

STATISTICS FOR MIDDLE SCHOOL (pupils ages 10-14) TEACHERS

Mike Perry and Gary Kader

INTRODUCTION

The purpose of the paper is to propose *fundamental principles for designing a course* for middle school (pupils ages 10-14) teachers.

The *goal of a course* for middle school teachers is to present statistical concepts and practices underlying the middle school curriculum in statistics/data analysis so that teachers develop an understanding which will enable them to adapt both statistical and pedagogical ideas to a classroom setting.

A *comparison with standard courses* (in the USA) would conclude that this is a course for teachers with emphasis on pedagogy but not a "teaching methods" course, a "content" course but not a "statistical methods" course, a "concepts" course but not a "liberal arts statistics" course.

There is a *different emphasis on topics*. The syllabus outline may look the same as a statistical methods class but the emphasis and instructional approach are different.

A course for middle school teachers requires a *different instructional approach* - the *integration of content and pedagogy*. The traditional notion of content should be expanded by the inclusion of practice. Stressing problem-solving ensures that presentations and instructional activities have the spirit of genuine statistical practice. Ideas should evolve in the context of statistical problem-solving activities, which encourage the development of statistical reasoning skills. Thus the practice of statistics is a means of supporting learning. The idea is endorsed by Cobb (1999): "The challenge as we formulated it was to transcend what Dewey called the dichotomy between process and content by systematically supporting the emergence of key statistical ideas while ensuring that the successive classroom mathematical practices that emerged in the course of the teaching experiment were commensurable with the activities of proficient data analysts. As Biehler and Steinbring noted, an exploratory or investigative orientation is not merely a means of supporting learning but is central to data analysis and constitutes an instructional goal in its own right."

A course for middle school teachers requires a *different focus*. A focus on *curriculum* is required to make a transparent connection between the statistics presented in a course for teachers and the statistics presented in the school classroom. A focus on *pedagogy* is needed so that teachers learn statistics in the same ways that we hope they would practice in their own classrooms; teachers should learn using the same types of activities that their students will experience. A focus on *learners* is required so that teachers understand the diversity of statistical perceptions of students in the school classroom.

CURRICULUM

A focus on *curriculum* should emphasize those statistical concepts and practices which underly the middle school curriculum in statistics/data analysis. This requires considerably more attention to fundamental ideas which in turn implies a reduced emphasis on specialized statistical methodology used in research. For instance, most statistical methods courses give little attention to many of the basic statistical graphs such as dot plots. These are basic tools in the middles

school class and thus a teacher needs experience using elementary graphical representations in both problem-solving and in concept development.

As another example, a statistical methods course might treat "measuring variation" in a rather perfunctory fashion by quickly presenting a few ideas such as standard deviation and interquartile range. A course for teachers should examine the idea of "measuring variation" more closely. The presentation should include different notions of variation and how these might be measured; the standard deviation measures variation about the mean but there are other types of variation with suitable corresponding measures.

As a final example, the treatment of probability in the middle school is predominantly experimental rather than mathematical. The teacher's background should include some combinatorial development of probability; it should, however, emphasize the experimental and simulation approaches which are usually not emphasized in the typical statistical methods course.

PEDAGOGY

A focus on *pedagogy* should employ an activity-based approach. Two types of activities should be employed: activities based on problem-solving and activities based on concept development.

Problem-solving activities should be based on a process composed of the four components: (1) Formulate a question, (2) Collect suitable data, (3) Analyze the data, (4) Interpret results

Problem-solving activities are important because they emphasize the complete cycle in the statistical process. They provide a motivational context for introducing and practicing the application of statistical ideas and techniques. Following is an example of a problem-solving activity from Kader and Perry (1994).

The Hat Shop Problem

Most hats and caps come in "one size fits all", or sometimes in "small, medium, or large." Only the more expensive hats come in sizes and we expect these to fit properly. If you have never worn one of these finer hats then you may not be familiar with the way hats are sized.

A merchant who deals in fine hats for men must decide what styles to keep in stock and how many hats of each size to order. The merchant needs to know which sizes are most common and which occur least often. In other words, the merchant needs to know the distribution of hat sizes.

European hat sizes correspond to head circumference measured in centimeters. Hat sizes in the USA and Great Britain correspond to the diameter in inches. At any rate, to obtain information about hat sizes it is necessary to obtain information about head circumferences.

Questions

Are some head sizes more common than others?
Which sizes are most common and which are the least common?
How often do different hat sizes occur?
What is the distribution of hat sizes?

Data

We obtained a sample of 125 students who were enrolled in introductory statistics classes.

Analysis

The analysis focuses on the distribution of head sizes.

This problem motivates the introduction and practical use of histograms and can in a natural way lead to use of the normal distribution as a model of head circumferences.

Interpretation

A histogram based on class intervals which correspond to hat sizes gives directly estimates of percentages needed to stock for each hat size.

Concept development activities develop statistical concepts underlying the problem-solving process. The activities described by Kader (1999) are based on manipulating dot plots to examine how distributional characteristics relate to the mean and median, the deviations from the mean, and the value of the mean absolute deviation (MAD). The objective of one of the activities is to see how skewing a distribution affects the mean differently than the median. The objective of another activity is to experience and understand the property that deviations from the mean sum to zero. Another activity is aimed at understanding how the MAD responds to changing the degree of variation in the data values.

The activity described in Perry and Kader (1998) is intended to examine patterns in sequences of random outcomes. Students play the game of push-penny by pushing a coin across a lined board. There are different sets of rules for keeping score based on how often the coin lands on a line. Outcomes are recorded and plotted to examine sequences of outcomes. This experimental activity allows students to experience the trial to trial variation produced by randomness and to illustrate three fundamental principles: The Central Statistical Principle; The Law of Large Numbers; and Probability Distribution.

LEARNERS

A focus on *learners* should consider *learner perceptions*. Students' intuitions about statistical concepts vary and affect the way teachers should present ideas to them. Loosen (1985) points out that many students have an intuitive concept of variability which has nothing to do with "how much the values differ from the mean." Their perception has to do with "how often the observations differ from one another." These students are basing their choice on an intuitive concept of variability - *unalikability* – the lack of data points of the same size or the lack of clusters of values of the same size. These learners do not think of variation as *how much the values differ*, rather as *how often they differ*.

This is an important lesson for teachers; we may be talking about one concept of variation while our students are thinking about another! This may result in proposing a measure for one type of variation that is in direct conflict with a student's perceptions. Perry and Kader (2004) pursue this idea further and point out that teachers should be taught in a fashion that emphasizes the concept of variation first and then shows how a particular of measure variation indeed measures what we have in mind or does not measure what we perceive.

The course should emphasize the *evolution of ideas*. In many presentations, definitions (often as formulas) just appear, often without adequate explanation or justification. Teachers, in particular,

need to understand that statistical techniques are inventions. They begin with a need to describe or understand a statistical principle and evolve from initial ideas.

For instance, many introductory level texts begin the presentation of Pearson's correlation coefficient with a definition -and the definition is a rather imposing formula. A few comments are made about the interpretation, the student is asked to calculate a few and the text continues on from there. These presentations never hint that the coefficient was invented. The evolution of the idea began with a need to quantify something and evolved from initial ideas to the formula that appears in the book. Holmes (2003) illustrates how the concept measuring the strength of association with a correlation coefficient evolves to the resulting formula in the text. The idea begins with the notion of positive association being "above average values of Y tend to occur with above average values of X" and "below average values of Y tend to occur with below average values of X". Negative association is described in a similar fashion. This idea is represented graphically in a scatterplot by lines drawn through the (Mean-X, Mean-Y) point, which provides a view of the data in quadrants. Association is described in terms points in the quadrants. A simple measure of association based on the number of points in each quadrant is developed and its shortcomings are pointed out. Pearson's coefficient is then developed as the cure for these shortcomings.

Multiple representations of data play an important role in the statistical problem-solving process. The data analysis/interpretation phase requires a thorough understanding of data representations. Different representations of distributions also affect students' understanding of concepts. One representation may clarify an idea for one student whereas a different representation may be more instructive for another learner.

There are three types of representations-physical, numerical, and graphical. An example of developing the concepts of quartiles for a finite set of distinct measurements on a continuous variable is illustrated in Session 4 of The Statistics, Data Analysis, and Probability course of the Annenberg/CPB (2002) Learning Math project website (www.learner.org/channel/courses/learningmath/data/session4/index.html). The lesson begins with a physical representation of thirteen measurements represented by thirteen spaghetti noodles of varying lengths. The definitions of the quartiles are developed by rearranging and manipulating the noodles. Once developed from the physical objects, the mapping to numerical measurements can be developed and the algorithms demonstrated with the physical objects can be described using the numbers. The final step is to associate both the physical and numerical representations with the graphical representation in terms of a box plot.

It is important for teachers to have a thorough understanding of the inter-relationship between different graphical representations of a data set. For instance, we might use both a histogram and a box plot to examine a set of measurements. These two graphs both use intervals but in different ways; the distinction is a source of confusion for many learners. The histogram uses intervals of fixed length which contain varying proportions of the measurements; the box plot uses intervals of varying lengths which contain the same (at least approximately) proportion of the measurements.

Teachers need *models for data analysis*. Data analysis includes the proper interpretation of results. The interpretation is a form of "argument". Statistical arguments involve reasoning with or about representations of data. Reasoning depends on proper querying of the data representation. The research on reading and querying data representations provides a paradigm for levels of questioning which correspond to levels of reading the data.

The model below is adapted from the work of Fran Curcio (1987) discussed in Friel (1997) and is influenced by the notion of "interrogation of data" discussed by Wild and Pfannkuch (1999). The models provides for three levels for *querying data*.

Direct Queries --reading *within* the data

An explicit and specific question for which the answer comes directly from the data representation.

Derive Queries --reading *between* the data

A question for which the answer depends on interpolating and finding relationships in the data, making comparisons of data values, or applying arithmetic operations to the data.

Inference Queries --reading *beyond* the data

A question whose answer depends on inferring a conclusion from the data. This involves extrapolating, predicting, or generalizing from the data.

Friel (1997) gives the following example.

Students brought several different foods to school for snacks. One snack that lots of them like is raisins. They decided they wanted to find out just how many raisins are in ounce boxes of raisins.. They opened their boxes and counted the number of raisins in each of their boxes. The students were presented with a dot plot showing the information the class found.

The following questions were considered.

“How many boxes of raisins had more than 34 raisins in them? “ This is a *Read between the data* question. The answer requires the answer to the previous question and the answer to four more similar questions. The desired result is the sum of these five frequencies.

“Are there the same number of raisins in each box? This is a *Read between the data* question. The answer requires a comparison. The respondent must note that at least two boxes contain different numbers of raisins. For instance, four boxes contain 34 raisins, four boxes contain 35 raisins.

“If the students opened one more box of raisins, how many raisins might thy expect to find?” The answer requires *Reading beyond the data*. We are making a prediction about an unopened box of raisins.

A Statistics Course for Middle School Teachers

The Statistics, Data Analysis, and Probability course of the Annenberg/CPB (2002) Learning Math project presents a course based on the principles discussed in this paper and contains both problem-solving and concept activities, some of which are in an interactive format. The course can be viewed at the project website (www.learner.org/channel/courses/learningmath/index.html)

REFERENCES

- Annenberg/CPB (2002). *Learning Math: Statistics, Data analysis and Probability*. Corporation for Public Broadcasting: Washington, D.C.
- Cobb, P. (1999). "Individual and Collective Mathematical Development: The Case of Statistical Data Analysis", *Mathematical Thinking and Learning*, **1**(1), 5-43.
- Curcio, Frances R. (1987). "Comprehension of Mathematical Relationships Expressed in Graphs" *Journal for Research in Mathematics Education* 18 (November), 382-93.
- Friel, S.; Bright, G.W.; and Curcio, F. (1997). "Understanding Students' Understanding of Graphs", *Mathematics Teaching in the Middle School* **3**(3).
- Holmes, P. (2001). "Correlation: From Picture to Formula", *Teaching Statistics* **23**(3), 67-70.
- Kader, G. and Perry, M. (1994). "Learning Statistics with Technology", *Mathematics Teaching in the Middle School* **1**(2).
- Kader, G. (1999). "Means and MADs", *Mathematics Teaching in the Middle School* **4**(6), 398-403.
- Kader, G and Perry, M. (1998). "Push Penny -What is Your Expected Score", *Mathematics Teaching in the Middle School* **3**(5).
- Loosen, F.; Lioen, M.; Lacante, M. (1985). "The Standard Deviation: Some Drawbacks of an Intuitive Approach", *Teaching Statistics* **7**(1), 2-5.
- Perry, M. and Kader, G. (2004). "The Coefficient of Unalikability", *Teaching Statistics* **26**(1) .
- Wild, C.J.; Pfannkuch, M. (1999). "Statistical Thinking in Empirical Enquiry", *International Statistical Review* **67**(3), 223-265.