DATA-DRIVEN LEARNING STRATEGY IN INTRODUCTORY STATISTICS COURSES

Xiaoyi Ji
Utah Valley University, USA
Xiaoyi.ji@uvu.edu

A data-driven, statistical thinking-driven, and problem-solving-driven method for teaching descriptive statistics in introductory statistics is introduced. It emphasizes collecting different types of real data first, and then visualizing data based on the characteristics of the data combined with statistical thinking, finding summarized information by reading data distribution, and finally learning problem-solving skills with understandable math formulas.

INTRODUCTION

An introductory statistics textbook, in general, starts with some important concepts. The first lesson to the expectant or uneasy students would be compiled of definitions such as data, statistics, descriptive statistics, inferential statistics, variable, the level of measurement of a variable and so on described with meticulous statistical languages. It's the same as pouring out unfamiliar words about various statistical teaching methods, teaching techniques, statistical software, and assessments of students learning outcomes to a person who has never been a statistics teacher before. Rather than starting with teaching concepts and definitions, would it be a good approach for teachers to introduce students to collect various types of data designed with questions before learning statistical concepts and basic statistical methods in teaching descriptive statistics in introductory statistics courses? The study is designed to provide students chance to collect six different types of real data and guide them to understand the basic statistical concepts using the data, visualize the data based on its characteristics and corresponding statistical questions, understand the data through reading its distribution, and investigate summarized information of the data through data distribution.

LEARN STATISTICAL CONCEPTS THROUGH COLLECTING VARIOUS TYPES OF REAL DATA WITH CORRESPONDING STATISTICAL QUESTIONS

A data-drive technique in an introductory statistical class is manifested in collecting real data by students themselves and learning basic statistical concepts and methods which rely on the data naturally. In the first day of class of introductory statistics, students are advised to collect six different types of data with specified orders in the following:

- Data set I: Access Canvas and download the data named "time taken for homework 16", which I refer to as "Time Taken" later in this paper.
- Data set II: Toss a pair of dice 100,000 times and record the sum of the face numbers on the pair of dice in R.
- Data set III: Introduce your grade in school to the class and record the information.
- Data set IV: Download "The Age of Death in a City" from https://catalog.data.gov/dataset/public-health-statistics-life-expectancy-by-community-area-0a0c4.
- Data set V: Record your grades from last semester with course number and credit hours.
- Data set VI: The age and corresponding blood pressure of family members who are 18 or older.

The first question to ask students could be: where does the data come from? The above six data sets are from different resources: websites (data I and IV), *simulation* in R (data II), *observational studies* (data III and V), and *sample survey* (data VI). The six data contain *single variable data, bivariate data, grouped data, weighted data, simulated data* and *multivariate data*, which cover most of the types of data in introductory statistics courses. Carefully chosen data sets provide the necessary background to students for understanding the statistical concepts at the first day of the class. For instance, a sequence of statistical concepts can be explained when students study the Data set I (see Table 1) with a statistical question of "how much time did UVU STAT 1040 online students spend doing the homework assignment from chapter 16?" The collection of

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

students who take the online course is a *population*, and the collection of observed students in the data is a *sample*. The information is referred to as *data* in statistics, even though individual *observation* is not available in the data set. The data can be interpreted as a *grouped data* for the *variable* of "Time Taken". The variable, time taken for homework, is a *quantitative* and *continuous variable* because the values of "Time Taken" are numerical values with infinite uncountable possibilities. The number of students is *frequency*, and the data structure appears as *frequency distribution* in a spread sheet.

Table 1. Basic statistical concepts covered from each data set

Data Set	Example observation(s)	Basic statistical concepts covered
I	Time Taken Students $1 \leq t < 10$ 8	Explained above.
II	5, 9, 12, 7, 7	Quantitative data, discrete, simulation, statistical software
III	SO, SR, JR, SR, JR	Qualitative data, ordinal level measurement
IV	70.9, 76.9, 64, 74.2, 73.4	Quantitative data, continuous, sample
V	A, 4-hour ENG 1010,	Multivariate, qualitative versus quantitative, weighted data
VI	(43, 128), (23, 118)	Bivariate data, independent and dependent variable

VISUALIZING DATA IN A MANNER THAT MAKES IT EASY TO READ AND BE HELPFUL TO ANSWERING STATISTICAL QUESTIONS

Going through the steps listed in the textbook is a typical method of teaching data displaying. However, guiding students to find a way to picture data to convey information for solving statistical questions themselves adds a special meaning to the class which inspires their curiosity of learning new things in a natural. A sequence of questions related to statistical thinking is listed in learning visualization of each data set (see Table 2). For example, the questions for "Time Taken" are in the following:

- What questions are you interested in when you look at the data set?
- How many observations are there in the data set? Is it large or small?
- How do we represent the possible values of "Time Taken" in a graph? Is another axis needed to represent the number of students?
- Does the data set look like a step function learned in high school?
- Can you modify "the step function" to be a better plot for picturing the meaning of the data set?
- If the question is interested in the percentage of students for time taken of homework, how do you manipulate the data?

Table 2. Learning data display by answering a list of questions without any rules

Data #	Learning data display by answering a list of questions
I	• Draw a histogram (Figure 1) in R based on the questions answered above
II	• Is it possible to rearrange the data with size of 100,000 in the form of data I by hand? • Who can do it in a second?
III	• Is it possible to rearrange the data in the form of data I with four different categories? • Horizontal axis is not a number line, is it? Why? • Is it better if bars have gaps?
IV	• Is it possible to rearrange the data in the form of data I? How many classes choose to do so? • Is it possible to write down the observations by looking at the histogram of the data? • Is there any plot that the observations can be written down based on the plot? • How to draw a stem-and -leaf plot?
V	• What is an independent or dependent variable in the data set? • Can you draw a point if you know $(\boldsymbol{x}, \boldsymbol{y})$?
VI	• (See on next page)

The "Time Taken" is listed first out of the six data sets because it is the best example for students to find a statistical method to picture the data with existing math knowledge. The concepts of frequency and relative frequency distribution taught in the first session are reinforced in displaying of the distribution. Furthermore, it is helpful for learning other data sets by analogy, while Data set II, III, and IV can be rearranged as the form of data I and then drawn a bar graph or histogram.

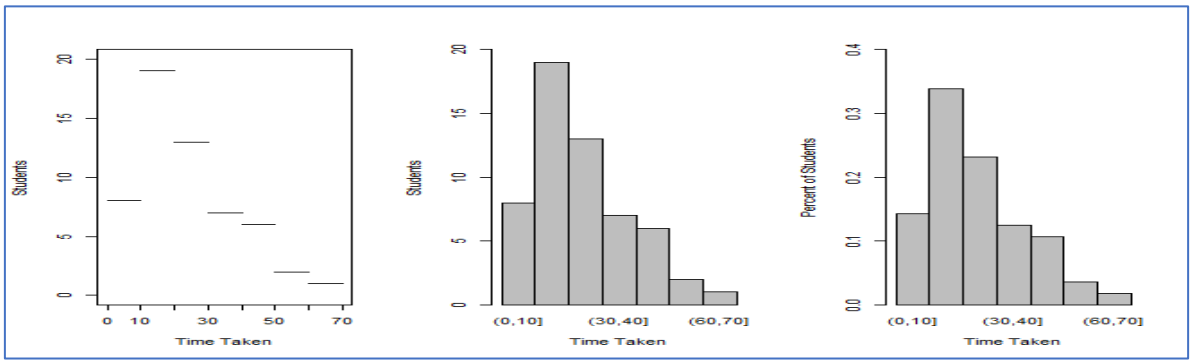

Figure 1. Histogram drawn in R

It is important to highlight displays in which different representations of the same data can convey different information and to emphasize the importance of selecting representations suited to the statistical questions. Data set VI is a multivariate data set. The questions for the multivariate data are listed in the following:

- Are there any two variables that are seemingly related to each other?
- What display may help us to investigate the association between grade and credit hours, or grade and courses?
- Is there any way to visualize three variables in a plane? Can we use the different size of circles to indicate credit hours of classes?

With the help of answering the questions above, the scatter plots (Figure 2) are used to show the association between credit hours and grades, and the bubble plot is introduced for three variables visualization.

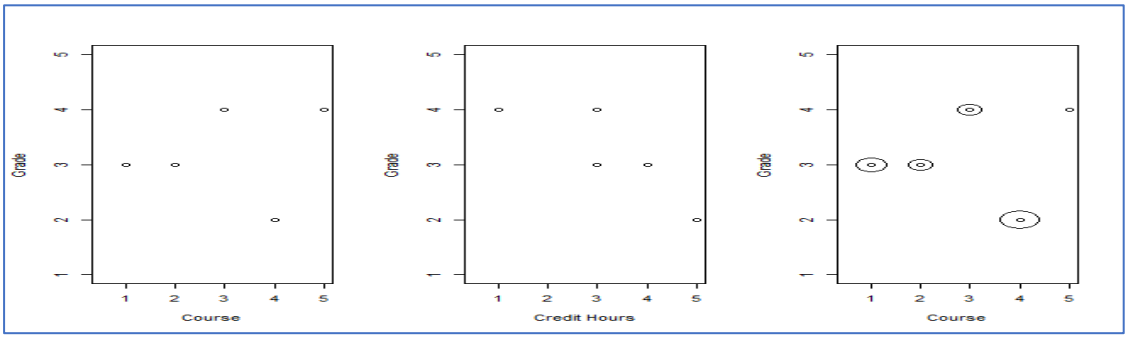

Figure 2. Scatter plots and bubble plot

The displays of the six data sets contain histogram, bar graph, scatterplot, and bubble plot. The shapes of displays of the six data sets contain skewed to left and right, symmetric.

FINDING SUMMARIZED INFORMATION OF DATA THROUGH READING DATA DISTRIBUTION RATHER THAN MATH FORMULAS

To guide students to investigate summarized information of data time taken for homework, the following questions are asked in the discussion:

- What causes the data distribution to have a long tail on the right?

- Since the individual observations are unknown, is there a way to find an approximated value for the Individuals in each class?
- Can we find the mean time using $\frac{\sum x_i}{n}$? If no, can you find an approximated mean?
- Can the formula of grouped data mean be generated based on this example?
- Is it possible to identify median, or the range of the median?
- Compare mean and median, which is larger? Why is it larger? What conclusions can be drawn for the data distribution if it is skewed?

Finding summarized information by reading data distribution and answering the above questions as following:

- Exam 2 covers chapters 10-16. The fact that students who take the online class do the homework at the last minute before the due date causes the long tail to the right.
- Midpoints can be used to approximate the time taken in the classes. Also, the approximated total time taken is half of the total area of the bars (Figure 3). The mean is half the total area of bars divided by total number of students.
- The mean is about 24. The median is between 10 and 20.
- The mean is greater than the median. The mean is affected by the long tail on the right.

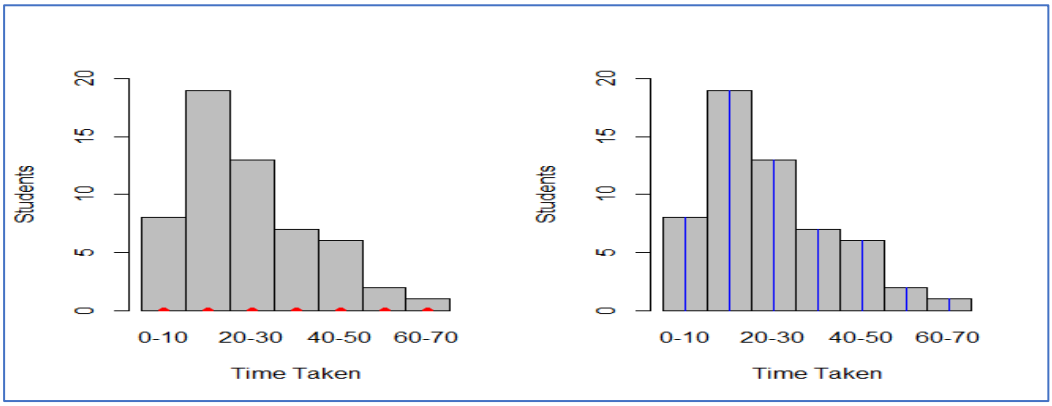

Figure 3. Histograms representing time taken

The concepts of linear correlation can be introduced in processing data IV. With the help of the bubble plot in Figure 2, no students incorrectly use $\frac{\sum x_i}{n}$ to compute mean for the weighted data. The different sizes of circles show that the grades are weighted differently by the credit hours. The weighted mean is computed naturally by average weighted observations.

CONCLUSIONS

Most of my students have expressed their positive and encouraging opinions to the approach of the data-driven learning strategies in this study. Therefore, it's suggested to introduce students with collecting the data sets carefully chosen to include basic statistical concepts and statistical questions first, and then learning statistical concepts, investigating basic statistical methods in the processes of exploring the data distributions and problem-solving skills in teaching introductory statistical classes.

REFERENCES

Mertler, C. A. (2014). *The Data-Driven Classroom: How do I use student data to improve my instruction?(ASCD Arias)*. ASCD.

Sullivan, M., & Verhoosel, J. C. M. (2013). *Statistics: Informed decisions using data*. New York: Pearson.

TeacherVision. https://www.teachervision.com/displaying-data