Winter 2013, Volume 3, Issue 4

THE STATSWHISPERER

The StatsWhisperer Newsletter is published by Dr. William Bannon and the staff at StatsWhisperer.™ For other resources in learning statistics visit us on the web at: www.statswhisperer.com

The Most Important Questions Not Being Asked in Statistical Research

The most important questions *not being asked* in statistical research today relate to important *checks of data integrity*, which are smaller details within the larger process of data analysis. Checks of data integrity are small details in a historic list of secondary considerations that have dramatic consequences. For example, Civil War era surgeons discovered the small detail of hand washing prior to surgery cleansed germs toward dramatically enhancing patient outcomes. The detail of hand washing was small, but the effects were far reaching and impactful.

Analogous to the negative impact germs have upon outcomes in surgery, a quantitative study is often hindered by *insidious* problems within the data that *frequently result in incorrect statistical findings*. The identification and treatment of these insidious problems within the data is dependent upon making adequate checks of data integrity.

One could state that making checks of data integrity prior to statistical analysis is the equivalent of a surgeon hand washing prior to surgery. In each case, the potentially harmful factor (i.e., germs or insidious problems within the data) is being treated in preparation of the critical procedure (i.e., surgery or statistical analysis). Once more, just as positive outcomes in surgery are dependent upon thorough surgeon hand washing, producing accurate findings in statistical analysis is dependent upon making thorough checks of data integrity.

INSIDE THIS ISSUE

However, although checks of data integrity are essential, these details are largely ignored or not reported in the application of statistical procedures within the field of quantitative research. For example, within the professional peer-reviewed quantitative literature, one could extract a random number of published studies and find articles that pay little or no attention to making checks of data integrity. Thus, this newsletter issue has termed checks of data integrity to be the most important questions not being asked in statistical research.

Bannon (2013) recommends four primary areas in which checks of data integrity should be made, each of will be reviewed within the current newsletter issue. Furthermore, this newsletter issue will review the reasons that these checks are often omitted within quantitative research, as well as possible solutions to promote a wider application of these procedures.

What Are Checks Of Data Integrity?

Integrity is an important quality. People like to deal with other people who maintain their integrity. People like to drive on a road that also maintains its integrity. In fact, many of us make it a practice to avoid both people and roads that do not maintain their integrity. This is an intelligent choice, which should be extended into your approach to data analysis. Specifically, if you have not performed all necessary checks of data integrity, you should avoid using those data in inferential analysis.

Checks of data integrity are concerned with one central question, which is: Are the data appropriate for statistical analysis? In general, when performing any task, one needs the right materials at hand before the right product can be produced. Essentially, checks of data integrity substantiate a researcher is working with the right materials in a quantitative study, before beginning the process of data analysis. Certainly, we could never start with data that are not appropriate for analysis and expect to produce a quality study.

Currently among experts in quantitative analysis, there is not an *agreed upon* number of categories in which checks of data integrity should be made. However, Bannon (2013) suggests four primary areas where checks of data integrity are essential, which are:

1) Statistical Power

2) Test Assumptions

3) Missing Data

4) Measurement Tools

This newsletter issue will present each of these topics briefly, along with a description of how small factors within each area can easily distort statistical findings. The description of how each area can influence study findings is done to help illustrate why it is essential to question if each area has been addressed within a study. However, before addressing each area, we will review why these checks are not focused upon, as well as possible solutions to this challenge.

The Anonymity of Checks of Data Integrity

Although checks of data integrity are essential in quantitative analysis, in a great deal of published research these tasks reflect a certain level of anonymity. Specifically, within the literature, many researchers begin the report of study findings at the descriptive or inferential level of statistical analysis without describing if/how necessary checks of data integrity were addressed.

Essentially, a trend has developed where researchers focus upon statistical analysis, but not on data preparation. There are likely many reasons for this trend. For example, many individuals seem to have the impression that inferential statistical methods are so complex that all attention should be focused upon performing these techniques. Subsequently, performing the checks of data integrity prior to inferential analysis receives little attention. Although this trend may be understandable, it reflects an approach to quantitative research that is not fundamentally sound. To be blunt, this approach to statistical research is a method of producing research findings that are most likely inaccurate.

Another reason, this trend may exist is that there is not a formally agreed upon model of data analysis that instructs researchers universally on

The Anonymity of Checks of Data Integrity (continued)

the fundamental considerations that must be addressed when conducting a data analysis study. Essentially, there is not a standard model of data analysis that instructs statistical researchers on the need to perform checks of data integrity prior to conducting inferential analysis.

The absence of a standard model of data analysis reflects the fledgling nature of statistical research, as well developed fields have standard methods or models of operation. For example, the field of medicine has standard methods of operation that a surgeon conducting an appendectomy is expected to follow. Specifically, it would be inappropriate for a surgeon to deviate from the predetermined *model* of operation when performing an appendectomy that he or she is expected to follow as per the medical community.

Essentially, prior to performing the operation, the surgeon knows there are standard methods of operation that must be followed, as well as where to find those requisites listed (e.g., AMA website). If these requisites were not specified by a standard model of operation listed in an accessible location, the globe would be fraught with surgeons performing appendectomies in many diverse ways. In essence, without this standard model, surgeons would be performing these operations with varying levels of fidelity relative to how this procedure is meant to be performed. Thus, the presence of a recognized standard model of operation is a huge asset to patients, surgeons, and all related parties.

Similarly, the specification of a standard model of data analysis would provide similar benefits to those working in and served by the field of statistical research. Specifically, a standard model of data analysis would serve as a guide that specifies the agreed upon fundamental steps that a researcher must follow to maintain fidelity to statistical research. The absence of such a model is a primary cause of the currently large number researchers performing analysis in diverse ways with varying levels of fidelity to the statistical procedures used, including the presence or absence of checks of data integrity.

Bannon (2013) has suggested the following seven steps as a standard model that informs researchers of the fundamental steps needed to conduct a legitimate quantitative study:

Step 1: Study Map

Step 2: Data Entry

Step 3: Checks of Data Integrity

Step 4: Univariate Statistical Analysis

Step 5: Bivariate Statistical Analysis

Step 6: Multivariate Statistical Analysis

Step 7: Write-Up and Report

Within this list of seven steps, step 3 (bolded) describes the need to perform checks of data integrity. Thus, through this model, the reader clearly sees that Checks of Data Integrity (step 3) are the appropriate considerations after data entry (step 2) and prior to univariate (a.k.a., descriptive) statistical analysis.

Therefore, if these seven steps were used as a standard model of data analysis, attention would be called to the need to perform all necessary checks of data integrity. By extension, the *insidious* problems within the data that distort findings might be more readily identified and treated, which would produce a higher quality body of statistical research.

The Anonymity of Checks of Data Integrity (continued)

The first step in promoting a standard model of data analysis is to realize that the necessary details of a data analysis study are just as important as the necessary details within a surgical procedure. For example, one small detail gone awry in a surgery might result in a patient fatality. However, one small detail gone awry in a statistical analysis is no less serious.

Specifically, decisions regarding the provision of care, policy, allocation of resources, etc. are frequently based upon statistical findings. Since one small detail left unaddressed in analysis, such as a check of data integrity, could produce inaccurate findings, this one small omitted detail could result in the incorrect care being provided to patients, incorrect policy decisions, and/or the misappropriation of resources. Thus, small errors in analysis related to checks of data integrity may result in problems at a societal level.

In short, the anonymity of checks of data integrity would be reduced through the use of a comprehensive model of data analysis supporting a universal standard for conducting fundamentally sound statistical research, which incorporated making these checks.

The Four Checks of Data Integrity

In this section, the four areas in which checks of data integrity are necessary will be mentioned. Each area is appropriate for an entire newsletter. Subsequently, this section will present a rather truncated discussion. A broader and more technically comprehensive description of each of the four areas in which data integrity must be checked will be included in future, more specialized, newsletter issues.

This section of the current newsletter will present a short description of each area, along with an example of how a small detail within each check of data integrity might result in statistical findings that do not accurately reflect the data. The purpose of this section is twofold. First, the text is meant to illustrate why individuals conducting quantitative studies should implement checks of data integrity.

Second, this section is also geared to help individuals reading published statistical findings in research articles, books, and reports, to grasp the implications of whether or not checks of data integrity have been included in the report or manuscript.

Specifically, as each type of check of data integrity is discussed in this section, imagine you are reading a published quantitative study. Furthermore, imagine that the quantitative study does not describe if the respective check of data integrity was conducted. Then consider how the believability of the findings might be impacted in light of the fact that the check of data integrity was not reported. This is a simple exercise meant to underline how important checks of data integrity are within an analysis.

1) Statistical Power

The first check of data integrity concerns statistical power (Bannon, 2013). Statistical power can be thought of as the probability that a statistical procedure will indicate that the study hypothesis is true (and reject the null

The Four Checks of Data Integrity (continued)

1) Statistical Power (continued)

hypothesis) when the hypothesized relationship between variables truly exists. Briefly, an analysis needs a sufficient level of statistical power to reveal a true relationship that exists between variables. Another way of putting it is that a statistical test might indicate a true relationship between variables does not exist, solely because the data did not incorporate enough statistical power to accurately portray the relationship.

For example, assume a true (statistically significant) relationship exists between variables A and B, but also assume this is not known. Subsequently, imagine a statistical test is being applied to examine if a relationship between variables A and B exists, which uses a sample of *30 study participants*. Lastly, imagine the statistical test indicates that a true relationship between variables A and B *does not exist* (even though in reality it does).

However, now imagine while all other factors remained the same, the statistical power of the test was increased through doubling the number of study participants from 30 to 60. As a result, the same statistical test might then indicate the true relationship between variables A and B *does exist*.

This is a simple illustration of how findings are distorted by a lack of statistical power. Specifically, initially (using the sample of 30) it seemed that a true relationship did not exist between variables A and B (even though it did), simply because the analysis did not incorporate enough statistical power to reveal the relationship. However, when power was increased (using the sample of 60), the statistical analysis did indicate the true relationship between variables A and B.

Thus, when conducting analysis or reading published research it is imperative to assess if a sufficient level of statistical power exists to detect a relationship between variables. For example, if an analysis indicates a significant relationship between variables does not exist, it is beneficial to ask if the finding is accurate or alternately has there been a failure to detect a true relationship because the analysis did not incorporate a sufficient level of statistical power?

2) Test Assumptions

The second check of data integrity concerns test assumptions (Bannon, 2013). The number and type of test assumptions differ by statistical procedure. Parametric tests incorporate test assumptions, such as normality, homoscedasticity, multicollinearity, linearity, no undue influence of outlier scores, and so on. A failure to check any one of these assumptions can lead to inaccurate findings. As an example, consider the assumption where outlier scores (scores that are extremely low or high) do not have an undue effect on study findings. Using two parametric statistical tests, Bannon (2013) illustrated how a finding can be distorted by a small number of outlier scores.

Test 1: In the first test, an analysis indicated the mean score reflecting *Happiness* (Score range: 1–5) was significantly higher among 47 study participants that lived with a dog (Mean score =3.02, *SD*=.79) relative to 53 study participants that lived with a cat (Mean score=2.48, *SD*=.62) at a statistically significant level, *t*(98)=3.8, *p*<.001.

The Four Checks of Data Integrity (continued)

2) Test Assumptions (continued)

Test 2: Next, to illustrate the impact of outlier scores on study findings, three outlier scores were added to the two groups involved with the analysis. Specifically:

Three low outlier scores (scores = 1.0) were added to the 47 study participants that lived with a dog (total=50). Three high outlier scores (scores = 5.0) were added to the group of 53 study participants that lived with a cat (total=56).

The same parametric statistical analysis was repeated, which examined if the level of *Happiness* was significantly different between these two groups. This analysis indicated that there was not a statistically significant difference in mean *Happiness* scores between these two groups.

This is an example of how a few outlier scores can distort the results of a statistical analysis. Specifically, there was a true difference in *Happiness* scores between these two groups (the group that lived with a dog and the group that lived with a cat). However, these few outlier scores made it appear as if this difference did not exists.

Thus, consider that an analyst might have produced the results of Test 2 first, which indicated a statistically significant difference in *Happiness* scores *did not* exist. If that analyst did not check for the influence of outlier scores (to identify the outlier scores entered in Test 2), he or she would have accepted distorted findings that did not reflect the data.

However, if he or she did produce the results of Test 2 first, but then made a check of integrity where the influence of the outlier scores were identified and removed, the true variable relationship in Test 1 would have become evident.

Subsequently, in light of the impact of outlier scores, when conducting or reading the results of a statistical analysis, it may be difficult to regard the findings as credible as they might be if an assessment of outlier scores is not made or mentioned.

For example, if an examination of outlier scores is not made or mentioned in a study, how can one be certain the statistical results presented are not distorted via the influence of a small number of outlier scores?

As mentioned earlier, the assessment of an undue influence of outlier scores is just one of many test assumptions that may be required to perform this needed check of data integrity. Each of the other test assumptions also carries the potential to distort statistical findings. Therefore, a clear check of test assumptions, as well as report of the results of these procedures within a published manuscript is vital.

3) Missing Data

The third check of data integrity concerns missing data (Bannon, 2013). Missing data values have a significant potential to not only distort study findings, but also the generalizability of findings. However, we will only discuss the distortion of study findings here.

Most often, the term missing data refers to a survey item where the study participant did not provide a valid response. As per Bannon (2013) to assess the seriousness of missing data within a study, the researcher must first *define* how missing data is measured within a study. Next, the *percentage of study participants with missing data*, as well as *how much data each*

The Four Checks of Data Integrity (continued)

3) Missing Data (continued)

case is missing should be computed. Next, the *pattern of missing data* (i.e., MCAR, MAR, NMAR) should be indicated. Lastly, based upon all these other considerations, a *method to account* for the missing data values should be applied.

Each of the steps described above have the potential to distort study findings. For example, the choice of the method to account for the missing data values may play a significant role in producing study findings. For example, one of the most commonly skipped survey items is the question regarding *annual household income*. Traditionally, many researchers have entered the mean annual income score for study participants that do not provide a valid response to this item.

For example, within a sample of 10 study participants, 9 may report their annual income and 1 may not. Subsequently, if the mean annual income for the 9 that did report their annual income was $50,000, that number would be entered as the annual income for the 1 study participant that did not report annual income.

This is often a poor method to account for missing data as study participants that do not provide data regarding annual income tend to have income levels extremely lower or higher than average. Therefore, imputing the mean value (i.e., the average income) of the cases that did provide data is perhaps the worst method of estimating a missing value.

For example, let's assume a study were considering if *Annual Income* was associated with *Happiness*. Also, assume that study participants with extremely high annual incomes did not report valid responses and were ascribed the mean income level of all study participants. Lastly, assume that those study participants with extremely high income levels (that did not provide a valid response) also evidenced the highest levels of *Happiness*.

In this instance, because the mean income levels were entered for the study participants with extremely high income levels who did not provide valid responses, the analysis would report that the highest levels of *Happiness* are associated with the mean level of *annual income.* However, the true relationship between *Annual Income* and *Happiness* would be quite to the contrary.

If one is reading a peer-reviewed quantitative research article that does not mention the amount of missing data in a study or how missing data were accounted for, how could one be certain this undesirable form of mean imputation was not performed. This is an important question as through this common scenario, it is relatively easy to see how choices made to account for missing data values can distort study findings.

Therefore, to assure data integrity, a thorough examination of, as well as method of accounting for, missing data values is essential. Furthermore, these elements should always be described in a published research report or manuscript to bolster confidence in statistical findings.

4) Measurement Tools

The forth check of data integrity concerns measurement tools (Bannon, 2013). Assessing measurement tools largely refers to examining the validity and reliability of an instrument.

The Four Checks of Data Integrity (continued)

4) Measurement Tools (continued)

The subject of measurement tools is perhaps the easiest issue to understand regarding how checks of data integrity impact study findings. Briefly, if variables are not measured well, the value of a finding is dubious at best.

One of the simplest examples of how measurement tools impact data integrity is that some measurement tools can be assessed fully in terms of reliability and validity, while others do not. For example, if *Happiness* was measured using one single question such as *Do You Feel Happy?* (on a scale of 1-5), there would not be a great deal of room to measure the validity or reliability of the measure. Although, if *Happiness* was measured using multiple items that were combined into a composite score, the validity and reliability might be quite simple to assess. Of the two examples, the second would likely reflect a greater level of data integrity.

However, when statistical findings are reported, there is often little room to mention the richness of the measures. For example, in the scenarios above where *Happiness* is measured via a single item or richer composite score, the results of analysis would likely be stated in the same manner, such as "Analysis indicated *Happiness* was related to…" Therefore, an assessment of study measurement tools is an important check of data integrity.

Final Comments

Essentially, a broad implementation of the checks of data integrity described in this newsletter would be a significant point of professional evolution within the field of quantitative research. It is worth mentioning that the examples of how these checks can distort study findings are but a few elementary examples. Also, the newsletter discussed one challenge at a time and did not consider the possibility of a cumulative effect of many checks of data integrity going unaddressed.

However, even without these more dire scenarios, hopefully, at this point it is clearer how these checks have an enormous impact on the accuracy of statistical findings, as well as why these checks of data integrity are termed the most important questions not being asked in statistical research today in the current newsletter.

REFERENCES

Bannon, W. M. (2013). *The 7 Steps of Data Analysis: A Manual for Conducting a Quantitative Analysis*. New York: StatsWhisperer Press. Only available at:

http://www.statswhisperer.com/products/

Thanks for your interest in our newsletter!

Contact us at: wb@statswhisperer.com with any questions or for information on:

Subscribing to this newsletter visit:

http://www.statswhisperer.com/newsletter/

Books and products visit:

http://www.statswhisperer.com/products/

Webinars and seminars visit:

http://www.statswhisperer.com/seminars/