



**Handouts for
Elementary Data
2008-2009**

Recording Sheet for Data Tasks

Use this sheet for notes, as a place to write your questions, and as scratch paper.

Statistical Investigation

By Gemma Mojica & Dr. Katie Mawhinney

In many mathematics classrooms the process of statistical investigation has not received as much attention as it needs because the focus has traditionally been placed on the analysis of data, particularly just on measures of center, and the various methods of displaying data. There is a tendency to spend the most time on the mechanics of statistics, such as how a mean is calculated or how a box plot is made. One possible explanation for this may be that historically statistics has been seen and taught as a subtopic within mathematics, instead of a separate field in its own right. Even with the inclusion of data analysis and probability by the National Council of Teachers of Mathematics (NCTM) in the *Principles and Standards for School Mathematics* (NCTM, 2000) as a content standard, the organization has looked to statistics and mathematics education experts to further define data analysis and probability in the K-12 classroom (Franklin, Kader, Mewborn, Moreno, Peck, Perry, & Scheaffer, 2005).

The purpose of this article is to highlight vital topics in data analysis and statistics that may have previously been neglected or glossed-over in the interest of focusing on the mechanics of statistics. Beginning with the PCAI Model of Statistical Investigation (Graham, 1987), teachers are provided a framework for engaging students in statistical investigation that delves deeper into other important aspects of statistical investigation, essential for a student's understanding of data analysis and statistics. The four components of the PCAI model may emerge linearly, as listed here, or may include revisiting and making connections among the components. An outline of the model follows:

- 1.) Posing a question.** Selecting a question that is...
 - motivated by describing summarizing, comparing, and generalizing data within a *context*.
 - measurable—variables (numerical or categorical) should be able to be measured.
 - based on data available within the time frame of the investigation.

- 2.) Collecting the data.** Determining...
 - methods of collecting data.
 - the population.
 - if a sample will be collected.
 - the type of sample (i.e. random, convenience, census, etc.).
 - whether a sample is representative or biased.
 - sample size.
 - if more than one sample will be collected, or if class data could be pooled to increase sample size.

- 3.) Analyzing data.** Describe and summarize data...
 - using relevant summary statistics, such as the mean, median, mode.
 - using tables, diagrams, graphs, or other representations.
 - by describing variation.

- 4.) Interpreting the results.**

- Relate analysis to original question and context.
- Make decisions about the question posed within the context of the problem based on data collection and analysis.

Below are some examples that may help to clarify the components of the model.

1.) Posing a question

This initial step is what drives the real world use of statistics and provides the context for the statistical investigation. Selecting a clear, measurable question motivates the investigation process and influences all the remaining steps in the investigation process. For example, if seventh graders focus on issues relating to weather, one might ask, "How much rainfall will our town receive next month?" Considering this question as the motivator for data collection, a student may notice that different places within the same town receive different amounts of rainfall at any given time, so that a better question might be, "How much rainfall will our school, Middle Road Middle, receive next month?" Then the planning for a structured method of data collection can begin.

2.) Collecting the data

When a posed question relates to a particular group, students should be able to identify the population under investigation and determine when a census (polling the entire population) or a sample (a subset of the population) should be taken. For example, if the question is, "What is the average height of a student in Ms. Jones' first period class?" a census may be possible and reasonable. Where as answering the question, "What is the average height of a student in Middle Road Middle School?" may be best approached with surveying a subset of the population if the student population of the school is large. Once the decision is made to sample the population, a student must make careful plans to collect data, in order to understand the results of the investigation. Will students be chosen at random to be measured? If a random sample is desired, how will students be chosen (i.e., how will they be assigned a number so that a random number generator may be used)? If a convenience sample is chosen, how would a sample of all eighth graders versus all sixth graders affect the results? Should more than one sample be taken? Such issues of data collection can motivate the student to return to and possibly revise their initial question, to make collecting data a more reasonable task. If students plan to select a sample, they should also consider whether their sample will be representative of the population from which it is drawn, or whether the sample will be biased.

3.) Analyzing data

Once data is collected, there are numerous characteristics to consider that can be used to describe the distribution of data. Data can be described by measures of center, such as the mean, median, mode, and midrange. Data can also be described by spread, using the range, quartiles, and inter-quartile range. Attending to variability is also a useful way to describe data. Students can attend to variability by considering the minimum and maximum values, the range, the inter-quartile range, and unusual data points, like outliers. Variability will be further discussed within this article.

Though measures of center are often useful, particularly for questions like, "What is the average height of a student in Ms. Jones' first period class?" there are many reasons for considering other aspects of data and the relationships between these characteristics. For example, it would be

important to note that Tim Thumb, Tom Thumb's shorter first cousin, is a student in Ms. Jones' first period class. Since his height is an outlier of the data set of heights of students, a more accurate representation for the mean height could be the one calculated without his measurement. Recognizing Tim's height as an outlier requires looking at the distribution of the data set and understanding that most heights fall between 3 feet and 6 feet, within a range of 3 feet. Further, there are many possible situations where reporting the mean or even the median value of a data set is not appropriate for answering the initially posed question.

Alongside this discussion of different data characteristics, students should consider the most or more appropriate method of displaying their data. Different displays can highlight different data characteristics, while masking others. If the student studying heights in Ms. Jones' first period doesn't want to completely exclude Tim's smaller stature, a histogram or dot plot may be the chosen method of display to show that most of the heights were in fact, distributed between 3 and 6 feet though the mean was lower than expected. Thus, the mechanics of creating the possible data displays is needed for the larger task of decision-making within the process of statistical investigation, instead of being the main focus of classroom instruction.

As mentioned above, variability is an important feature of data. The GAISE Report highlights the importance of considering variability within the process of statistical investigation:

Statistical thinking, in large part, must deal with this omnipresence of variability; statistical problem solving and decision making depend on understanding, explaining and quantifying the variability in the data. It is this focus, variability in data, that sets statistics apart from mathematics (Franklin et al., 2005, p. 5).

What is variability? Both teachers and students have experiences with variability, though for teachers, a deeper understanding of variability within statistics is needed. Franklin and Garfield (2006) describe different types of variability—

Measurement variability

- Measurement variability occurs because repeated measurements can be different.
Example: measuring an individual's blood pressure daily over a 30 day period
- Variability can also be a result of measurement error.
Example: using a small ruler to measure a large distance (accuracy & precision)

Natural variability

- Variability is innate in nature and is due to differences in individuals.
Example: women's heights—all of the women in the room are not the same height, or men's average height versus women's average height

Induced variability

- Variability is the result of factors not related to natural variability.
- Variability may be due to the design of an experiment.
Example: Suppose a class of students wants to study the growth of a particular plant or flower. The plants or flowers from the same variety will grow to different heights. Location may be a factor that influences the height of the plants or flowers.

Sampling variability

- When samples are randomly selected from a population, statistics vary from sample to sample

Example: Polling a group of voters to find out which candidate they support would likely yield different results from a different sample of the same size.

Though the students are not yet required to differentiate between these different types of variation, they are capable of understanding why such factors can influence data. Discussions of variability in the classroom should remain informal, but should be present when appropriate as a piece of the statistical investigation process.

4.) Interpreting the results.

This final step listed in the PCAI Model may seem repetitive at first glance. Interpreting why the data characteristics calculated are what they are may occur alongside the analysis. However, as a teacher planning a lesson may refer back to the main goal of the lesson, a statistical investigator should look back to the original context that motivated the question, determine if the best method of data collection was chosen, determine if the best method of analysis was chosen including the most important data characteristics highlighted, and finally whether or not the question was well-posed and answered.

The PCAI Model for statistical investigation may seem to be a very involved process for data analysis and statistics in the middle grades. However, in North Carolina's revised Curriculum Standards, K -5 teachers are encouraged to focus on the components of the model through the competencies in the Data Analysis and Probability Strand. Attention to variability is also part of the K-5 Standard Course of Study in North Carolina within this strand. This discussion is offered here as knowledge for the teachers of data analysis and statistics in the middle grades, not as something to be copied precisely and handed out to classes. Ideally, the model can help teachers provide for middle grades students a means by which their experiences with statistics can move beyond the mechanics of statistics to a better conceptual understanding of data they are and will be exposed to in their everyday lives and careers, as citizens of the 21st century. Perhaps, after giving some thought to such a model, a teacher might be able to consider the question, "What is the average height of the students in my first period class?", and create a deep and meaningful lesson that engages students in statistical investigation. Once the teacher has guided the students through the mechanics they need to investigate this question, the teacher could then ask students to consider all the different data characteristics and ask them to decide which are important and why. This teacher could also have students create various displays of the data and ask the students to decide which are more appropriate and why. And finally, the teacher could have more than one team of students collect all of the heights of students in the class and facilitate a discussion about why the different teams may have different results, building on a natural understanding of variability and beginning to formalize the idea within a statistical investigation.

References

- Franklin, C., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2005). *A curriculum framework for K-12 statistics education*. GAISE report. American Statistical Association. Retrieved September 3, 2006 from www.amstat.org/education/gaise
- Franklin, C. A. & Garfield, J. B. (2006). The GAISE project: Developing statistics education guidelines for grades pre-k-12 and college courses. In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook* (pp. 345-375). Reston, VA: National Council of Teachers of Mathematics.
- Graham, A. (1987). *Statistical investigations in the secondary school*. Cambridge, UK: Cambridge University Press.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

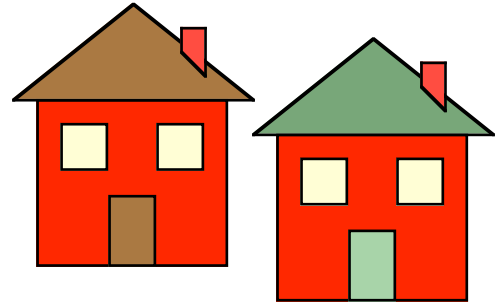
Data for Number of Roofs Lived Under

Set 1:

3, 3, 4, 4, 4, 5, 6, 6, 6, 7, 7, 7, 11, 11, 12, 12, 12, 12,
13, 13, 13, 13, 18, 18, 18, 19, 19, 19, 20, 20

Set 2:

2, 3, 4, 5, 5, 6, 6, 7, 7, 7, 9, 9, 9, 10, 10, 10, 10, 10,
11, 11, 11, 12, 12, 12, 12, 13, 13, 13, 14, 14, 14, 15, 19, 21, 43



1. Make line plot displays for the two sets of data.

2. What kinds of summary statements might we make about the two sets of data using only the median?

