Data Quality's Role in Sound Inference: Critical, Yet Commonly Ignored in Teaching and Practice, and Due for Increased Emphasis

Sylvia Kuzmak
Rise Coaching and Consulting LLC, NJ, USA

Introduction

Data quality's role in sound inference is critical, as expressed by the saying, "Garbage in, garbage out." Assessment of data quality is dependent on the questions being addressed, and ensured through appropriate data collection procedures given the questions addressed. Three illustrations of the trend of diminishing emphasis on the importance of data quality in teaching are provided: (1) diminished emphasis in the GAISE College Report from 2005 to 2016, (2) CAUSE cartoon contest winning caption overlooking data collection procedures and overvaluing data cleaning, and (3) published model class exercises overlooking basic data quality principles. With the rise of big data and concomitant lack of control and involvement in details of data collection by the analyst, the breadth of importance of data quality is commonly ignored. Four ways to provide increased emphasis on data quality in teaching are identified, expressed through reference to the three illustrations, and drawing from Kuzmak (2016).

Illustration 1: GAISE College Report (2016) has Dropped Focus on Data Quality

Message on data quality featured prominently in the GAISE College Report (2005) [and dated from the 1992 Cobb Report]:

- One of "the basic elements of statistical thinking" is understanding "*The importance of data production.* Recognizing that it is difficult and time-consuming to formulate problems and to get data of good quality that really deal with the right questions. Most people don't seem to realize this until they go through this experience themselves. (Introduction, p.8).
- "GOALS FOR STUDENTS IN AN INTRODUCTORY COURSE: WHAT IT MEANS TO BE STATISTICALLY EDUCATED. … Students should understand the parts of the process through which statistics works to answer questions, namely, How to obtain or generate data …" (p. 12, underlining added for emphasis).

Focus on Data Quality has been dropped in the GAISE College Report (2016):

- The description of "statistical thinking" no longer calls out the understanding of "the importance of data production" and the need "to get data of good quality."
- The "Goals for Students in Introductory Statistics Courses" no longer has an item to "understand the parts of the process through which statistics works to answer questions" including the initial fundamental step of "how to obtain or generate data." Numerous mentions of "data" appear in the report, but not with focus on the quality of the data.

Recommendation:

1. The GAISE College Report has and will continue to have great influence on statistics education practice. Data quality is critical to sound inference, and explicit emphasis on data quality should be added to the report, including addressing the importance of ensuring and evaluating data quality. (See topics identified in Recommendations for Illustration 3.)

Illustration 3: Published model class exercises with student-generated data

Consider the following class exercise, typical of published model class exercises:

In a class of 26 college students, students are asked to pair up, and to interview each other to ask each other their planned major and record their response. The data on each student's planned major are pooled together. A bar chart showing the frequency of planned majors for the class is generated.

Several characteristics of this exercise foster poor data quality:

- The question(s) that the study aims to address is(are) not defined or shared beforehand. The question of interest will influence how the data should be collected, e.g., whether there should be a response category of "not decided," or rather, one's "most likely major" should be the response when one is undecided; and whether, in the case of a double major, both majors should be recorded for the person, weighted as one-half each.
- There are 26 different data collecting personnel in this exercise, who have not predetermined a uniform data collection procedure addressing response options, definition of terms, whether there is privacy of response, etc. So the data collection will be inconsistent.
- Is the sample representative of the population about which one wants to draw conclusions? (For example, has a sports or theatrical performance event conflicting with class time led to certain segment(s) of the class to be absent at the time of data collection?)

Recommendations:

3. Explicitly teach the critical importance of data quality, and that data quality should be evaluated for any prospective data source; that data quality should not be assumed.

One lesson is to provide a data set and pose questions related to descriptive statistics and relationships between variables. When students provide responses based on statistics calculations from the data set, the teacher informs them that they are all wrong. Subsequent graphical review of the data set reveals some clearly "bad" data, e.g., from dropped digits or numbers based on different assumed units of measure. Statistics are recalculated based on the "cleaned" data, with vastly different results. This was an exercise in an undergraduate statistics course at Princeton Univ. when the author was an undergraduate, when John Tukey headed the Statistics Dept. and at the advent of Tukey's Exploratory Data Analysis.

4. Teach how to ensure data quality, including how to avoid the data quality problems in the 'Illustration 3' exercise above. Make the point that some techniques are data domain specific.

5. Regularly remind students to consider data quality, with every exercise being an opportunity for such reminding. Include caveats when exercises are done that oversimplify data collection and evaluation. If the message of needed attention to data quality is not regularly reinforced, one is implicitly teaching indifference to data quality, which is a misapplication of statistics.

Illustration 2: CAUSE Cartoon Caption Contest has Flawed Messaging on Data Quality

Winning (but Flawed) Cartoon Caption: "Don't Forget to Clean Your Dirty Data!" (Feb. 2018)

- Data quality is ensured by appropriate and consistently applied data production/ collection procedures, and is dependent on the questions being addressed.
- If sufficient data quality is not built in to the data set, "data cleaning" does not apply, unable to compensate for "garbage in, garbage out."

Better Caption: "As goes the lab, so go the data." (Honorable Mention, Feb. 2018)

Recommendation:

2. Great cartoon to introduce lessons on: The Importance of Ensuring and Evaluating Data Quality. (See topics identified in the Recommendations for Illustration 3.)

Conclusion

More emphasis is needed on data quality in teaching and practice! The recommendations in this poster are a start!

Reference

Kuzmak, S. (2016). Mapping Knowledge for Probability and Statistics Application: Mathematical and Non-Mathematical, *pp.5-8*. International Association for Statistical Education (IASE) Roundtable Conference Proceedings, https://iase-web.org/Conference_Proceedings.php?p=Promoting_Understanding_of_Statistics_about_Society_2016