Best Practices in Data Cleaning: Debunking Decades of Quantitative Mythology

Resources for further exploration

Jason W. Osborne, Ph.D.

Chapter #	Title	Chapter resources:
1.	Why Data Cleaning Is Important: Debunking the Myth of Robustness	1. Review the author instructions for journals generally considered to be top tier or most respected in your field. See if any of them explicitly instruct authors to report testing assumptions, data cleaning, or any of the other issues we raise. 2. On our book's website, I provide links to author instructions from journals in various fields. Which journals or fields have the most explicit author instructions? Which have the least explicit instructions? Can you see any differences in the articles contained in journals that have more explicit directions for authors? 3. Review a recent study of yours (or your advisor) where statistical assumptions were not tested and where the data are still available (we all have them, and I am as guilty as everyone else). As you work through this book, apply the various data cleaning techniques and test all assumptions for all statistical tests used in the study. Perhaps all the assumptions are met and your results now have even more validity than you imagined. Congratulations! Perhaps after cleaning the data and testing assumptions, your results are changed. Sometimes that can be a positive outcome, or sometimes that can be disappointing. 4. If you have an interesting example of results and conclusions that changed after revisiting a data set and testing assumptions, I would love to hear from you at jasonwosborne@gmail.com. Send me a summary of what you found, and how things changed. I may add it to the web site so others can learn of your findings!
SECTION I: BEST PRACTICES AS YOU PREPARE FOR DATA COLLECTION		
2.	Power and Planning for Data Collection: Debunking the Myth of Adequate Power	1. Explore the data sets sampled at *N* = 20 (in .ZIP compressed format) from the population posted online to see if any of the samples that produced seriously misestimated effects can be salvaged through conventional data cleaning methods.

		2. Download the "population" data set and take a sample of your own. Calculate the correlation coefficient and see how close you are to the "true" population estimate of rho = .43. Take a smaller or larger sample and see whether your results change. 3. Take a recent study that you were involved in (or one from a good journal in your field, preferably a study your advisor published recently). Using the freely available software G*Power, calculate how much power that study had to detect a small, medium, or large effect (for example, using t-tests, *d* = 0.20, 0.50, and 0.80 for small, medium, and large effect sizes). Are you satisfied with that level of power? If not, calculate what sample size you (or your advisor) would have needed to gather in order to have sufficient power. 4. Review articles in a top journal in your field. Note how many articles mention the term *power* and in what context. What percentage appear to have performed an a priori power analysis? 5. Find an article in your field that concludes null findings (e.g., no relationship, no difference between groups). Do the authors discuss whether they had sufficient power to detect effects of reasonable magnitude? If not perform your own test to see if their conclusions are warranted. If power to detect reasonable effects is not high, their conclusions might be suspect.
3.	Being True to the Target Population: Debunking the Myth of Representativeness	1. Experiment with how ceiling and floor effects can distort results of analyses. Download the data set from the website for this book. In it you will find several variables that have generally strong correlations, such as family socioeconomic status and student achievement. While neither variable suffers from restriction of range, we can simulate a restriction of range issue. o Explore how the correlation between reading achievement (BYTXRIRR) and socioeconomic status (BYSES1) is influenced by restricting analyses to students whose parents have less than a high school education or more than a graduate school education (use BYPARED to select parent education level). 2. Review articles from a respected journal in your field (or from a list of recent articles published by your advisor). See if you can identify any of the following issues raised in this chapter. o Use of extreme groups analysis. o Mismatch between measures used and sample (possible floor or ceiling effect). o Whether there is potential restriction of range (almost any convenience sample, such as college students, will have strong possibilities of this). o Whether there might be aggregation errors (i.e., groups that were combined that might have been analyzed separately). o Whether the purpose of the article is met through the sample. If not, can you describe a sampling methodology that would have better met the goals of the study?66 Best Practices in Data Cleaning 3. Examine data you have collected for a research project (or one your advisor or colleague has collected if you do not have data on-hand) for evidence of ceiling or floor effects, restriction of range, combining of groups that may not be

		homogenous, and so on. If you do not find evidence of such effects, simulate them as I did for the examples in this chapter and explore how your conclusions would change with less ideal sampling.
4.	Using Large Data Sets With Probability Sampling Frameworks: Debunking the Myth of Equality	1. Examine a study in your field that utilized a public data set like the ones described in this chapter. Did the authors use best practices in accommodating the sampling issues? o National Center for Educational Statistics (NCES) in the United States. For example, the Education Longitudinal Study 2002 (ELS 2002), National Education Longitudinal Study of 1988 (NELS 88), and Third International Mathematics and Science Study (TIMSS). Available through the NCES website or the Interuniversity Consortium for Political and Social Research website. o Centers for Disease Control and Prevention (CDC), such as the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES). Available through the CDC website. o The Bureau of Justice Statistics, including the National Crime Victimization Survey (NCVS). Available through the bureau's website. 2. Find a data set in your field of interest that utilized complex sampling. Through reviewing the user manuals, identify the weighting variables and what design effects you might need to account for. Find out how to utilize this information in the statistical software you most commonly use. 3. Pick a relatively simple analysis (simple one-way ANOVA or simple correlation) and perform analyses of interest to you using both appropriate handling of the complex sampling and inappropriate handling of the sampling. Compare results to see how serious an error you are likely to make if you fail to appropriately model sampling in your analyses. If you do not have access to other data sets, earlier in the chapter I mentioned popular data sets in a variety of social science disciplines. 4. Pick a commonly used data set in your field that requires the use of complex sampling. Perform a search of scholarly articles published using that data set and describe what percentage of the authors appropriately modeled the sampling issues. If you find interesting data, share it with me and I will post it on the book's website.
SECTION II: BEST PRACTICES IN DATA CLEANING AND SCREENING		
5.	Screening Your Data for Potential Problems: Debunking the Myth of Perfect Data	1. Data sets mentioned in this chapter are available for download on the book's website (grades, horse-kicks). Download them and practice screening for nonnormality in the software you prefer. Identify how to perform a K-S test (with or without the Lilliefors correction) or the S-W test. 2. Explore a recent data set from your research, your advisor's research, or from a journal article you admire. Do the variables meet assumptions of normality according to the various methods discussed in this chapter?

		3. Discuss basic data cleaning with another scholar in your field. Ask whether that person routinely screens data for normality. If not, ask why not. If so, ask what methods that person relies on to determine whether the assumption is met. Other resources: Z score table
6.	Dealing With Missing or Incomplete Data: Debunking the Myth of Emptiness	1. Download from the book's website some of the missing data sets I discuss in this chapter, and see if you can replicate the results I achieved throughvarious means. In particular, I would challenge you to attempt multiple imputation. 2. Choose a data set from a previous study you conducted (or your advisor did) that had some missing data in it. Review how the missing data was handled originally. (**I also have another data set online that you can play with for this purpose.)** • Conduct a missingness analysis to see if those who failed to respond were significantly different than those who responded. • Use imputation or multiple imputation to deal with the missing data. • Replicate the original analyses to see if the conclusions changed. • If you found interesting results from effectively dealing with missingness, send me an e-mail letting me know. I will gather your results (anonymously) on the book's website, and may include you in future projects. 3. Find a data set wherein missing data were appropriately dealt with (i.e., imputation or multiple imputation). Do the reverse of #2, above, and explore how the results change by instead deleting subjects with missing data or using mean substitution. Other : • SPSS syntax for creating example data sets • SAS multiple imputation syntax • Part of NCES data used to generate examples-SPSS format
7.	Extreme and Influential Data Points: Debunking the Myth of Equality	1. Data sets from the examples given in this chapter are available online on this book's website (Univariate examples, Correlation examples, ANOVA examples). Download some of the examples yourself and see how removal of outliers generally makes results more generalizable and closer to the population values. 2. Examine a data set from a study you (or your advisor) have previously published for extreme scores that may have distorted the results. If you find any relatively extreme scores, explore them to determine if it would have been legitimate to remove them, and then examine how the results of the analyses might change as a result of removing those extreme scores. *And if you find something interesting, be sure to share it with me.* I enjoy hearing stories relating to real data. 3. Explore articles from well-respected journals in your field (some links are here). Note how many report having checked for extreme scores, and if they

		found any, how they dealt with them and what the results of dealing with them were (if reported).
8.	Improving the Normality of Variables Through Box-Cox Transformation: Debunking the Myth of Distributional Irrelevance	1. Explore how to implement Box-Cox transformations within the statistical software you use. Download one (or more) of the example data files from the book's website and see if you use Box-Cox transformations to normalize them as effectively as I did. Remember to use best practices, anchoring at 1.0. • Horse kicks data • Grades data • AAUP faculty salary and institution size data 2. Using a data set from your own research (or one from your advisor), examine variables that exhibit significant nonnormality. Perform an analysis prior to transforming them (e.g., correlation, regression, ANOVA), then transform them optimally using Box-Cox methods. Repeat the analysis, and note whether the normalization of the variables had any influence on effect sizes or interpretation of the results. If you find an interesting example, e-mail me a summary and I may feature it on the book's website. • Box - Cox syntax in SPSS
9.	Does Reliability Matter? Debunking the Myth of Perfect Measurement	1. Download the spreadsheet from the book's website that allows you to explore correcting simple correlations for low reliability. Enter a correlation, and the two reliabilities for each of the two variables used in the correlation, and examine the effects of good or poor reliability on effect sizes (particularly the percentage variance accounted for). 2. Examine a good journal for your field. Can you, like me, easily find an article reporting results where alpha was .70 or lower? Find one of these articles, correct a correlation for low reliability using the information from the article (and the spreadsheet available from this book's website). How would the author's results have looked different if the variables were measured with perfect reliability? Send me an e-mail with what you find, and I may share it on the book's website.
SECTION III: ADVANCED TOPICS IN DATA CLEANING		
10.	Random Responding, Motivated Misresponding, and Response Sets: Debunking the Myth of the Motivated Participant	1. Think about how you could include a measure (a question, an item, a scale) that would help you determine if any of your subjects are not responding thoughtfully to your measures. Create a plan to do so to examine the quality of the data you might be collecting. What actions could you take to examine this issue in your own research? 2. Examine the data set presented on the book's website. Can you identify the participants engaging in random responding? What happens to the results when they are eliminated from the analysis? Variables: • Score_t1= pretest score • Score_t2=posttest score • bad-answers= number of inappropriate answers to target items on pretest

		• bad_answers_t2= inappropriate answers on posttest
11.	Why Dichotomizing Continuous Variables Is Rarely a Good Practice: Debunking the Myth of Categorization	1. Download the data set from the book's website. Compare an analysis (such as simple correlation or regression) with continuous variables and dichotomized variables. How do interpretations and effect sizes suffer when continuous variables are illegitimately dichotomized? • math_gr8 and math_gr10 are continuous variables • math8 and math10 are categorical variables 2. Look through the best journals in your field and see if you can find an example in which an author dichotomized a continuous variable. What was the justification for doing so? Do you agree it was a legitimate analytic strategy? 3. Using one of your own (or your advisor's) data sets, explore how dichotomization can alter or damage power and effect sizes.
12.	The Special Challenge of Cleaning Repeated Measures Data: Lots of Pits in Which to Fall	
13.	Now That the Myths Are Debunked . . . : Visions of Rational Quantitative Methodology for the 21st Century	