SOME DIFFICULTIES OF LEARNING HISTOGRAMS IN INTRODUCTORY STATISTICS

Carl Lee
Central Michigan University, MI 48859, USA Carl.Lee@cmich.edu

Maria Meletiou-Mavrotheris
Cyprus Ministry of Education meletiu@spidernet.com.cy

Keywords: statistics education research; graphical representations; histogram; bar graph; variation

Abstract:

The findings reported in the article came from a study where we examined over 160 students' final examination papers, which included questions specifically designed to investigate different levels of understanding about the construction and interpretation of histograms employed to demonstrate the concept of variability. The article describes the four main difficulties in constructing and interpreting histograms identified by the study. It also briefly discusses implications for research and provides suggestions for instructional remedies to help improve students' ability to construct and interpret histograms.

1. INTRODUCTION

Research has suggested that student misconceptions are quite difficult to change (Garfield, and Ahlgren, 1988). Moreover, some of the misconceptions (such as representativeness) seem to vary with problem context (Garfield & delMas, 1990). Student difficulties in learning statistical concepts and in overcoming misconceptions may in part be due to the overlooking of some basic representations of variation and data production (Meletiou & Lee, 2002). The histogram is among the main graphical tools employed in the statistic classroom for assessing the shape and variability of distributions. Introductory statistics courses have been traditionally using the histogram both as a tool for describing data and as a means to aid students in comprehending fundamental concepts such as the sampling distribution.

Because comprehension of histograms is the basis for the concepts of variability and distribution, which are the core concepts of an introductory statistics course, overlooking student difficulties with histograms might have dramatic consequences on the teaching and learning of statistical concepts. In the article, we present findings from a study specifically designed to establish better understanding of the main difficulties students encounter in the construction, interpretation and application of histograms. After providing an overview of the study design, we present findings from the study. Four main categories of mistaken beliefs about histograms identified by the study are discussed. Implications for research and instruction follow.

2. DESIGN OF STUDY

2.1 Motivation for Study

Our past experience was suggesting that understanding of histograms is not as trivial as one might think. For example, in a previous semester, our college-level introductory statistics students had done extremely poorly in the following question given to them at the end of the course (Lee, 2000):

When constructing a histogram for describing the distribution of salary for individuals forty years or older but not yet retired:
a) What goes on the vertical axis?
b) What goes on the horizontal axis?
What would be the proper shape of the salary distribution? Explain why.

Analysis of students' responses suggested that most of them confused the histogram with the scatterplot of salary vs. age, thinking that *'the graph is skewed-to-the right because as people approach retirement, their salary gradually drops'*. This observation came as a surprise. Since histograms appear very frequently in the media and other contexts, one would assume that a college student would have good understanding of this important type of graphical representation. The surprising observation motivated us to conduct the current study, in order to investigate more closely student difficulties in constructing and interpreting histograms.

2.2 Context and participants

The site for the study was an introductory statistics course in a mid-size Midwestern university in the United States. One of the authors, Lee, was the instructor of the course. A total of 162 students participated in this study over a three-semester period starting in the Fall 2001 semester. About 75 percent of

students in these classes were Business majors, while the remaining 25 percent specialized in some other non-science major. The prerequisite for the course was College Algebra, and only few students had taken mathematics courses at the Precalculus level or higher. Approximately 55 percent of students were female. Most students were Sophomores or Juniors. Only a very small percentage of the students were adult learners.

2.3 Instruments, Data Collection and Analysis Procedures

Commonly used research methodologies for investigating student understanding and reasoning are either small scale qualitative interview studies and/or large scale quantitative assessments. In order to dig deep into the process of learning and the reasons behind student responses, qualitative interview methodology has in late years been applied much more frequently than large scale testing methodology. Interview studies are able to investigate a small number of students in depth, and the data resulting from such studies are usually much richer than data provided by large scale quantitative assessments. However, it often takes a huge amount of time to interview students. As a consequence, it is very difficult to conduct a large scale interview study which can be inferred to a general population. The current study attempted to take the advantage of both quantitative and qualitative methodologies by following a mixed-methods approach. We employed both qualitative and quantitative techniques to gather data from correspondents. Linking the depth of qualitative data with quantitative breadth provided us with complementary information and a more holistic picture of students' thought processes.

To investigate student reasoning about histograms, we designed four tasks related to construction, interpretation and application of histograms in real world scenarios. In order for us to be able to identify patterns of student difficulties and misunderstandings, the problems asked students to not only provide answers to several questions, but also to explain the reasoning behind these answers.

The four tasks that formed our assessment are shown in Figure 1. Tasks A and B were included in the first test given to students in the course, which covered topics including descriptive statistics, graphical methods and probability. Tasks C and D were included in the final exam.

Figure 1: Assessment tasks

Task A:

An insurance company is interested in the cholesterol levels of individuals in our community that are 40 years of age or older. A random sample of 100 individuals was chosen from this population and the following information was collected:

Sample size = 100 Average cholesterol level = 158 (mg)
Median cholesterol level = 160 (mg) Standard deviation = 20 (mg)

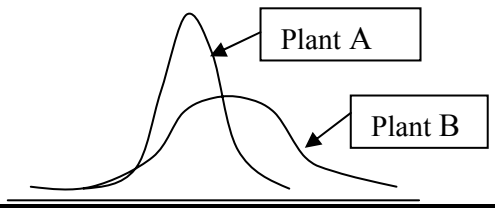
Q1: Based on the information above, the shape of the distribution of cholesterol levels for individuals in the community of age 40 or older is more likely to be _____.
Explain the reason:

Q2: When constructing a histogram for the cholesterol level data
What goes on the horizontal axis?_____
What goes on the vertical axis?_____

Q3: An individual has a cholesterol level of 188 mg. Is this an unusually high cholesterol level? Why ?

Task B:

The following graph shows the distribution of the width of window frames manufactured at two different plants.



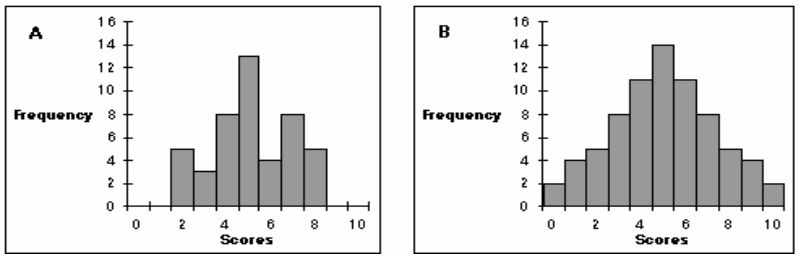
Which of the following statements is correct? (Choose one).

(a) Width distribution of frames manufactured at Plant A is skewed.
(b) Width distribution of frames manufactured at Plant A has a similar variation to the width distribution of frames manufactured at Plant B.
(c) Width distribution of frames manufactured at Plant A has a larger variation than that of the width distribution of frames manufactured at Plant B.
(d) Width distribution of frames manufactured at Plant A has a smaller variation than that of the width distribution of frames manufactured at Plant B.

Explain the reason for your choice:

Task C: (taken from Garfield, delMas, & Chance, 1999)
Which of the following distributions show more variability? (Check one):
(a) A has more variability (b) B has more variability

"Choosing distribution with more variability" task



Now, check the statement or statements that led you to select the choice above:

(a) Because it is bumpier
(b) Because it is more spread out
(c) Because it has a larger number of different scores
(d) Because the values differ more from the center
(e) Other (please explain) _____

Task D:
When constructing a histogram for describing the distribution of salaries for individuals that are 40 or older and have not yet retired,

Q1: Explain:
(a) What goes on the vertical axis?
(b) What goes on the horizontal axis?

Q2: What would the shape of the salary distribution more likely be? Explain why.

Q1 of Task A investigates student understanding of the relationship between graphical and numerical representations, while Q2 their ability to construct a histogram, and Q3 their ability to apply the concept of distribution in a real world situation. Task B tests whether students understand the relationship between the variation and graphical presentation of a distribution. Task C examines several possible fallacies about histograms, including the belief that a 'bumpier' distribution with no 'systematic pattern' has a larger variability (Garfield, del Mas & Chance, 1999) or that a distribution's variation is determined exclusively by its range. Task D is the task that we had given at the end of a previous semester and had unexpectedly found out that most of the students confused the histogram with a scatterplot of salary vs. age. Q1 of Task D is very similar to Q2 of Task A, only situated in a different context.

We conjectured that students would be less likely to confuse the two types of graphical representations early in the semester before being formally introduced to scatterplots. In order to check this conjecture, while Task D was again given to students at the end of the course, Task A was included in the first test.

Since the tasks designed for this study were included in class exams, they were graded as part of students' test scores. A detailed analysis of student responses to the four tasks was conducted separately at a later stage. In addition to performing qualitative data analysis of the written assessments, we also looked at them in purely quantitative terms, drawing conclusions about the performance of the study participants as a whole. Students' solutions were coded with respect to correctness, but also with respect to their attention to numerical and graphical aspects of each problem.

We paid particular emphasis in searching for common mistakes and categorizing these mistakes. Since our focus was on students' reasoning, and, especially, the types and patterns of their mistakes, the effect of important external factors such as instructor, technology, or student ability was less of a concern in this study and was not taken into consideration in our analysis.

3. RESULTS

The purpose of the study to discover the sources of student difficulties in constructing and interpreting histograms with the ultimate goal of using research findings to develop instructional aids for improving student learning. Indeed, we have, through the years, been carrying out classroom-based research (Chance & Garfield, 2002) and using the research findings to modify our instructional approach (Lee 2003; Meletiou & Lee 2002). We have actually noticed that the frequency with which student mistakes occur has been decreasing. Nonetheless, the general categories of wrong reasoning seem to remain the same. In this section, we brief discuss the main categories of mistaken beliefs regarding histograms unveiled by the current study.

Tables A.1, A.2 and A.3 summarize student responses to Q1, Q2 and Q3 of Task A, respectively. Although over 80 percent of the students recognized that the standard deviation plays a critical role in determining the shape of a distribution, almost one-fifth of them had the wrong belief that as long as the mean is smaller than the median, the distribution must be skewed. Five percent of the students, not only thinking deterministically but also confusing the two types of skewness, concluded that the distribution must be skewed to the right because the mean is smaller than the median (see Table A.1). We summarize the type of faulty reasoning revealed through our analysis of students´ responses to this question as follows:

Students often think deterministically when interpreting the distribution of a real world dataset when given the data summary of average, median and standard deviation. Whenever the average differs from the median, regardless how small the difference is relative to standard deviation, the shape of the corresponding histogram is interpreted as being skewed.

Similar to our previous experience, we again found that deciding what to put on the horizontal and what on the vertical axis when constructing a histogram is quite challenging for students. For the majority of students, the histogram is a two-dimensional graph and it must therefore have two variables, one that goes on the horizontal and one that goes on the vertical axis. Consequently, in the specific task of constructing the histogram of the cholesterol level for individuals aged 40 or older, although the condition '40 years or older' only restricts the target population, 36 percent of the students concluded that the histogram should show the relationship between cholesterol level and age.

In Table A.2, the first two wrong answers commonly given by students in our study seem to reflect this wrong belief. Another variable that several students perceived as one of the two variables to appear on the histogram is the individual's case id. A plot of cholesterol level based on the id sequence seemed logical for these students. In summary, two types of mistakes identified when asking students to decide what variables to include in a histogram were:

Students perceive histograms as two-dimensional graphs that must have two variables, and thus tend to interpret a histogram as a two-variable scatterplot.

Students have the tendency to perceive histograms as displays of raw data on Y with each bar standing for an individual observation and individual case or time on X (case or time series plots).

In applying the Empirical Rule, students gave a variety of reasons to support their incorrect answers, several of which suggested a tendency to think deterministically. Reasons in the 'Incorrect reasoning' of the table A.3, are indications of students´ deterministic mindset.

Findings from Task B are summarized in Table B. About 9 percent of the students concluded that the distribution of window frame width from Plant A is larger than that of window frame width from Plant B, giving reasons such as: *'The distribution from Plant B is*

taller', 'Large values on Y axis for Plant A', 'Higher on Y means widths differ more'. These reasons suggest yet another type of misunderstanding:

> *When comparing two histograms with regards to their variability, instead of looking at the horizontal axes of the histograms to compare their spread, students tend to look at their vertical axes and compare differences in the heights of the bars (i.e. differences in frequencies among the different categories).*

Table C summarizes the results of Task C, which investigated students' reasoning about the relationship between variation and the shape of a distribution. As already noted, Task C was similar to Task B but had a different context and was included in the final exam, whereas Task B was included in the first test the students took.

In Task C, we identified similar types of mistaken beliefs to those described in Task B. Students tended to compare values on the vertical axis, and to conclude that the variable which has *'more varied values on Y'*, *'less pattern on Y'* or *'is more random on Y'* has a larger variation. In addition, over 22 percent of the students in our study justified their selection of distribution B as the one having a larger variation by choosing the statement *'Because it has a larger number of different scores'*. This is a wrong justification for a correct answer, suggesting that these students thought of the different groups shown on histograms as being individual scores. Had we not asked students to justify their answer, we would not have been able to discover this flaw in their reasoning, which concurs with findings in Q2 of Task A, of students having the tendency to perceive histograms as displays of raw data with each bar representing an individual observation rather than grouped sets of data. Our findings agree with those reported in Chance, Garfield and delMas (1999) and in Meletiou (2000), who have also found that students often confuse 'bumpiness' of a histogram with 'variability'.

Task D was given in the final exam, after students had been introduced to scatterplots and had applied them in their study of regression and correlation topics. Our conjecture that the recently used scatterplot would have had an impact on students' tendency to confuse histograms with scatterplots does not seem to hold. Results for Q1 of Task D are similar to those for Q2 of Task A. In both questions, students made similar mistakes, although additional types of mistakes were also discovered in the final exam. In particular, 7.9 percent of the students wrote that *'on X goes Age, on Y goes frequency of Salary'*, a response that did not occur in Test 1. This may be due to the emphasis given by the instructor on pointing out that histograms and scatterplots are different graphical representations, and that the vertical axis of a histogram shows the frequency or relative frequency of values falling in the corresponding interval. Although students did seem to take this into consideration, they were still not able to overcome their wrong impression of the histogram as being 'a two-dimensional plot' requiring two variables; one on the horizontal and one on the vertical axis. One would need to conduct an interview study to confirm whether this is indeed the case.

In Q2 of Task D students gave a variety of reasons to justify their conclusions about the shape of the distribution of salaries for individuals who are 40 years of age or older. One major fallacy observed again, was that found in Q2 of Task A, of students confusing histograms with scatterplots. Some of the students completely ignored the fact that what the question was asking them to construct was a 'histogram', and gave answers totally based on the relationship between age and salary. Some of the reasons for concluding that the distribution is symmetric are troublesome, especially the incorrect application of Central Limit Theorem. The reasons students provided for choosing 'skewed-to-left' or 'skewed-to-right' indicate that these students not only confused histograms with scatterplots, but also, more seriously, had little understanding about distributions.

4. CONCLUSIONS

In this study, we attempted to address a concern we experienced previously, that is, of students having poor understanding of one the most commonly used graphical tools, the histogram. We developed four test items specifically designed to investigate students' reasoning about histograms, and analyzed 162 students´ responses to these items. Based on this analysis, we have identified four main types of student difficulties in constructing, interpreting and applying histograms in different real world contexts:

(1) Perceiving histograms as displays of raw data with each bar standing for an individual observation rather than as presenting grouped sets of data.
(2) Tending to interpret histograms as two-variable scatterplots or as time sequence plots.
(3) Tending to look at the vertical axes and compare differences in the heights of the bars when comparing the variation of two histograms.
(4) Tending to think deterministically when interpreting a distribution in real world contexts.

The fact that people often encounter histograms in the media and elsewhere does not necessarily mean that

they understand them. Histograms – as well as bar graphs and other graphs – are a transformation from raw data into an entirely different form. Understanding of this transformation is challenging, and statistics instruction needs to find ways to support it. Such a transformation changes the data representation – a process that Wild & Pfannkuch (1999) have defined as transumeration – and is one of the fundamental frameworks for statistical reasoning, for better understanding of variation, distribution and many other important statistical concepts.

If students have such major difficulties and misunderstandings about histograms, it should not come as a surprise that the majority fail to comprehend challenging concepts such as the sampling distributions. Having good understanding of spread when visually interpreting a distribution displayed in a histogram is necessary to be able to fully grasp the meaning of the concept of sampling distribution. Thus, it is important to reinforce understanding of histograms in the teaching of statistics. As the research literature tells us little about how understanding of histograms and other graphical representations develops, a possible direction of future statistics education research is to find ways to help students recognize the different functions of the horizontal and vertical axes across different graphical representations (Friel, Bright, Frierson, & Kader, 1997). This is essential since, as findings from this study point out, understanding of histograms and their relation to variation is one of the stumbling stones in statistics instruction.

Advances of technology provide us with new tools and opportunities for the teaching of statistical concepts including the use of various graphical representations. These new technological tools are, in fact, designed explicitly to facilitate the visualization of statistical concepts providing an enormous potential for making statistical thinking accessible by all students. Meletiou and Stylianou (2003) developed a course which has at its core element a technological tool, Fathom (Erickson, 2000) and investigated the effects of this technology-based course on students' understanding of graphical representations of data. Preliminary findings suggest improved comprehension of histograms and other graphical representations. Studies on the impact of technology for understanding variation and multiple graphical representations are worthy of pursuing.

Table A.1: Reasoning about the Distribution Shape of Cholesterol Levels using the Relationship Between Measures of Center and Measures of Variation.

Task A, Q1	**(N=162)**	
Answer	*%*	*Reason for selection*
Left-skewed	14.2%	Mean < median
Mounded-shaped	81.8%	Even though mean < median, the difference is very small when compared to s.d.
Right-skewed	5.0%	Mean < median

Table A.2: Common Mistakes when constructing a Histogram.

Task A, Q2	**% Correct (N=162): 34%**
	Most common wrong answers (% of N = 162)
What goes on the horizontal axis? What goes on the vertical axis?	X: Age, Y: Cholesterol level (28%) X: Cholesterol level, Y: Age (8%) X: Individual id, Y: Cholesterol level (22%)

Table A.3: Adequate and Incorrect Reasoning to Decide Whether a Cholesterol Level is Rare or Not.

Task A, Q3	**% Correct (N=162): 70%**
Adequate reasoning (some examples)	
1) 158 + 2(S.D.) = 198 > 188 2) Because of z-score = 1.5 within the 'normal' range of $\bar{x} \pm 2s$ 3) 188 does not fall outside 2 s.d. of mean 4) There is still about 13% of the people with a cholesterol level higher than 188 mg. 5) 188 is within the 95% range or within two s.d. of the mean.	
Incorrect reasoning (some examples)	

1) 188 falls outside of 1 s.d. of the mean. 2) Because the mean is far away from the median 3) Considering that the average is 158, this is relatively high. 4) Because 188 is rare. 5) Everything higher than 160 is too high. 6) Because the average is only 158. 7) 188 is on the right side of the mounded-shape curve. 8) Very far away from the median. 9) 138 to 178 would be considered normal. 188 is high.

Table B: Summary of Results from Task B for Investigating Student's Reasoning about the Relationship Between Variation and Shape of Distribution.

Task B: Relationship Between Variation and Shape of Distribution		
Answers	**%**	**Reasons**
c) Distribution from Plant A has a larger variation than that from Plant B	8.6%	1) Distribution from Plant A is taller, larger variation. 2) Large values on Y axis for Plant A. 3) Higher on Y means widths differ more.
(d) Distribution of width from Plant A has a smaller variation than that from Plant B	88.9%	1) The range of width from Plant B is wider. 2) Distribution of Plant B is more spread out. 3) Distribution of Plant B is flatter. 4) The peak width for Plant B is larger

Table C: Summary of Students' Reasoning about the Relationship between Variation and Distribution

Task C: Which graph has larger variation (N=150)?	**% Correct: 71.3%**
Reasons	**% of reason selected**
(a) Because it is bumpier	1.7% (Chose A)
(b) Because it is more spread out	52% (Chose B)
(c) Because it has a larger number of different scores	22% (Chose B), 6.7%(Chose A)
(d) Because the values differ more from the center	14.7%(Chose B), 9.3%(Chose A)

Table D.1: Common Mistakes in the Construction of the Histogram.

Task D, Q1	**% Correct (N=151) : 43.0%**
Question	Most common wrong answers (% of N =151)
What is on the horizontal axis? What is on the vertical axis?	X: Age, Y: Salary (33.1%), X: Salary, Y: Age (7.3%) X: Age, Y: Frequency of salary (7.9%), X: Individual, Y: Salary (7.3%)

Table D.2: Adequate and Incorrect reasoning for determining the shape of the salary distribution for individuals aged 40 or older who have not yet retired.

Task D , Q2
Incorrect reasoning
Reasons for concluding that the distribution is Skewed-to- right [Wrong reason for the right answer]: 1) People´s salaries will top off at a certain percent. They will not continue to make more and more. 2) Salary tends to increase as age increases. 3) As we get older, fewer and fewer people are working. 4) Fewer people will retire at age 40 and more will retire at an older age. **Reasons for concluding that the distribution is Skewed-to-left:** 1) As you get older, you make more money [most common incorrect reasoning] 2) A lot of employees over 40 years old are paid well.

3) The longer you work, the more you should make.
4) A few make low but most make higher salaries.

Reasons based on the perception of the histogram as a scatterplot:
1) Strong positive correlation, the older the higher the salary.
2) Upward, because younger people won't make much money.
3) Increasing slope, because the older the higher salary.
4) Downwards, because salaries decrease when getting closer to retirement.

REFERENCES

Chance, B., and Garfield, J. (2002). New approaches to gathering data on student learning for research in statistics education. Statistics Education Research Journal, 1(2), 38-41. [Available at http:/fehps.une.edu.au/serj].

Chance, B., Garfield, J., & delMas, B (1999, August). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. Presented at the 52nd Session of the International Statistical Institute, Helsinki, Finland.

Erickson, T. (2000). *Data in Depth. Exploring Mathematics with Fathom*. Emeryville, CA: Key Curriculum Press.

Friel, S.N., Bright, G.W., Frierson, D., & Kader, G.D. (1997). A framework for assessing knowledge and learning in statistics (K-8). In I. Gal and J. B. Garfield (Eds.), The assessment challenge in statistics education (pp. 55-63). Burke, VA: IOS Press.

Garfield, J., and Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics, *Journal for Research in Mathematics Education,* 19, 44-63.

Garfield, J., delMas, B. (1990). Exploring the stability of students' conceptions of probability. In J. Garfield (Eds), *Research Papers from the Third International Conference on Teaching Statistics.* University of Otago, Dunedin, New Zealand.

Garfield, J., delMas, B., & Chance, B.L. (1999). Tools for teaching and assessing statistical inference: Simulation software. Available at htpp://www.gen.umn.edu/faculty_staff/ delmas/ stat_tools/stat_tools_software.htm

Lee, C. (2000). Misconception Vs. missed conception in introductory statistics. Presented at the MAA 2000 Meeting, May, 2000, Mt, Pleasant, MI.

Lee, C. (2003). Using the PACE Strategy to Teach Statistics. To appear in J. Garfield (Eds), MAA Notes, 2003.

Meletiou, M. (2000). *Developing Students' Conceptions of Variation: An Untapped Well in Statistical Reasoning*. Ph.D. Dissertation, University of Texas, Austin, May, 2000.

Meletiou, M. and Lee, C. (2002).Teaching students the stochastic nature of statistical concepts in an introductory statistics course. *Statistical Education Research Journal*, 1(2), 22-37. [Available at http:/fehps.une.edu.au/serj]

Meletiou-Mavrotheris, M., and Stylianou, D. (2003). Graphical Representation of Data: The Effect of the Use of a Dynamical Statistics Technological Tool. To appear in: *Proceedings of the Sixth International Conference on Computer Based Learning in Science*.

Wild, C.J., and Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, Vol. 67(3), 223-265.