Making Decisions with Data

Gail Burrill, Michigan State University

Abstract Information is transforming the way people live and the way they do business. Making decisions based on this information is an increasingly critical skill. The data analysis strand in *Principles and Standards for School Mathematic*s describes the important ideas that are central to understanding how to process information and to use it wisely. School work in data analysis and statistics has primarily centered on making plots, organizing data, reading graphs, and calculating statistics. At the secondary level students should apply this knowledge and the skills they have developed, putting them into a context where they learn to make good decisions based on well-designed experiments with probability and randomness as tools in making these decisions. NCTM has produced *Navigating Through Data Analysis: Grades 9-12* to help teachers make this approach a reality. This paper briefly describes the book and discusses in detail one chapter's activities, where students formulate questions for data from a given context, choose appropriate graphs to analyze the data, and use simulation and informal inference procedures to find answers and make decisions.

Introduction

H. G. Wells wrote "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write" (Huff, 1954). The National Council of Teachers of Mathematics (NCTM) recognized this importance by identifying data analysis and probability as one of the content standards in *Principles and Standards for School Mathematics* (NCTM, 2000). *Principles and Standards* suggests that in the early grades students should spend time on exploration and techniques to graph data and calculate and interpret summary statistics. They should begin informal experiences with inference. In upper secondary school, data analysis should build on the foundations established in the lower grades. Students should apply the procedures they have learned and deepen their understanding of what it means to analyze data. The curriculum for all students should also focus more on informal inference-finding ways to generalize about patterns employing the exploratory skills they learned in earlier grades in the context of drawing conclusions and making decisions.

To help teachers visualize the standards in their classrooms, NCTM sponsored the Navigation series, a set of documents framing the standards in each of the content areas and grade levels described in the Standards, giving example s and discussions about their implementation in the spirit of the Standards. This paper briefly describes *Navigating Through Data Analysis: Grades 9-12,* which focuses on the use of statistical techniques and process of reasoning to make decisions, and expands on one chapter in more detail.

In the 9-12 data analysis book, students investigate simple random samples, the use of random samples to reduce the possibility of bias, and the relation between sample size and variability. The next section contains activities that explore an actual case study involving potential discrimination and, through simulation, students decide whether the data support a certain hypothesis. The third chapter, the focus of this paper, investigates a case study involving cholesterol levels, explores the relationship between two variables, and considers a variety of approaches for analyzing the data. This chapter also uses simulation to understand how likely it is that an observed outcome will occur by chance. The next section embeds the earlier examples in the context of how the data were collected by considering the design of studies. The final chapter suggests several activities that teachers might assign to give students the opportunity to put statistical thinking into practice. The examples use data analysis and sampling distributions generated through simulations as bases for informal inference in the context of real situations.

NCTM believes that learning is more than acquiring knowledge. Learning implies the ability to use knowledge in meaningful ways, to deploy knowledge to accomplish a task. To learn "with understanding, actively building new knowledge from experience and prior knowledge" (NCTM, 2000, p. 11), students must be involved in developing ideas and exploring conjectures. The following guidelines for teaching statistics are offered to help foster this kind of learning. Students should

- understand the need for and be able to formulate well-defined questions,
- begin with a picture of the data,
- be involved with organizing and analyzing the data,
- have hands-on experience producing data,
- identify assumptions,
- experience randomization,

- explore and experiment before using formal algorithms,
- develop a conceptual understanding of statistical concepts, and
- focus on the big ideas, not the rules.

These guidelines frame the development of the problem and solution strategies described below.

The Problem: Diet and Cholesterol

Table 1 contains data from a study investigating the effect of dietary change on cholesterol levels of 24 hospital employees who voluntarily switched from "a standard American diet" to a vegetarian diet for one month. The data represent their cholesterol levels (in milligrams per deciliter denoted mg/dL) both before and after the dietary change. The purpose of switching to the new vegetarian diet is assumed to be to decrease cholesterol level. The focus is on the data; issues of data quality and selection of sampling units are considered later in the book.

Table 1: CholesterolLevels Before and After Changing Diets

Before (mg/dL)	After (mg/dL)	Before (mg/dL)	After (mg/dL)
195	146	169	182
145	155	158	127
205	178	151	149
159	146	197	178
244	208	180	161
166	147	222	187
250	202	168	176
236	215	168	145
192	184	167	154
224	208	161	153
238	206	178	137
197	169	137	125

(Rosner, 1986)

As indicated in the guidelines, formulating a problem that is relevant for a set of data is critical in the decision making process. ("Students should formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them (NCTM, 2000, p. 401).") Certain data analysis techniques will illuminate different aspects of the data; whether these aspects answer the relevant questions is often overlooked by the beginning researcher. What questions can be answered from the study? In this case the questions might be: Is the diet effective? Will the diet make a difference for any patients? If so, which ones? Can a doctor predict, with some degree of certainty, the change in a patient's cholesterol after taking the diet? Once the question of interest has been clearly identified, the first element in analysis is to find a graph that may help answer that question. Understanding the strengths and weaknesses of graphs in relation to specific questions is essential for making sense of the situation. What can you learn from a box plot that you cannot learn from a dot plot and how will what you learn help you?

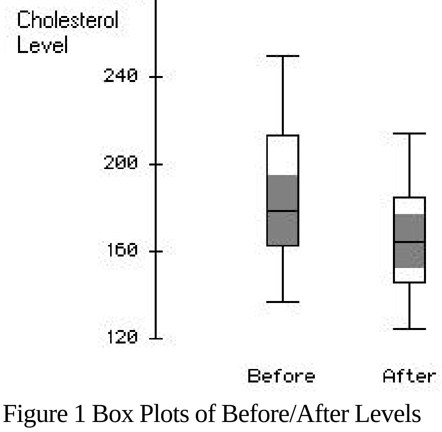
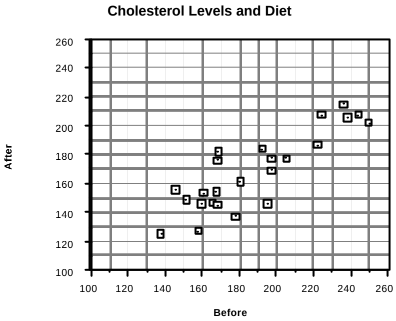


Figure 1 Box Plots of Before/After Levels

Some students begin by asking, "Does the diet make a difference?" They construct box plots of 'Before' cholesterol levels and 'After' cholesterol levels to answer the question. However, using box plots treats the data as two independent samples; students fail to recognize that this display ignores the pairing of before/after (Figure 1). While the after plot shows overall lower levels, it is not possible to see for whom this change occurred. The overlap makes any *improvements* in level difficult to determine from this analysis. ("Students should understand histograms, box plots, and scatter plots, and use them to display data (NCTM, 2000, p. 401).")

Some students make a scatter plot of (*Before*, *After*) to answer the question of whether the diet makes a difference (Figures 2, 3).



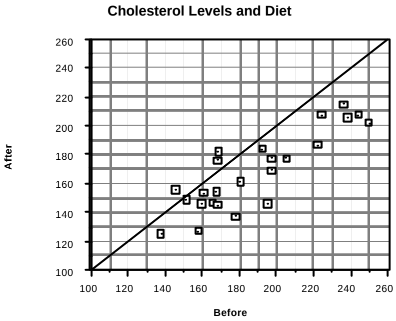


Figure 2 Scatter Plot (*Before*, *After*)

Figure 3 Scatterplot with Before = After line

If there were no difference, the initial cholesterol levels would not change with a new diet, and the points would all be on the line *Before* = *After* (the $y = x$ line), which provides a reference for analyzing how the data reflect changes in cholesterol due to diet.

Some students are concerned about the degree of change. What was the largest decrease anyone experienced? What was the least improvement? How many individuals' levels decreased? The variable of interest for these questions is change in level: *Improvement* = *Before* – *After*. A histogram can show the change in level. Figure 4 indicates that a typical Improvement is about 20 mg/dL after the diet, where 20 is the center of the distribution.

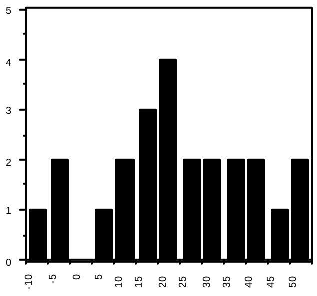


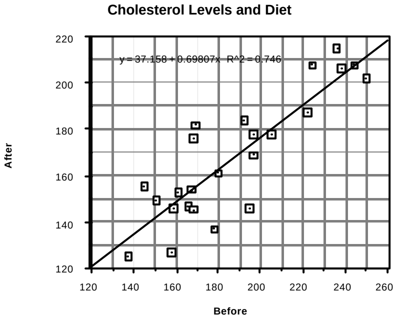
Figure 4 Histogram of *Improvement*

Using this plot, thinking about the center and spread, you can predict the Improvement of a randomly selected individual with some sense of how that prediction may vary, although the initial level is not taken into account. Note that simply calculating the mean change (about 20) does not give any indication of the variability around that mean. Are some changes very small and others very large?

Are the changes small and centered with the rest of the data? The graphs make the variability in the data visible.

The previous plots of single variable distributions did not take advantage of the fact that the *Before* and *After* scores are paired by the individuals from which they were taken. How does a current cholesterol level affect what might happen if someone uses the diet? Can a doctor predict the change in cholesterol level for a patient if the diet were used? Can an individual expect his cholesterol level to drop? If so, how much? Such questions require a plot that gives information about the relationship between *Before* and *After*.

A typical student approach is to fit a line to the data and to base conclusions only on the slope of the resulting equation (Figure 5.) The scatter plot shows that *After* and *Before* seem to have a linear relationship



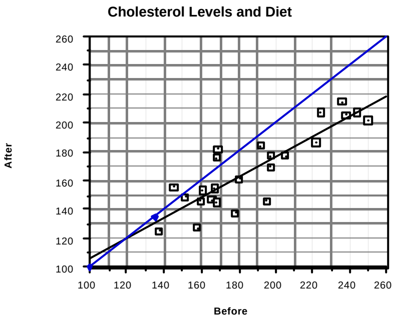


Figure 5 Least Squares Line *After* vs. *Before*

Figure 6 Least Squares Line with *After* = *Before* Line

(“For bivariate measurement data, students should be able to display a scatter plot, describe its shape and determine regression coefficients, regression equations…. (NCTM 2000, p. 401).”) The equation of the least squares regression line is approximately: *After* = 37 + 0.7(*Before).* This equation and its corresponding line predict the **average** *After* level for given *Before* levels.

Combining the regression line and the *After* = *Before* line (Figure 6) adds more to the analysis, helping answer the question: Does the diet work and for whom? The region below the *After* = *Before* line represents improvement. Points above the line represent increased cholesterol levels, and points on the line represent no change. The point at which the least squares line meets the *After* = *Before* line, about (123, 123), represents the point at which the predicted effect of the diet is “no change.” This me ans that for initial cholesterol levels greater than 123 (the “*Before*-coordinate” of the intersection point), the diet is, on average, effective. For lower *Before* levels, it is not. The plot also shows that the larger the initial cholesterol level is, the greater the predicted improvement after the diet because, as initial cholesterol level increases, the vertical distance from the *After* = *Before* line and the regression line increases.

This graph has two important ideas. First, the *After* = *Before* line can be used to identify improved cholesterol levels. Second, comparing the two lines shows that the average amount of improvement with the diet is related to the initial cholesterol level before the diet. People with higher initial levels tend to have larger decreases in levels, on average, than do people with lower initial values. Could a doctor predict a cholesterol level for a patient knowing his initial cholesterol level, say 220? The doctor might say 37+0.7(220) or 191. However, the graph of Figur e 6 indicates that, rather than a single value, there is a range of possible outcomes; there is variability around the line. The least squares line is not perfect at prediction. The diet does not account for all the variation in the *After* cholesterol levels. In some cases, the range of likely *After*-diet cholesterol levels includes the no-*Improvement* line. Thus, a doctor might use the information from the plot to predict a level and a range in which that level might reasonably fall.

Inferences

An important question remains. There seems to be improvement but is that improvement enough to make the diet a desirable treatment? Inference asks the question “Are the observed improvements possibly due to chance, or are they real? (“Students should develop and evaluate inferences and predictions that are based on data (NCTM 2000, p. 401).” If the diet actually has no effect, you might expect that increase and decrease occur purely randomly. What is the probability that at least 21 volunteers (out of 24) would improve in a study like the one under consideration? If someone’s cholesterol level is measured several times, the measurements would be likely to vary within some interval. Is it possible that the “improvement” observed in the two columns of data from the study represent just the natural variability of individual cholesterol levels? That is, could the diet actually have no effect? Inference can help answer: Did “enough” of the volunteers actually improve? And was the amount of improvement actually enough to be meaningful? Both questions can be answered with simulations, and the underlying reasoning process is identical in both cases.

The “number who improve,” a success, can be considered the statistic of interest. If the diet actually has no effect, you would expect about half the outcomes to be successes and about half to be failures since each outcome is due only to chance, not due to the diet. In a set of 24 volunteers you would expect about 12 to improve. In this case, 21 volunteers improved. What are the chances that as many as 21 would improve if the diet really does not work? A simulation where the variable to observe is “number who improve” can be used to investigate the question. (“Students should use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions (NCTM 2000, p. 401).”) A sampling distribution of “number who improve” generated for sets of 24 simulated “volunteers” with each volunteer having a fifty-fifty chance of improving can be used to estimate the likelihood that at least 21 volunteers would improve if the diet had no effect. One such sampling distribution is illustrated in Figure 7. Note that in this simulation, the estimated probability of having 21 or more

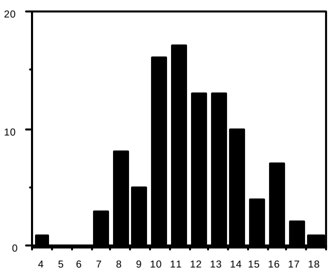


Figure 7: Sampling distribution of Simulation

successes by chance alone is zero. Answering the second question, using the actual amounts of improvement, is discussed in the actual chapter and is beyond the scope of this paper.

Conclusion

The vast amount of data generated in today's information world makes it crucial that all students understand how to use data to make intelligent and justifiable decisions. . The message in *Navigating Through Data Analysis: Grades 9-12* is the relevance of data analysis and the need for secondary students to go beyond learning procedures and skills to use them to make reasoned decisions. Diet and Cholesterol involved making a decision in a situation where there was variability and uncertainty. Students are given the opportunity to work through an analysis applying procedures and techniques they have already learned, evaluating what each might bring to the situation. In a literate society individuals should understand the need to do a careful analysis of the data they receive at work and in their personal lives to make reasoned decisions. Giving students real situations where they can apply their knowledge will help those in statistics education fulfill their responsibility to see that this goal is reached.

References

Hoff, D. (1954) *How to Lie with Statistics*. New York: Norton W. W. & Company Inc.

Lott, J. & Burrill, G. (Eds.) (in press). *Navigating Through Data Analysis: Grades 9-12*. Reston, VA: National Council of Teachers of Mathematics

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author

Rosner, B. (1986). *Fundamentals of Biostatistics*. Boston, MA: Duxbury Press.