

Teaching Students to Use Summary Statistics and Graphics to Clean and Analyze Data

John Holcomb, Cleveland State University
Angela Spalsbury, Youngstown State University

Copyright © 2005 by John Holcomb and Angela Spalsbury, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

ABSTRACT:

Textbooks and websites today abound with real data. One neglected issue is that statistical investigations often require a good deal of “cleaning” to ready data for analysis. The purpose of this dataset and exercise is to teach students to use exploratory tools to identify erroneous observations. This article discusses the merits of such an exercise and provides a team project, problem data, cleaned data for instructors, and reflections on past experiences. The main goal is to give instructors a prepared project for their students to perform realistic data preparation and subsequent analysis. The data for this project involve categorical and continuous variables for subjects age 65 and over testing calcium, inorganic phosphorus, and alkaline phosphatase levels in the blood. The project described in this article involves summary analysis, but the cleaned data could also be used for projects on independent samples t-tests, analysis of variance, or regression.

Keywords: Activity-based learning, Team projects, Data cleaning;

1. Introduction

For many years, statistical educators have emphasized the need for students to analyze real data in context (Cobb 1992, Cobb and Moore, 1997). Textbooks have risen to this challenge, and many provide references from appropriate literature as to why the data were collected and analyzed. A missing component in this effort is that real data are not often immediately ready for analysis. The data in some sense need to be “cleaned.” In fact, statistical programmers at pharmaceutical companies, medical research institutions, and survey companies spend a great deal of time readying data for analysis. This type of work consists of trying to minimize missing data and verifying the unusual values of some variables.

The goal of the project described in this article is to provide students with an opportunity to use exploratory data analysis techniques taught in the beginning of the course to identify what we call “problem values” for a team homework project. The data for this project come from a clinical investigation of calcium, inorganic phosphorus, and alkaline phosphatase levels in the blood for subjects age 65 and older where the blood was tested at several different commercial laboratories.

As implemented by the authors, the team project has one submission from each team, and the instructor evaluates the project 50% on statistical accuracy and 50% on written presentation.

Teams were self-selected, and the project counted toward 5% of the final grade. Previous literature has addressed the use of projects or case studies in statistics education reform. Holcomb and Ruffer (2000) describe their use of projects in introductory statistics. The project described in this article is similar in structure and graded in the same way projects described therein. Other specific examples of using projects in class include Fillebrown (1994), Halvorsen and Moore (2000), Ledolter (1995), Love (1998, 2000), Sevin (1995), Radke-Sharpe (2000), Smith (1998), and Vaughan (2003). Articles on the benefits of scientific writing include Radke-Sharpe (1991), Samsa and Oddone (1994), and Stromberg and Ramanathan (1996).

Before doing statistical analysis it is very important that the data have as much integrity as possible. Data cleaning is a two-step process including detection and then correction of problem values (Mason et al.). Some common sources of problem values are

- missing data
- 'blank,' coded as 0
- typing errors on data entry
- column shift-data for one variable column was entered under the adjacent column
- decimal point errors
- inconsistent coding.

Data cleaning issues are especially important when working with large databases, in particular with databases that include name and address information. Duplicate or incomplete entries are the bane of data managers. These issues are beyond the scope of this paper (cf. Kimball, 1996).

Elementary tools that can be used to clean data include descriptive statistics and statistical graphs. During analysis of a dataset, researchers faced with problem values that are missing, appear erroneous, or are not possible, often need to consult the original record containing the observation. Standard textbook limitations prohibit the inclusion of original records. The advance of technology allows us to provide this information (or mimic this process) on the web. We describe in this article a project that involves placing pseudo records for each observation on the web so that students can have the valuable experience of checking for reliability in the data as well as learning how to correct data transcription errors. It may be helpful to note that this exercise is not designed for use imputing missing values. An instructor who wishes to explore imputation, however, could adapt the project easily. For information regarding data imputation, see Little and Rubin (2002).

2. Pedagogical Uses

The dataset used for class was compiled by Boyd and Delost from the Department of Health Professions at Youngstown State University, and Holcomb (at the time) in the Department of Mathematics at Youngstown State University. The objective of the study was to determine if significant gender differences existed between subjects 65 years of age and older with regard to calcium, inorganic phosphorous, and alkaline phosphatase levels (Boyd et al., 1998). (Note that these three variables are subsequently referred to as "outcome variables" for brevity.) The researchers performed a retrospective chart review of laboratory procedures performed in 6

different physician practices. The data consisted of 178 subjects representing 92 males and 86 females aged 65 or older. Using two way analysis of variance with the factors of sex and laboratory (six levels), the study found significant differences in the means for the two sexes with respect to the outcome variables. The authors urged laboratories to create separate reference ranges for healthy individuals over age 65 categorized by gender.

Although the original data for this study had observations needing investigation, we massaged the original data to include data problems and issues that have arisen in other research projects. This was done to expose students to a range of cleaning issues that can arise when preparing data for analysis.

The discrete variables in this dataset are **sex** (1 for male, 2 for female), **lab** (one of 6 labs where blood was analyzed), and **agegroup** (five, 5 year categories, starting at 65 and ending at 89). The continuous variables in this dataset are **age** (years), **alkphos** (alkaline phosphatase, international units per liter), **cammol** (raw calcium, milligrams per deciliter), and **phosmmol** (inorganic phosphorous, milligrams per deciliter).

Students downloaded the messy dataset from a class web site and were told that there were problems with the data. They were instructed to identify unusual and missing values with their exploratory data techniques and then check the values against the patient record. The patient record is a grid of observation numbers available on the web that mimics actual patient data values. Thus if a student wanted to confirm the age of observation #34, he or she would click on obs34 at <http://academic.csuohio.edu/outlier/bigtable.htm>. The goal for the students was to find 27 incorrect data values and 2 missing data values in the variables **age, sex, alkphos, lab, cammol, phosmmol, and agegroup** and then correct these data values. They also needed to identify 4 missing values that are truly missing and not available in the original record.

We wanted students to use statistical techniques to determine which values needed confirmation by looking at the patient record. Of course, students have the option of clicking on every patient record on the web and locating any wrong values in the dataset with that approach. One reason in using this dataset for the assignment is that the sample size of 178 observations discourages such activity, without being too large for us to create the html files for each of the patient records. However, at least one team in every class has checked every observation on the web.

Another advantage of this dataset is that it deals with physiological variables (levels of various minerals and enzymes in blood) of which most of our students have no familiarity. Statisticians often have to work with variables for which they have little knowledge. This project illustrates the interdisciplinary nature of statistics and that collaboration with subject-matter experts is often important.

This article describes the class activities of cleaning the data and the summary assignment for students to complete once the data are cleaned. This data, however, can also be used for a variety of subsequent projects. Independent samples t-tests could be used with the grouping variable of gender. One way analysis of variance (ANOVA) can be applied to the lab or the agegroup variables. Two way ANOVA can be used to take into account variability from the

factors of gender and lab. Lastly, students can use age (in years) and regression analysis to predict calcium, inorganic phosphorus, or alkaline phosphatase levels.

3. Using Summary and Graphical Techniques To Find The Problem Values

At this point in the course, the students have been taught to summarize discrete variables with a frequency table and to summarize continuous variables with the mean, median, standard deviation, minimum, and maximum.

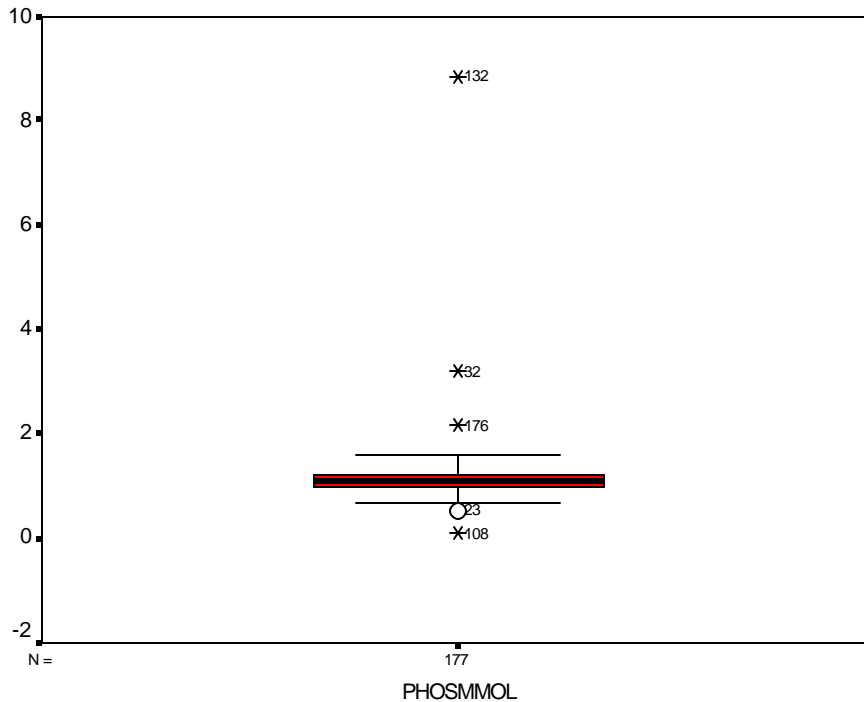
For example, creating a frequency table for **lab** yields the following output in SPSS:

Figure 2: Frequency table for lab with problem values

		LAB				
		Frequency	Percent	Valid Percent	Cumulative Percent	
Valid	1	88	49.4	49.7	49.7	
	2	41	23.0	23.2	72.9	
	3	16	9.0	9.0	81.9	
	4	13	7.3	7.3	89.3	
	5	11	6.2	6.2	95.5	
	6	6	3.4	3.4	98.9	
	21	1	.6	.6	99.4	
	43	1	.6	.6	100.0	
	Total		177	99.4	100.0	
	Missing	System	1	.6		
Total		178	100.0			

The student should immediately see the problem values of 21 and 43. The problem values for the continuous variables are also obvious, but in completely a different manner than the discrete variables. By inspecting the boxplot for inorganic phosphorus (**phosmmol**) the student should realize that something is wrong.

Figure 3: Boxplot of **phosmmol** with problem values



If one uses the web grid of observations, the student sees the correct value of **phosmmol** is .84 for observation 132. Checking observation 108 shows the correct value of **phosmmol** is 0.9 and the student should begin to surmise that the outliers need to be checked. Students could also examine a histogram and see that something looked amiss as well. With the variable age, the boxplot looks even stranger. One corrupt value given for age was 771. From inspection it is unclear if the true age is 77 or 71, but obviously 771 is not correct. Students need to find the correct age by clicking on the web patient record for that observation number.

In regard to missing data, one of the missing values is the value of **agegroup**, which can be determined from the actual age. Another article (Johnson, 1996) addressing a similar issue had students correct erroneous data values by using redundant variables available within the dataset. (In this case height, weight, and body mass index (BMI) were given, allowing the student to infer the apparent correct height for a man using the BMI and weight.) There is also a missing **alkphos** value which can be found by looking at the patient record. The remaining four missing values are not available on the patient record, thus exhibiting what often occurs in real research. A complete table of problem values (with the variable name and observation number) with the correct value appears in Appendix B.

Instructors should note for the continuous variables that all problem values are either missing or detectable as outliers. However, not all outliers are problem values. For example, even when the dataset is cleaned, outliers still exist in all three of the outcome variables. This is good reinforcement for students who sometimes feel all outliers are erroneous or need to be removed.

Once the student believes that their dataset is cleaned and has integrity, they can then complete the official summary. We require students to provide a summary of the discrete variables. We

also require them to give summary statistics and to construct side-by-side boxplots for the continuous variables to compare the levels of the minerals and enzymes between men and women and between laboratories (see Appendix C).

4. Modifications from Experience

After grading one such set of assignments, we have decided that in the future, we will require students to catalogue the problem observations in the data and the changes that they made to clean the dataset. It is clear from the summary statistic values if the students found all of the problem values, however, it will make grading easier if a paragraph summarizing the changes made appears in the beginning of the report. Teaching students to record the changes that they made to data is good training, since real analysis requires keeping track of changes made to source data. This will also assist in grading if discrepancies appear between the student's calculations and the instructor's calculations.

We would like for students to come to the realization that cleaning data may or may not affect the conclusion. With this in mind, in future assignments, we will require students to compare one or more of the project analyses between the original and the cleaned data (i.e., sensitivity analysis). As one can guess, the erroneous age values greatly inflate both the mean and standard deviation. With the outcome variables, the results are not so clear. Note that the suggested changes are already incorporated into the assignment as presented in Appendix C.

One issue that we faced was that students often asked us if they had found all of the problem values. This is problematic because in real statistical analysis, the analyst never really knows if he or she has found all incorrect data values. When presented with the number of problem observations a group found, we often told the students if they had found them all or not. This is not ideal, and we hope to avoid doing this in the future. Since all of our students found all of the problem values, it appears that there is little difficulty in locating all of the problem values. When the first listed author gave the team class time to begin the assignment, he noted that all the student teams found the problem values within an hour. Also, he observed different student groups updating each other on their progress in finding the problem values. This kind of cross-class collaboration is not necessarily harmful, and it is probably impossible to avoid.

The most important issue that we faced with the written project reports was the lack of substantial commentary on the graphs and tables. Students were reluctant to infer any conclusions from the summary graphics. We feel this is most likely a result of a variety of factors, one of which may be that this was the first time students were asked to express in words their statistical thinking. To address this concern, we have decided that in the future we will give a simple assignment prior to this project. The assignment will contain a dataset with only a few well understood variables. We will require students to perform a summary analysis with at least one or two well written paragraphs analyzing the data. This assignment will also teach students how to cut and paste statistical graphs and tables, with appropriate labels, into their word processing program.

5. Getting the Data

The file calcium.dat contains the data with the problem values. The file calciumgood.dat contains the data with the problem values corrected. The observation grid can be found at <http://academic.csuohio.edu/holcombj/outlier/bigtable.htm>. The file calcium.txt is a documentation file that contains a brief description of the dataset and the purpose of the assignment.

Acknowledgements

The authors wish to thank the reviewers and the Datasets and Stories editor for their helpful comments.

Appendix A: Key to variables in calcium.dat and calciumgood.dat

Calcium.dat

Columns	Variable	Comment
9-11	OBSNO	Patient Observation Number
21-22	AGE	Years
33	SEX	1=Male, 2=Female
42-44	ALKPHOS	Alkaline Phosphatase International Units/Liter
55	Lab	Lab: 1=Metpath; 2=Deyor; 3=St. Elizabeth's; 4=CB Rouche; 5=Youngstown Osteopathic Hospital; 6=Horizon
63-66	CAMMOL	Calcium mmol/L
74-77	PHOSMMOL	Inorganic Phosphorus mmol/L
88	AGEGROUP	Age group 1=65-69; 2=70-74; 3=75-79; 4=80-84; 5=85-89 Years

Calciumgood.dat

Columns	Variable	Comment
9-11	OBSNO	Patient Observation Number
20-22	AGE	Years
32-33	SEX	1=Male, 2=Female
42-44	ALKPHOS	Alkaline Phosphatase International Units/Liter
54-55	Lab	Lab: 1=Metpath; 2=Deyor; 3=St. Elizabeth's; 4=CB Rouche; 5=Youngstown Osteopathic Hospital; 6=Horizon
62-66	CAMMOL	Calcium mmol/L
74-77	PHOSMMOL	Inorganic Phosphorus mmol/L
88	AGEGROUP	Age group 1=65-69; 2=70-74; 3=75-79; 4=80-84; 5=85-89 Years

Appendix B: Description of problem values

Observation Number	Variable	Problem Value	Correct Value
6	lab	43	4
11	age	771	71
21	cammol	3.2	2.2
22	alkphos	Missing	64
25	cammol	25.3	2.53
26	cammol	20	2
27	cammol	22.3	2.23
28	sex	22	2
28	cammol	24.3	2.43
29	cammol	25	2.5
30	cammol	23.3	2.33
31	cammol	24	2.4
32	cammol	25	2.5
32	phosmmol	3.21	1.23
33	cammol	25	2.5
34	cammol	23.5	2.35
35	cammol	22.5	2.25
36	cammol	24.5	2.45
37	cammol	23.3	2.33
42	phosmmol	Missing	Missing
53	age	699	69
60	alkphos	9	97
78	lab	21	2
79	agegroup	Missing	4
85	cammol	Missing	Missing
105	age	Missing	Missing
108	phosmmol	0.09	0.9
120	sex	21	1
123	age	730	73
132	phosmmol	8.84	0.84
149	cammol	1.05	2.05
170	lab	Missing	Missing
173	sex	12	2
176	phosmmol	2.16	1.26

Appendix C: Assignment given to Students

The objective of the study for which you will analyze the data was to determine if significant gender differences existed between subjects 65 years of age and older with regard to calcium, inorganic phosphorous, and alkaline phosphatase levels (Boyd et al., 1998). The researchers performed a retrospective chart review of laboratory procedures performed in 6 different physician practices. The data consisted of 178 subjects representing 92 males and 86 females age 65 or older. In the dataset, there are three discrete variables, **sex**, **lab**, and **agegroup**. The coding is as follows:

Sex	1=Male; 2=Female
Lab	1=Metpath; 2=Deyor; 3=St. Elizabeth's; 4=CB Rouche; 5=YOH; 6=Horizon
Agegroup	1=65-69; 2=70-74; 3=75-79; 4=80-84; 5=85-89 years

The other variables of **age** (years), **alkphos** - alkaline phosphatase (IU/L), **cammol** - calcium (mmol/L), and **phosmmol** – inorganic phosphorus (mmol/L), are continuous.

1. The first task of the assignment is to check the validity of the data. Determine if this is a "messy" dataset with variable values that appear incorrect. Attempt to recover the correct values by looking up the true values from the actual data records. Copies of these can be found on <http://academic.csuohio.edu/holcombj/outlier/bigtable.htm>. Be sure to catalogue the problem values in the data and the changes that were made to clean the dataset. Include a paragraph detailing the steps taken to clean the dataset.
2. Once the data are "clean", perform a summary analysis of the three discrete variables (**sex**, **lab**, and **agegroup**). For the variables **alkphos**, **cammol** and **phosmmol**, report the mean, median, standard deviation, min and max broken down by **sex**. Also summarize the variables **alkphos**, **cammol** and **phosmmol** in a similar way with the factor variable as **lab**.
3. Construct side by side boxplots of the variables **alkphos**, **cammol**, and **phosmmol** with the factor variable as **sex**. Next construct side by side boxplots of the **alkphos**, **cammol**, and **phosmmol** continuous variables with the factor variable as **lab**.
4. Compare the mean and standard deviation of **age**, **alkphos**, **cammol**, and **phosmmol** from the messy dataset with the mean and standard deviation from your cleaned dataset. Does cleaning the data make a difference? Explain.
5. Using your summary statistics and your side-by-side boxplots, do you believe a significant difference exists in **alkphos**, **cammol**, or **phosmmol** levels with respect to **sex**? Why or why not? Do you believe a significant difference exists in **alkphos**, **cammol**, or **phosmmol** levels with respect to **lab**? Why or why not?
6. Suppose Mr. and Mrs. Contrarian are married and Mrs. Contrarian has lower calcium than Mr. Contrarian. She refuses to believe the results of the study that men tend to have lower calcium

than women because she has lower calcium than her husband. Using your results to question #3, explain to Mrs. Contrarian the flaw in her thinking.

7. One of the objectives of this research was to propose a reference range of values that are to be considered “normal” for calcium, inorganic phosphorus, and alkaline phosphatase. Looking at the results for **cammol** alone for each of the labs, explain why a single reference range is so difficult to establish.

REFERENCES:

- Boyd, J., Delost, M., and Holcomb, J., (1998), "Calcium, phosphorus, and alkaline phosphatase laboratory values of elderly subjects," *Clinical Laboratory Science*, 11, 223-227.
- Cobb, G. (1992), "Teaching Statistics," in *Heeding the Call for Change: Suggestions for Curricular Action*, ed. L. A. Steen, pp. 3-43, Washington, DC: Mathematical Association of America.
- Cobb, G., and Moore, D. S. (1997), "Mathematics, Statistics, and Teaching," *The American Mathematical Monthly*, 104, 801-823.
- Fillebrown, S. (1994), "Using Projects in an Elementary Statistics Course for Non-Science Majors," *Journal of Statistics Education* [Online], 2(2), (<http://www.amstat.org/publications/jse/v2n2/fillebrown.html>).
- Halvorsen, K.T. and Moore, T.L., (2000), "Motivating, Monitoring and Evaluating Student Projects," in *Resources for Undergraduate Instructors Teaching Statistics*, MAA Notes, No. 52, Washington DC: Mathematical Association of America.
- Holcomb, J. and Ruffer, R. (2000). "Using a term-long project sequence in introductory statistics," *The American Statistician*, 54, 49-53.
- Johnson, R. W., (1996). "Fitting Percentage of Body Fat to Simple Body Measurements," *The Journal of Statistics Education*, 4(1), (<http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html>).
- Kimball, R. (1996). "Dealing with Dirty Data," *DBMS*, 9(10), (<http://www.dbmsmag.com/9609d14.html>).
- Ledolter, J. (1995), "Projects in Introductory Statistics Courses," *The American Statistician*, 49(4), 364-367.
- Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: Wiley – Interscience.
- Love, T. E. (1998). "A Project-Driven Second Course" *Journal of Statistics Education* [Online], 6(1), (<http://www.amstat.org/publications/jse/v6n1/love.html>).
- Love, T. E. (2000). "A Different Approach to Project Assessment" *Journal of Statistics Education* [Online], 8(1) (<http://www.amstat.org/publications/jse/secure/v8n1/love.cfm>).

Mason, J., Gillenwater, K., Pugh, R., Kenefik, E., Collins, G., Whitaker, M., Volk, D. "Practical Analysis of Nutritional Data (PANDA)," Chapter 2, [Online], (<http://www.tulane.edu/~panda2/Analysis2/ahome.html>).

Radke-Sharpe, M. (1991), "Writing as a Component of Statistics Education," *The American Statistician*, 45, 292-293.

Radke-Sharpe, N., (2000), "Curriculum in Context: Teaching with Case Studies in Statistics Courses," in *Resources for Undergraduate Instructors: Teaching Statistics*, MAA Notes, No. 52, Washington DC Mathematical Association of America.

Samsa, G., and Oddone, E. Z. (1994), "Integrating Scientific Writing Into a Statistics Curriculum: A Course in Statistically Based Scientific Writing," *The American Statistician*, 48, 117-119.

Sevin, A. (1995), "Some Tips For Helping Students In Introductory Statistics Classes Carry Out Successful Data Analysis Projects" in *ASA Proceedings of the Section of Statistical Education*, 159-164.

Smith, G. (1998), "Learning Statistics By Doing Statistics" *Journal of Statistics Education* [Online], 6(3), (<http://www.amstat.org/publications/jse/v6n3/smith.html>).

Stromberg, A. J., and Ramanathan, S. (1996), "Easy Implementation of Writing in Introductory Statistics Courses," *The American Statistician*, 50, 159-163.

Vaughan, T. S. (2003), "Teaching Statistical Concepts With Student-Specific Datasets" *Journal of Statistics Education* [Online], 11(1) (<http://www.amstat.org/publications/jse/v11n1/vaughan.html>).

SUBMITTED BY:

John P. Holcomb, Jr.
Cleveland State University
2121 Euclid Ave., RT 1515
Cleveland, OH, 44115-2214
j.p.holcomb@csuohio.edu

Angela Spalsbury
Youngstown State University
One University Plaza
Youngstown, OH 4455
angie@math.yasu.edu