal digits in the last two digits for the Coulter RTS and Other data sets, restricted to values with at least 3 digits. We replicated this analysis and found similar results, with RTS having 12.4% of values with equal last digits and Others having 9.9% equal. Using a binomial distribution with probability .1 (meaning an assumption of uniform distribution) these outcomes have probabilities of 7.8 E -09 and .55 respectively, which is strong evidence to be suspicious of the RTS data set.

**Midratio Analysis**

To replicate the results of mid-ratio analysis, we calculated the mid-ratios for Colony counts in the RTS data and Other Investigators’ data. As in Pitt and Hill's initial findings, our results indicate an overwhelming predominance of midratio values in the 0.4-0.6 range in the RTS data, compared to a uniform distribution in the Others' data.

To ensure that the uniform distribution is what we should expect from genuine data, we also looked at the distribution of the midratio on the Outside Lab's data, which Pitt and Hill did not include in their results. While there were fewer triples to consider, the distribution of the midratio still appeared uniform, supporting Pitt and Hill's conclusion that the RTS data is unusual. In addition, we checked the uniformity of the midratio distributions in the Coulter counts. Pitt and Hill also did not include this analysis, and when we analyzed the Coulter counts, the mid-ratios did not reveal anything unusual.

Pitt and Hill then constructed a probability model for the mid-ratios based on the probability of a midratio falling into the interval [0.4,0.6], using the given Poisson parameters. The test of significance relies on the fact that for any value of lambda, the probability of falling into the ratio is less than 0.26. We applied this model to the Colony counts and conducted a binomial-based significance test. Out of 1361 triples, 824 of the midratio values fell into the range [0.4, 0.6], yielding a highly significant p-value near zero.

**Triplicate Probability Models**

To check the RTS findings against a theoretical distribution, the authors assert that each triple t can be modeled as a set of IID draws from a Poisson distribution with common parameter lambda\_t. The comparator distribution permits us to answer the question: what is the probability that from n triples, k will include their means as one of their values? Pitt and Hill derive a probability model from the properties of the Poisson distribution, which we did not check in detail. Instead, we incorporated their assumptions into a simulation, in which we drew 10,000 triples for each value of lambda from 1 to 2,000, and then estimated the rate of mean-inclusion for each lambda. This yielded a set of simulated probabilities that corresponded closely to Pitt and Hill’s “MidProb” table.

Using these probabilities, Pitt and Hill conduct three tests of significance. The first “crude” test simply finds the maximum probability (which happens to be where lambda = 4), uses that to calculate the binomial probability of n triples in which k contain their means. This essentially approximates an “upper bound” for the p-value. This is a straightforward calculation, and we were able to replicate the authors’ results.[[1]](#footnote-0)

The second, more refined, test relaxes the requirement of constant lambda. Instead, it treats the RTS data as a series of Bernoulli trials (where “success” indicates that a triple includes its mean). Each trial has a distinct probability of success derived from our “MidProb” table, with the lambda parameter set as each observed triple’s mean (lambda is the Poisson maximum likelihood estimator). This yields a Poisson binomial distribution (the distribution of the sum of independent Bernoulli trials with non-equal probability), whose density can be readily estimated. We did not precisely replicate the authors’ results here, because our simulated probabilities were slightly different to those in the paper, but we confirm their overall conclusion with respect to the RTS data. We are slightly less confident with respect to the data for other investigators, but again our conclusions are not materially different. As we report in the terminal digits analysis, our number of cases was slightly higher (k = 128; n= 597), leading to a p-value of 0.17 — some distance from their 0.58, but not a statistic which renders the data improbable.

Finally, the authors use the normal approximation of the Poisson binomial to calculate traditional z-scores. Here, our results closely mirrored the authors’; a z-score of around 34 is strong evidence against the null hypothesis.

**Criticisms of Analysis**

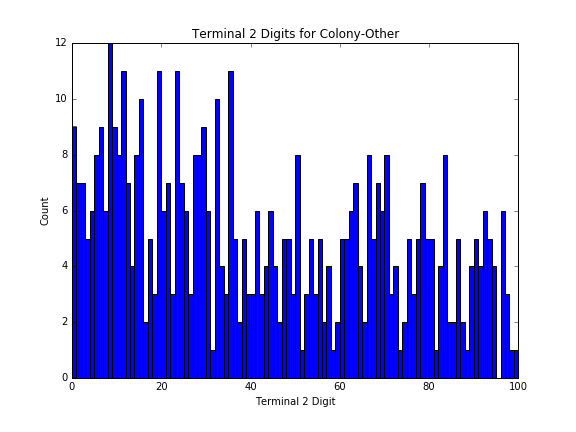
In a general sense, it is important that hypotheses used to detect fraud be formulated and specified *prior* to conducting any exploration of the data. From a statistical point of view, it is always possible, genuine randomness notwithstanding, to find *some* unusual pattern in a sample of data if enough features are analyzed. However, these unusual patterns do not reflect any true departure from randomness and should not be used as evidence that data are not genuine. That is, in statistical terms, it is still possible to obtain significant p-values for unusual characteristics of the data, but these p-values no longer have the same interpretation if hypotheses were not pre-specified.

To demonstrate how easy it is to find “unusual” features present in data, consider the data for the Other Investigators' Colony counts, which Pitt and Hill deem to be genuine. In the triplicates, it appears that the lowest value of the triple appears in the first column of the data quite often: in 230 of the 597 cases the lowest value is in the first column, a proportion of 0.385. In a “truly random” scenario, we would expect 1/3 of the lowest values to be in column 1. Thus, if we treat each triple as an independent random trial with probability 1/3, we can conduct a binomial test of significance on the 597 trials. Our sample estimate returns a p-value of 0.00437, indicating that the probability of seeing the smallest value in the first column is significantly higher than it would be due to chance alone. This is clearly an incorrect interpretation of the p-value, however, since we did not hypothesize that low values would be in first column until *after* observing this phenomenon.

This simple example serves to emphasize the importance of selecting metrics of fraud before exploring the data. The terminal digit analysis is a standard metric and it is logical to assume that this hypothesis would be specified before looking at any data. However, the midratio analysis seems to be a characteristic of the data observed after looking at the data. Pitt and Hill even acknowledge that their decision to test mid-ratios came *after* "[h]aving observed what appeared to us to be an unusual frequency of triples in RTS data containing a value close to their mean". This casts some doubt on the validity of the p-value; however, the significance of the midratio test was so strong that we feel confident that the conclusion is valid.

**Terminal Digit Analysis**

The terminal digit analysis is a good way to test the validity of the data, and does not rely on too many unrealistic assumptions about the way the data was generated. It also has the benefit of being a simple to understand and common approach one would try early on when addressing this question. On the other hand, the equal digit analysis is somewhat more aggressive than the terminal digit analysis. Since the majority of these numbers are only three-digit, it is much more likely that the final two digits will not be uniformly distributed in these data sets compared to just the terminal digit. In fact, we believe that is why the equal digit analysis was only applied to the Coulter data sets, as the Colony data sets were not of large enough average values to lead to guarantees of uniformity. For example, if you examine the distribution of all possible final two digit pairing in the Colony-Other Investigators data set, you can see that the empirical distribution does not look very close to uniform:



As you can see, the distribution is quite a bit heavier for the smaller values of the last two terminal digits. For this reason assuming that the last 2 digits of each number should be equal 10% of the time would not be valid, and it is good that the authors did not apply the equal digit analysis to the Colony data sets.

It is also worth wondering why a more general test of the uniformity of the final two digits such as a chi-squared test on all 100 possible two digit combinations was not used. We performed such a test and found that the probability for the Coulter-Other data set was only 0.06. This results leads us to question the Coulter Equal Digit probabilities given in the paper, as the 10% assumption for the equal digits relies on the overall distribution of last two digits being uniform. If the authors noticed that the RTS data set had a large number of data points with the last two digits being equal and *only then* applied the statistical test, that would reduce the confidence of that outcome in our minds. In our opinion the Terminal Digit analysis already shows that the data is not distributed as expected, and the Equal Digit Analysis only complicates this analysis with less persuasive evidence.

**Triplicate Probability Models**

In general, we agree with the logical formulation of the model in this section, but we are curious about the choice of distribution to model the count data. In general, the Poisson distribution is a natural choice to model count data because it is a generalization of the Binomial, as the Binomial parameter *n* goes to infinity and *p* goes to zero. However, in this scenario, it is unclear whether the number of colonies formed by surviving cells is an appropriate application of the limits of a Binomial.

In addition, we have concerns about the parameter estimation in each triple. The sample mean for each triple is indeed the maximum likelihood estimator of the Poisson lambda parameter, but with only three data points available to estimate this parameter, these sample means are unlikely to be reliable estimates of the true parameters. It might be better to incorporate shrinkage methods or use a Bayesian approach to share information across all colony triples to get more reliable estimates of each triple’s mean.

Since the strength of evidence in this scenario is so strong, we do not think these criticisms would change the conclusions, but they would be quite salient in scenarios where the evidence were not as strong.

**Location of the Mean in the Triples**

As an extension of this analysis, we noticed in the image of the RTS lab notebook provided that the mean value always appeared in the first column in the mean containing triples. However, this is only a sample of 6 triples, so we applied a chi-squared test to the count of the means occurring in each column, on the assumption that those count should be uniform. We excluded the outside lab results due to low counts and obtained the following results:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type | Investigator | Column 1 | Column 2 | Column 3 | Χ2 | P |
| Colonies | RTS | 220 | 373 | 97 | 166.25 | 7.92E-37 |
| Colonies | Others | 38 | 45 | 26 | 5.08 | 0.0788 |
| Coulter | RTS | 38 | 79 | 59 | 14.33 | 0.0008 |
| Coulter | Others | 13 | 13 | 10 | 0.5 | 0.7788 |

Once again we find that the RTS datasets do not conform to the uniformity assumptions while the other investigators have data which is consistent with the assumption of means uniformly distributed across the columns. We believe that this information in conjunction with the other evidence presented further enhances the case that the RTS data is not genuine.

**Conclusion**

While we have raise several technical points of consideration about Pitt and Hill's methodology, the evidence in general is strong enough for us to arrive at the same conclusion — the RTS data is likely anomalous.

In general, we found Pitt and Hill’s work to be both convincing and valuable for future work in detecting fraud. Crucially, their paper illustrates the importance of developing statistical methods to detect fraud in data that would otherwise appear genuine, and contributes valuable tools for detecting fraud in similar preclinical radiobiological data sets. While the triplicate probability model was designed specifically for this type of data, Pitt and Hill’s ideas can be applied more generally; this model provides a general framework for future methods to be developed for other types of datasets.

1. Our simulated probabilities also incorporated Pitt and Hill’s restriction that the range of any triple had to be at least 2. This appears to follow from their assumption that fabrication involved first selecting a desired mean, and then choosing two other numbers such that they were approximately equal distances from that mean. In general, it is not clear this must be true: the triple (0,1,1) contains its rounded mean (1), even though it has range < 2. This has a material effect on the estimated (or calculated) probabilities for low values of lambda. Indeed, since the maximum probability without the range limitation is close to 0.8, the RTS data would actually “pass” Pitt and Hill’s first (crude) test. [↑](#footnote-ref-0)