Ethics of Data Quality

ASA Ethics Case Study #3

Report Author:

Tom Hood

Date: October 26, 2018

Case Study Analysis

The ASA Ethical Guidelines for Statistical Practice assert the integrity of data and methods is of utmost importance. In this case study, we see a statistician troubled by a large company, serving both government and private clients with large amounts of data unavailable anywhere else, with no control processes or audit procedures in place to assure a uniform quality of data. In the study, the statistician claims “the data are probably ‘pretty good’ but are likely to vary widely in quality from one data set to another.” Being an ethical statistician, she proposes to management a series of services to institute data quality standards, procedures, and improve availability of analytic products using the data; despite her proposals receiving applause from management, it is continuously left unfunded. When discussing this with her colleagues, they mention many widely used data sets lack formal quality standards, and additionally even with quality control standards in place it can take years or even decades to identify and resolve the underlying data quality issues. The statistician is deeply troubled by this and is considering either adding disclaimers to each data product to inform customers about the lack of quality data control or taking other employment available to her.

Examining this case study, the most glaring issues pertaining to statistical ethics are the lack of data and method integrity and neglect of responsibility to clients. According to the ASA ethical guidelines, “the ethical statistician is candid about any known or suspected limitations defects, or biases in the data that may affect the integrity or reliability of the statistical analysis” (Page 2, Section B). In this case, the company knowingly submits invalid data with known inconsistencies, faulty assumptions, and lack of auditing, and does not report the known data limitations or provide any form of disclaimers to their clients. This behavior is not only unethical

but highly irresponsible and reckless. The case study notes that the data is shared with various government agencies, meaning that in some uses the data has severe or significant social impact and importance. Willfully neglecting their duties, the company also breaches their owed responsibilities to their clients and the public, most notably failing in their "[understanding] and [conformity] to confidentiality requirements of data collection, release, and dissemination and any restrictions on its use established by the data provider (to the extent legally required), protecting use and disclosure of data accordingly" (Page 4, Section C). Finally, while not as eminent an issue as the former two, the company has also neglected responsibility towards their statistician employees by forcing them into an unethical working situation. Employers, organizations, and anyone who otherwise employs someone to analyze data have an obligation to "understand and respect statisticians' obligation of objectivity…, recognize that the ethical guidelines exist and were instituted for the protection and support of the statistician and the consumer alike…, [and] support sound statistical analysis and expose incompetent or corrupt statistical practice" (Page 6-7, Section H).

Addressing the present issues within the company prompts us to wonder what the statistician should do in her given situation and offer any potentially solutions that might be implemented to correct the otherwise unethical behavior. First to advice the statistician, she has a responsibility to provide uncompromised and consistent data and should pursue options to enable her to do so. The first, and perhaps best, solution would be involving the legal department and ASA foundation and informing both parties of the unethical behavior present within the company. As mentioned earlier in the report, those employing any person to analyze data are obligated to understand and respect the statistician's obligation of objectivity, and

“employers, funders, or those who commission statistical analysis have an obligation to rely on the expertise and judgement of qualified statisticians for any data analysis. This obligation may be especially relevant in analyses known or anticipated to have tangible physical, financial, or psychological effect” (Page 7, Section H). While inconsistent data standards and lack of auditing is never acceptable, it is especially irresponsible when the data has such large social magnitude. Ideally, intervention from legal professionals and the ASA would result in the company providing proper funding and support to the statistician’s proposed data integrity services and thereafter ensure that appropriate data ethics are regularly practiced. A second solution is to advise the statistician to follow through with her plan to include disclaimers on all data products to inform customers about the severe lack of quality control. As a statistician, she has a duty to “[report] on the validity of data used, acknowledging data editing procedures, including any imputation and missing data mechanisms” and therefore is obligated to provide notice of inconsistent data quality to all consumers (Page 2-3, Section B). However, this solution may also place the statistician susceptible to punishment from her employer if she include disclaimers without permission, consent, or knowledge of her superiors, and it could be argued that her behavior in this case is insubordinate and worthy of termination, defamation, or revoking of certifications; for these reasons, this solution should not be selected with precedence over the previous one unless the statistician can receive some form of permission or consent form her employer to provide disclaimers. Additionally, this solution does nothing to actually correct the lack of data quality control or standards, rather it just serves as a method of damage control to mitigate the company’s liability. Her last solution is to leave her job and take up employment somewhere else, but even this solution has flaws. Firstly, she is still not solving

the unethical behavior at her current employment, and as a statistician she has a responsibility to report or remediate any unethical behavior she is aware of. Secondly, this solution is contingent on her having a different employment opportunity available or else she risks becoming unemployed. For these reasons, this solution should only be considered as a last resort and is not recommended for the statistician. In summary, the best solution for the statistician is to report all unethical behavior existing at her employment to her company's legal department and the ASA foundation and demand corrective action and the institution of responsible data practices for the company.

Statistical Analysis and Simulation

Premise:

I have acquired two data sets, one with quality control standards and one without, to perform statistical analysis on and simulate the effects of lacking quality control and auditing procedures on data sets. The sets contain Amazon stock data over a 10-year range (2004-2014). The set that met quality control standards contains samples from every day during the 10-year span (a total of 2,519 data points) and has consistent decimal truncation (all samples rounded to 2 decimal places). The set that failed to meet quality control standards has regular gaps between samples (a total of 1,506 data points) and does not have consistent decimal truncation (samples are rounded to 0, 1, 2, 3, or 4 decimal places). All statistical calculations are performed in R to prevent human error. All relevant tests and datasets have been included in this report for reader convenience.

Purpose:

The main purpose for analysis was to compare two data sets (one with quality control standards and one without) to better understand the effects lack of quality control can have. Before testing, I theorized that the lack of quality control standards would have a significant effect on the sample, to the extent that the mean values of the two data sets would not be the same. In this case study, the data in concern is used by several private and government agencies and is stated to carry significant social impact. This implies that erroneous data would also have severe consequences, thus highlighting the importance of reliable and quality data.

Comparison of Mean Closing Prices:

First, I created data subsets for stock closing prices, with *qc_closing* originating from the dataset with quality control standards, and *noqc_closing* originating from the set without any quality control standards or auditing procedures. I then derived a five-number summary for both subsets, shown below:

```
> summary(qc_closing)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.07   51.42  120.00  148.64  227.30  407.05
> summary(noqc_closing)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.91   69.86  122.75  146.82  232.42  360.84
```

At a glance, we see that the data is not identical, however the variation between the two sets is appears rather insignificant, likely due to the sheer size of the samples. To investigate further, I conduct a T-Test on the closing price means.

$$H_0: \bar{x}(qc_{closing}) = \bar{x}(noqc_{closing})$$
$$H_1: \bar{x}(qc_{closing}) \neq \bar{x}(noqc_{closing})$$
$$\alpha = 0.05, variances\ not\ equal$$

```
        Two Sample t-test

data:  qc_closing and noqc_closing
t = 0.54909, df = 4021, p-value = 0.583
```

The test yielded a p-value of 0.583, leading me to accept the null hypothesis that the true difference in means is 0, implying the means are equal. At first, I found these results shocking because it suggests that the lack of quality control standards has no significant impact on data sets; however, at mentioned previously, the large sample sizes likely made up for the poor data quality. This could be verified later by conducting a similar test on smaller samples. Additionally, in the case study, the statistician's colleagues inform her that many widely used data sets also lack quality control standards. The results of the this test demonstrate that it is safe to use poor quality data as long as the sample size is large.

Comparison of Mean Daily Percent Return:

For the second test, I again created data subsets for stock returns, with *qc_return* originating from the dataset with quality control standards, and *noqc_return* originating from the set without any quality control standards or auditing procedures. Like above, I then derived a five-number summary for both subsets, shown below:

```
> summary(qc_return)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-21.8200  -1.1300   0.0000   0.1125   1.3400  26.9500
> summary(noqc_return)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-14.600  -1.090   0.000   0.186   1.400  26.950
```

As shown earlier, the data does not appear to vary significantly. It is extremely likely that the following T-Test will yield results similar to the previous test.

$$H_0: \bar{x}(qc_{return}) = \bar{x}(\text{noqc}_{\text{return}})$$
$$H_1: \bar{x}(qc_{return}) \neq \bar{x}(noqc_{return})$$
$$\alpha = 0.05, variances\ not\ equal$$

```
        Two Sample t-test

data:  qc_return and noqc_return
t = -0.82487, df = 4021, p-value = 0.4095
```

The test yielded a p-value of 0.4095, leading me to accept the null hypothesis that the true difference in means is 0, implying the means are equal. As already stated, the results of this test again demonstrate that it is safe to use poor quality data, assuming that it contains a large sample size.

Conclusion:

The above tests show that the use of poor qualitiy data for statistical purposes will not produce erronesous results and is insigificantly different from well-kept, standardized data. However, the main focus of this report is on the *ethics* of using poor quality data, so despite its neglible effects, it still remains impermissable to use poor quality data. The statistician has an obligation to maintain data in a standardized fashion. Also mentioned earlier was the test results being a product of the size of the samples; it is unknown, but highly probable that as the sample size shrinks or the magnitude of lacking data quality increases, the tests will indicate a statistically significant difference bewtween the sample means. Thus, to summarize, it remains unethical to practice statistics using poor quality data, and should not be done under any circumstances.

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data_noqc = read.table("C:/Users/trhoo/OneDrive/Desktop/AmazonStock_NoQC.csv", header=TRUE, sep
 =",")
> data_qc = read.table("C:/Users/trhoo/OneDrive/Desktop/AmazonStock_QC.csv", header=TRUE, sep =",
")
> attach(data_qc)
> qc_closing = Closing_Price
> qc_return = Daily_Percent_Return
> attach(data_noqc)
The following objects are masked from data_qc:

    Closing_Price, Daily_Percent_Return, Date

> noqc_closing = Closing_Price
> noqc_return = Daily_Percent_Return
> summary(qc_closing)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.07   51.42  120.00  148.64  227.30  407.05
> summary(noqc_closing)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.91   69.86  122.75  146.82  232.42  360.84
> summary(qc_return)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-21.8200  -1.1300   0.0000   0.1125   1.3400  26.9500
> summary(noqc_return)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-14.600  -1.090   0.000   0.186   1.400  26.950
> return_ttest = t.test(qc_return, noqc_return, var.equal = TRUE)
> return_ttest

        Two Sample t-test

data:  qc_return and noqc_return
t = -0.82487, df = 4021, p-value = 0.4095
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2483126  0.1012424
sample estimates:
mean of x mean of y
0.1124940 0.1860292

> closing_ttest = t.test(qc_closing, noqc_closing, var.equal = TRUE)
> closing_ttest

        Two Sample t-test

data:  qc_closing and noqc_closing
t = 0.54909, df = 4021, p-value = 0.583
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.658330  8.282668
sample estimates:
```

```
mean of x mean of y
 148.6372  146.8250

> tps = replicate(1000, return_ttest$p.value)
> plot(density(tps), main = "Probability Density for Percent Return")
> tps = replicate(1000, closing_ttest$p.value)
> plot(density(tps), main = "Probability Density for Closing Price")
>
```