STATISTICS TEACHER
SUPPORTING THE TEACHING AND LEARNING OF STATISTICS
ASA NCTM
STatistics Education Web STEW
online journal of K-12 statistics lesson plans

THE MEAN AND VARIABILITY FROM THE MEAN

Christine Franklin (ASA, University of Georgia), Gary Kader (Appalachian State University), Tim Jacobbe (Southern Methodist University), and Kaycie Maddox (Northeast Georgia RESA)

Overview of Lesson

The Pre-K-12 GAISE report (Franklin, et al. 2007) emphasizes that students should be able to do more than just calculate summary statistics; they need conceptual understanding of those statistics that deepens as they progress through different levels of statistical literacy (Levels A, B, and C). The purpose of this activity is to develop students' conceptual understanding of numerical summaries used to describe the center and spread of a distribution of quantitative data.

In this lesson, students investigate how to interpret the mean at Level A (fair share value) and then at Level B (the balance point of a distribution). The students will also explore how to describe and interpret the variability in data from the mean at Levels A (using the number of steps from fair) and at Level B (using the MAD). What students discover in this lesson provides a foundation in their conceptual thinking for future use of the standard deviation in Level C (high school).

Type of Data

- One quantitative variable
- Static dataset provided by lesson plan authors

Learning Objectives

- **6.SP.A.2:** Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, spread, and overall shape.
- **6.SP.A.3:** Recognize that a measure of center for a numerical data set summarizes all of its values with a single number, while a measure of variation describes how its values vary with a single number.
- **6.SP.B.5.C:** Summarize numerical datasets by giving quantitative measures of center (median and/or mean) and variability (interquartile range and/or mean absolute deviation), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered.
  - Interpret the mean as the ***fair share*** value and as the ***balance point*** of a distribution
  - Quantify variability from the fair share value by using the ***number of steps to fair.***
  - Quantify variability from the balance point by using the ***mean absolute deviation (MAD)*** – the average distance of observations in a distribution from the mean

Audience

- This lesson has been tested with middle and high school students, pre-service and in-service teachers, and mathematics teacher educators.
- This lesson is also appropriate for upper elementary grade students.
- *Prerequisites:* Prior to this lesson, students should have experience identifying statistical questions, constructing graphical displays of quantitative data, and calculating the mean.

Time Required

- Two 50-75 minute periods

Technology and Other Materials

- Student Handout (one per student)
- Snap cubes
- Poster boards (with horizontal axes labeled ahead of time by the teacher)
- Post-it™ notes

Lesson Plan

The mean is a numerical summary for quantitative data used as a measure of quantifying the center of a distribution. The mean is calculated by adding up all the observations in a data set and dividing by the number of observations. But what does the mean tell us about the distribution of a quantitative variable and how can students interpret the mean at the different levels of their schooling—at elementary and then as they progress to middle school? In data analysis, it is at least as important to quantify the variability of a distribution as it is to quantify the center of the distribution. How are students expected to describe the variability in a distribution from its mean?

The lesson plan at Level A consists of two main parts. Students will be divided into groups for each part. For part 1, students will complete an investigation to develop conceptual understanding of the mean as the fair (or equal) value. For part 2, students will complete an investigation to develop conceptual understanding of variability as the fair share value, measured as the number of steps to fair. The handout at Level A will be completed within the groups. The class will come together at the end of Level A to summarize their findings.

The lesson plan at Level B consists of two main parts. Students will be divided into groups for each part. For part 1, students will complete an investigation to develop conceptual understanding of the mean as the balance point. For part 2, students will complete an investigation to develop conceptual understanding of variability from the balance point, measured first as the Sum of the Absolute Deviations (SAD) then as the Mean Absolute Deviation (MAD). The handout at Level B will be completed within the groups. This part of the lesson also prompts students to create their own distributions with a given balance point. The class will come together at the end of Level B to arrange the distributions created by each group with respect to variability from the mean (balance point).

What students discover will allow them to evolve in their conceptual thinking to utilizing the mean and standard deviation in practice once they progress to Level C (high school).

The focus of this activity is on developing students' conceptual understanding of numerical summaries used to analyze discrete, quantitative data that were collected to answer a statistical question. This is a GAISE Level A/B activity.

Introducing the Mean

One of the most commonly used numerical summaries used in statistical analysis is the mean. When we ask someone, "What do you think of when you are asked to describe the field of

statistics?" a typical response is "the mean." Before moving into the investigation, allow students to begin thinking about the idea of a 'mean.' Is this a term they have heard of or studied before and if so, what is their understanding of what the mean represents?

1. What is the mean? Students will give a variety of answers, most often stating the formula for finding the mean, saying the mean is the average, or the mean is a measure of center.

Let's consider question 1 in two parts:

2. How do you calculate the mean? Here, we expect students to provide the formula.

3. How do you interpret the mean? Students will struggle more with answering this question. We will investigate to discover the best way to interpret the mean.

Level A Investigation – Fair Share and "Steps to Fair"

Finding the Fair Share Value

A local school is interested in knowing how many people live in the household of each student. A statistical question they might ask is:

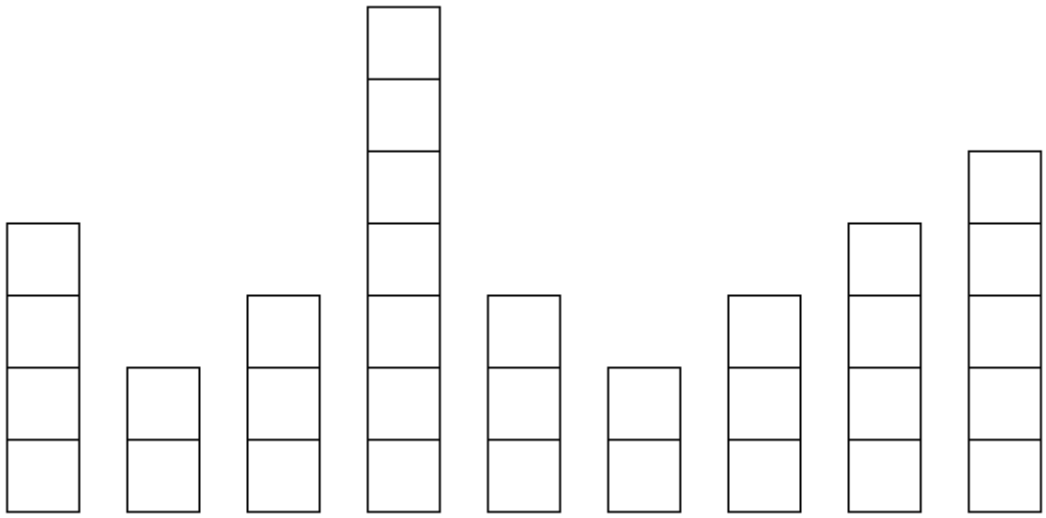
How do the number of people in a student's household at this local school vary?

4. What are three characteristics of this question that lead to classifying the question as a statistical question? Students are expected to name the population (households of students at the local school), the variable to be measured (number of people in a household), and to recognize the data collected on the number of people in a household will vary.

Understanding the time and effort involved with taking a census of all student households at the school, the principal decides to take a sample. As a pilot sample, nine children were selected to find out more about the size of their households. Each child was asked, "How many people are in the household where you live for most of year?" This is called the survey question, and it is used to help answer the statistical or investigation question posed earlier. Each child represents his/her family size with a collection of snap cubes.

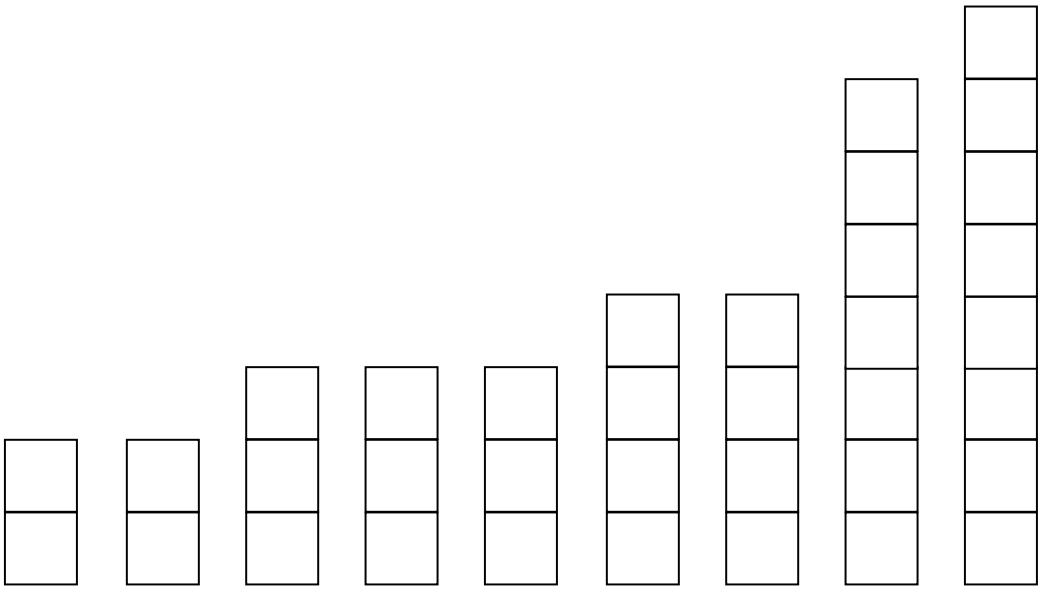
5. How many people are in the household where you live for most of the year? Represent your family size with snap cubes.

   Students in each group will create a snap cube stack to represent their family and then the group of students will create the distribution of the family sizes as represented by the different student snap cube stacks. For example, if we ask 9 students "how many people are in your family," the data for "family size" might be represented with snap cubes as shown below. The first set of four snap cubes represents a household with four family members, the second set of two snap cubes represents a household with two family members, etc.



6. How might we examine the data as represented by the snap cube stacks for these nine children? We expect students will first arrange the stacks in increasing order. We also want the students to recognize that the family sizes vary.

   If we put the original 9 stacks in order, we have the snap cube representation that follows below.

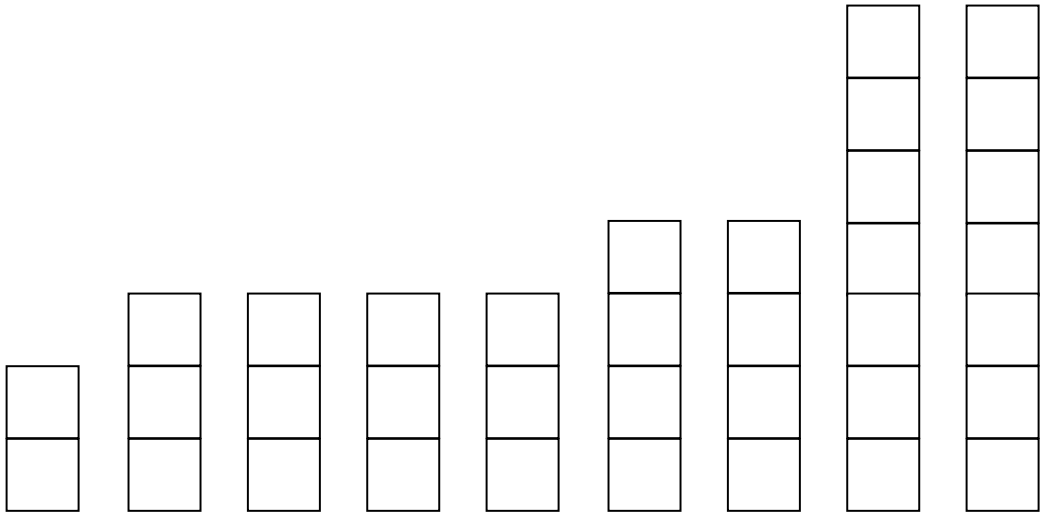


7. If all nine family sizes had been the same, there would be no variability. What if we used all our family members and tried to make all families the same size? How many people would be in each family? Here we anticipate students may use two approaches:

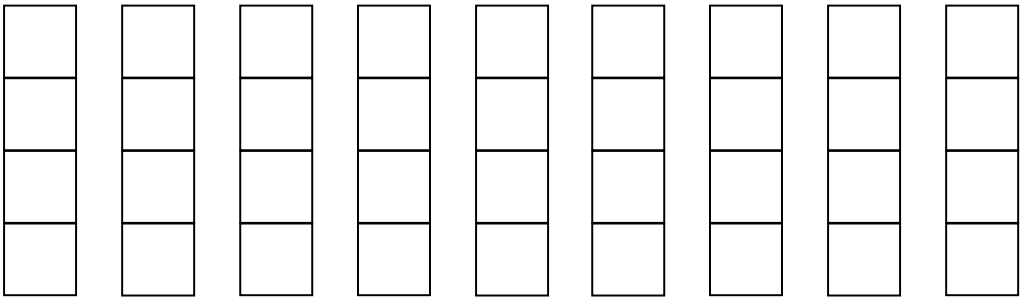
   (1) Disconnect all the snap cubes and redistribute them one at a time to the 9 students until all snap cubes have been allocated. In this case, there are 36 snap cubes. Redistributing them among the 9 children yields 9 stacks with 4 snap cubes each.

   (2) Removing one snap cube from the highest stack and placing it on one of the lowest stacks and continuing until the stacks are leveled out. Both methods yield a "fair share" family size of 4 which is the mean.

Removing a snap cube from the highest stack and placing it on one of the lowest stacks yields:



Continue this process until all the stacks are level, or nearly level when there is a remainder.



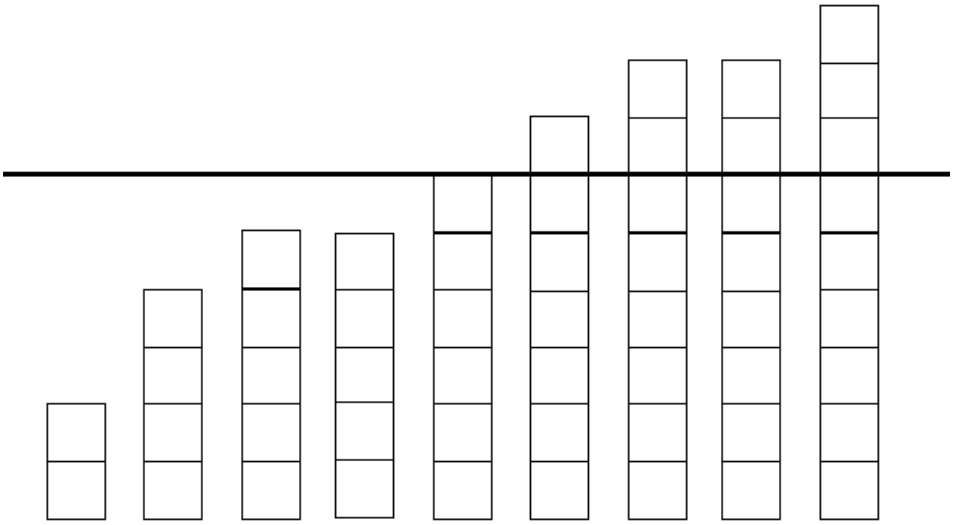
After the final move, all 9 stacks are level, revealing the fair-share family size to be 4. That is, if all 9 family sizes were the same, the number of people in the household would be 4. Note that the fair share value of 4 is the mean of these 9 family sizes and that there is no variability in these 9 family sizes.

Same Fair Share Value, Different Datasets

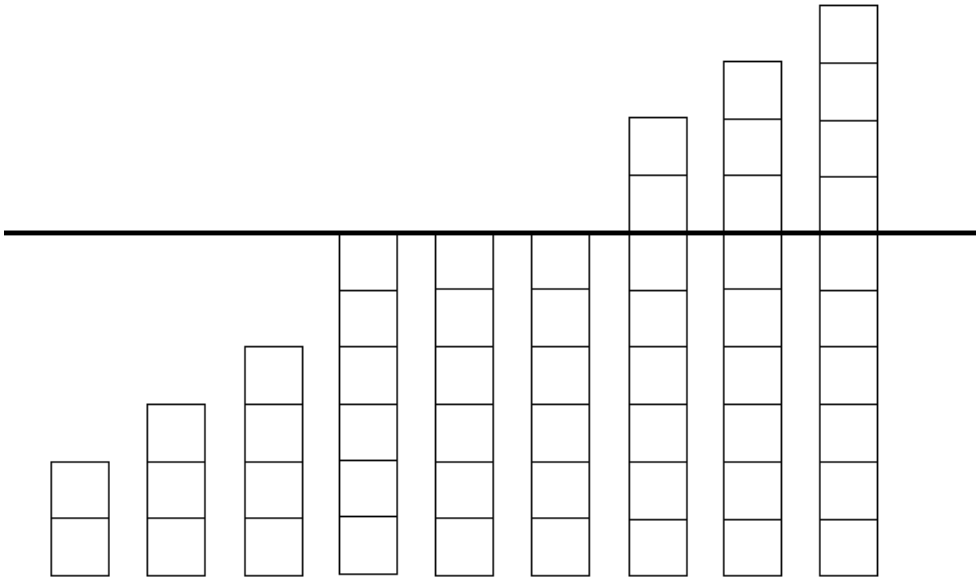
What if the fair-share value for nine children is 6? What are some different snap cube representations that produce a fair share value of 6? Let's consider two different distributions of family size where the fair-share value is 6.

For example, consider the following two groups of data on family size. Note that there are nine family sizes in each group. Also, the fair-share family size for each group is 6. Since the fair-share value for each group is 6, we cannot distinguish the two groups based on the fair-share value alone. A question we might ask is: *Which group is **closer** to being fair?*

Group 1



Group 2



8. How might we decide "how close" a group of family sizes is to being fair? We expect students may take one of two approaches: (1) To choose Group 2, since this group has the highest frequency of stacks of 6 snap cubes or (2) To choose Group 1, since this group has fewer snap cubes to move to level out all the stacks to the fair-share value of 6.

The first method is based upon which distribution has the most stacks that are the fair share value. The shortcoming of doing this method is it does not consider all of the observations in the distribution. The second method provides a numerical summary that does consider all the snap cube stacks of family sizes. This method of having fewer snap cubes to move can be thought of as counting the number of 'steps to fair' in moving snap cubes required to make the group fair. Fewer steps indicates closer to being fair and less variability. For Group 1, the number of steps is 8. For Group 2, the number of steps is 9. Group 1 is closer to fair and has less variability than Group 2.

9. One of the two groups is symmetric. Is it Group 1 or Group 2? Explain your choice. It is Group 2. From geometry, we see that we have rotation symmetry in that stacks of cubes above the leveling line at 6 can be rotated into the stacks below the leveling line at 6. This is not the case for Group 1.

Let's create other distributions of family sizes for nine children where the fair-share value is 6. You as the teacher will provide the students with snap cubes and certain conditions that must be met when creating the stacks of snap cubes for the 9 children. Each group is assigned one set of conditions.

Here are possible conditions to assign:

- Exactly one six and exactly two nines and exactly one three
- Exactly one ten and exactly two fours and no sixes
- More family sizes less than six than greater than six, no sixes.
- Not Symmetric, no sixes
- More family sizes greater than six than less than six, no sixes
- Not symmetric, exactly one six
- Exactly one five, one seven, and one six, not symmetric

After all the new distributions are created, determine the numbers of steps to fair for each group. Bring all the students together and let the students arrange the distributions in order based upon the least amount of variability from the fair share value to the greatest amount of variability from the fair share. The goal is that the students will utilize the number of steps to fair.

10. For each of the sampled distributions, interpret the results to answer the question, "How do the number of people in a student's household at this local school vary?" Students should comment that if all the families were the same size, the number of people in a household would be 6, which would be the fair-share or mean value. However, some groups have family sizes closer to fair than others. The students can then comment on those groups.

Summary of Level A

Students completing the first part of this lesson that investigates the mean and variability from the mean at Level A should understand:

- The notion of "fair share" for a set of numerical or quantitative data
- The fair-share value is the mean value
- The algorithm for determining the mean value
- The notion of "number of steps" to make fair as a measure of variability about the mean

Level B Investigation – The Balance Point and the Mean Absolute Deviation (MAD)

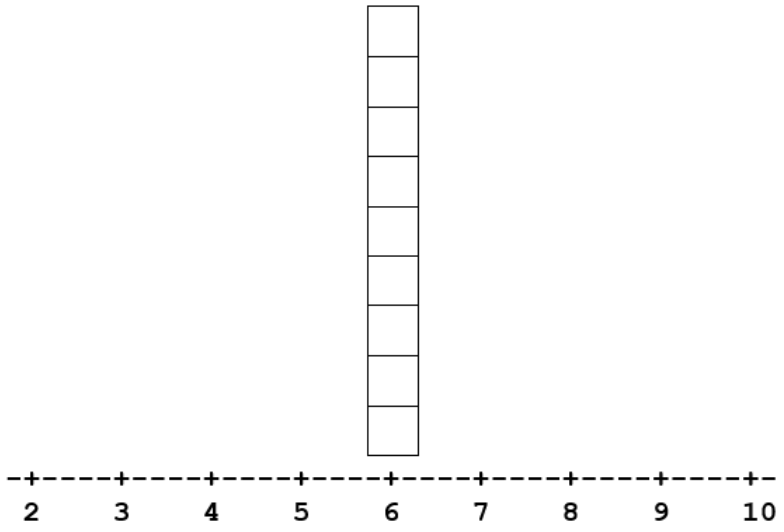
Students have learned how to use snap-cube representations for numerical data at Level A. They have also learned that the mean for a set of numerical data can be interpreted as the "fair-share" value and that a measure of variability in the data from the fair-share value is the "number of steps" required to make a snap cube representation "fair."

At Level B, students learn an alternative interpretation of the mean and another way to quantify the degree of variability from the mean in the data. These two ideas will be developed using the dotplot representation for data.

A local school is interested in knowing how many people live in the household of each student. We will use the same statistical question considered at Level A:

How do the number of people in a student's household at this local school vary?

As in Level A, a pilot sample of nine children were selected to find out more about the size of their households. Each child was asked, "How many people are in the household where you live for most of year?" Instead of using a physical snap cube representation, a more abstract representation is used where each child represents his/her family size with a Post-it™ note on a dotplot. Thus, a stack of 6 snap cubes is represented with one Post-it™ note at 6 on the dotplot. Suppose among this group of 9 children, there were 54 people in the households. Here is a possible representation of the distribution of family sizes for the nine children:



Each of the nine students had a family size of 6. Note that the fair-share value, or mean family size, for these data is clearly 6.

Same Fair Share Value, Different Datasets

At Level B, we want students to develop a second interpretation for the mean as the balance point of the distribution. To discover this interpretation, students can investigate what other distributions of the family size might look like for nine children.

11. Working in groups, create a new dotplot on poster board representing the distribution of nine families with a mean family size of 6. Use 9 Post-it™ notes to represent each family size. For this investigation, we will make a rule that no family size can be below 2, and no family size can be above 10. Each group is assigned one set of conditions to work with. The teacher might consider some of these conditions:

    - Exactly one six and exactly two nines and exactly one three
    - Exactly one ten and exactly two fours and no sixes
    - More family sizes less than six than greater than six, no sixes.
    - Not Symmetric, no sixes
    - More family sizes greater than six than less than six, no sixes
    - Not symmetric, exactly one six
    - Exactly one five, one seven, and one six, not symmetric
    - ***Symmetric, exactly two sixes, exactly one eight – this one can't be done – see if group realizes this fact

NOTE: To encourage students to form the distributions by using the balance point approach, provide the groups of students with a poster board that has the 9 Post-it™ notes placed at 6 instead of a blank poster board with only the horizontal axis drawn.

Each group should explain how they arrived at their distribution.

Students will take different approaches. If the students take a conceptual approach, the students will build the distribution while keeping in mind that the distribution needs to balance on each side of the mean value at 6. That is, the total distance from the mean for all the observations on one side of the mean must be the same as the total distance from the mean for all the observations on the other side of the mean. Other students may fall back on the familiar algorithm for determining the mean recognizing that the total of the 9 family sizes has to add up to 54 – so building the distribution will be by trial and error. Have each of the groups explain their reasoning to the class.

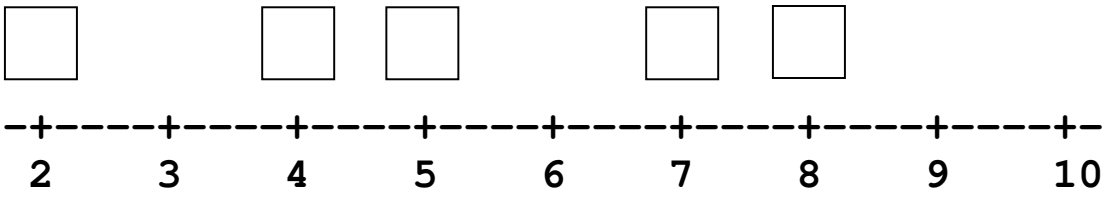
12. After all the groups have explained how they created their distributions, observe that each of the distributions has the same mean but different variability. As a class, arrange the distributions represented on the poster boards from the least amount of variability to the greatest amount of variability based on a visual inspection. One student from each group will display the group's distribution and the class will move their classmates until the class is satisfied with the ordering of smallest to largest variability in the distribution.

    Students will typically try to arrange initially by visually examining the overall distributions and what intuitively looks like least to most variability. Some students will

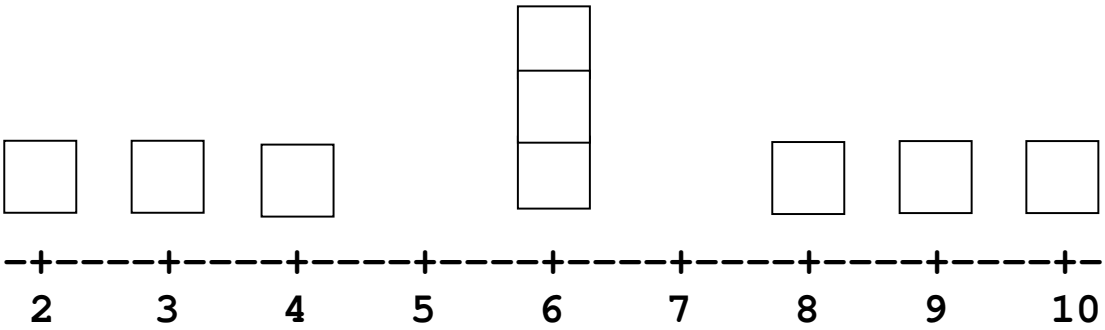
start to remember that at Level A, they used number of steps to fair and then utilize this by quantifying with a summary number the variability from the mean for each of the distributions. They will then begin counting how many units (the distance) an observation falls from the balance point for each distribution displayed. They can then add up the number of units for each distribution. These totals then determine how the distributions should be ordered from smallest to largest.

Two possible distributions for nine family sizes with a mean value of 6 are shown below.

Group 1



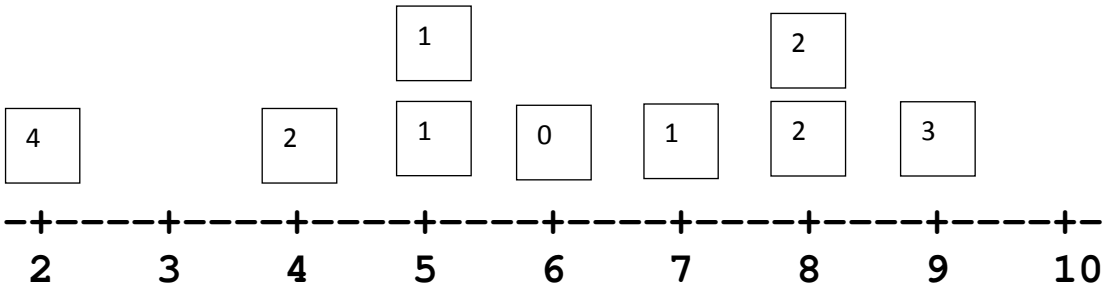
Group 2



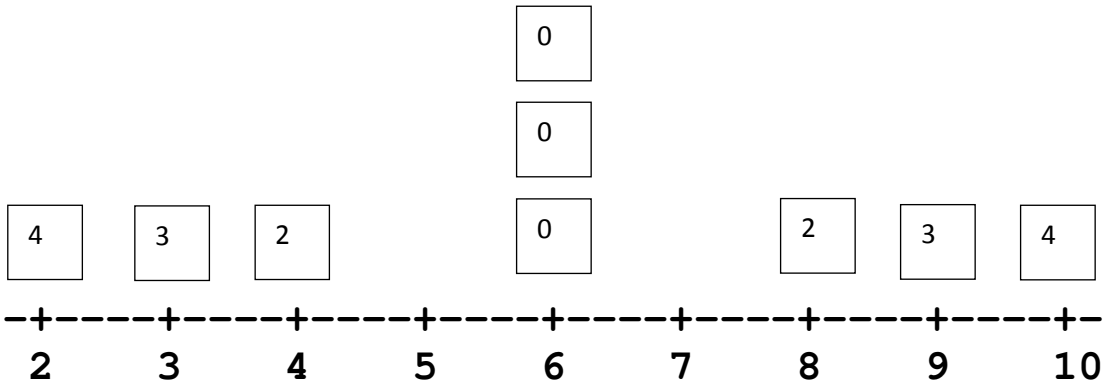
13. How do these two distributions compare? What is similar? What is different? We see that each distribution has the same mean of 6, thus the distributions can't be distinguished by reporting the mean. However, the amount of variability from the mean appears different.

14. How might we quantify the amount of variability from the mean value of 6? One way to quantify the amount of variability from the mean is to examine "how far" each data value is from the mean. Many students will recall their Level A knowledge of counting the number of steps to fair. They will count the number of units they must move a Post-it™ note to place the Post-it™ note at 6. Encourage students to write the number of units to move on the Post-it™ note.

Group 1



Group 2



What the students are calling the "number of steps", we can think of as the unit distance each individual data value is from the mean.

We are interested in which group has more variability from the mean "overall". One indicator is the Sum of these individual distances. For Group 1, the sum is 16; for Group 2 the sum is 18. Thus, there is more variability (from the mean) for the data in Group 2 than there is in Group 1.

Developing an Algorithm

The above Sum is determined by adding the distances for the individual data values from the mean. These distances are determined by first finding the deviation from the mean for each data value:

Deviation from the Mean = Value – Mean

The distance each value is from the mean is the absolute value of its deviation. That is,

Distance from the Mean = |Value – Mean|

The Sum of the Absolute Deviations provides an indication of how much the data vary from the mean. That is,

SAD = Sum of the Absolute Deviations = Sum[|Value – Mean|]

provides a measure of how much a group of data vary from the mean. The larger the SAD, the more the data vary from the mean.

Note that the data displayed in these two dotplots are the same data illustrated with snap cubes (Groups 1 and 2) in Level A. Recall that the "Number of Steps" to Fair Share is used as a measure of variability from fair share at Level A. For Group 1, the Number of Steps was 8; for Group 2, the Number of Steps was 9. The SAD for Group 1 is 16 = 2(8) and the SAD for Group 2 is 18 = 2(9). It can be shown that in general, the

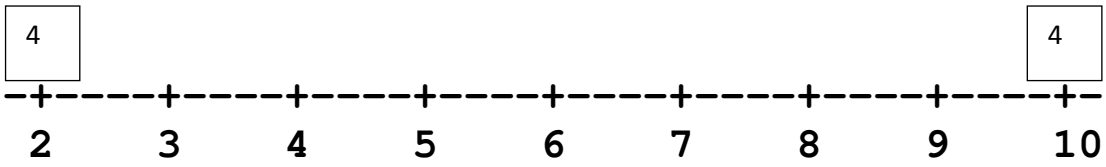
SAD = 2*(Number of Steps to Fair Share)

Also observe that the total of the distances for the values below the mean is the same as the total of the distances for the values above the mean. (In Group 1, this total distance on each side of the mean is 8; in Group 2 this total distance on each side of the mean is 9). For this reason, the mean is the ***balance point*** of the dotplot distribution.

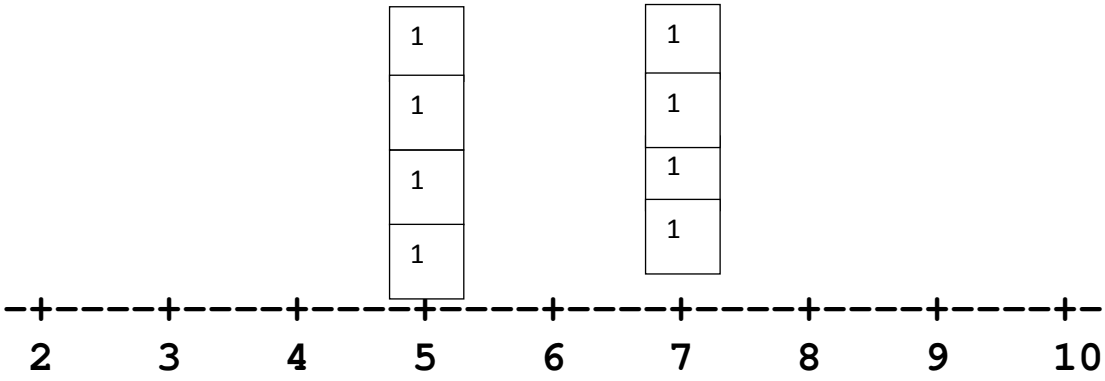
Adjusting the SAD for Group Size

The SAD provides a basis for comparing variation between two groups with the same number of observations in a sample. The following example illustrates a shortcoming in the SAD for groups or samples of difference sizes.

Group 3



Group 4



15. What is the mean for Group 3 and Group 4? What does this mean value tell you about the distributions for Group 3 and Group 4? Students should recognize the mean value for both distributions is 6. This value gives the balance point for each of the distributions.

16. What is the SAD for Group 3 and Group 4? What does the SAD tell us about each distribution? Do you see any potential issue with using the SAD to compare the variability of these two distributions? Students should recognize that the SAD for both distributions is 8. The SAD is a way to measure variability from the mean value of 6. However, by visually looking at the distributions, we can see the variability from the mean value of 6 is greater for Group 3 than for Group 4. The reason the SAD for Group 4 is as large as Group 3 is that the number of observations in the sample is larger (8 observations) than the number of observations in Group 3 (2 observations).

17. How could we adjust the SAD to take into account the number of observations in the sample? To adjust for the difference group sizes, determine the MAD (Mean Absolute Deviation), defined as:

$$\text{MAD} = \frac{\text{SAD}}{\text{Number of Data Values}}$$

The MAD provides for numerical data a measure of the variability from the mean. The larger the MAD, the more the data vary from the mean. The MAD measures the average distance or deviation of the observations in a distribution from the mean.

18. Determine the MAD for Group 3 and Group 4. What does the MAD tell us about how much the observations vary from the mean in each distribution? The MAD for Group 3 is four and the MAD for Group 4 is one. In Group 3, the family sizes vary on average from the mean of six by four whereas the family sizes vary on average from the mean of six by one. Group 4 has less variability than Group 3.

19. For the sampled Groups 1 and 2, interpret the results to answer the question, "How do the number of people in a student's household at this local school vary?" Students should comment that the mean value or balance point of the nine families is 6 and would be considered a typical family size. Since the two samples of 9 families is the same, we can use the SAD to measure how much variability from the mean the family sizes vary. We observe that Group 1 varies less with SAD of 16 compared to 18. Finding the MAD for each groups, we observed that for Group 1, the family sizes vary on average a deviation of 16/9 or 1.78 people from the mean of 6 compared to 18/9 or 2 people for Group 2 from the mean of 6.

Summary of Level B

Students completing the second part of this lesson that investigates the mean and variability from the mean at Level B should understand:

- The notion of "balance point" for a set of numerical or quantitative data

- The balance point of a distribution of numerical data is the mean value
- The Sum of the Absolute Deviations (SAD) is a measure of the amount of variability from the mean and it can be used when comparing groups if the size of the groups is the same.
- The Mean of the Absolute Deviations (MAD) is the average deviation of the observations in a distribution from the mean and can always be used to compare two or more groups regardless of the size of the groups.

Transitioning from Level B to Level C

The MAD is a precursor to the standard deviation, the measure used most often in practice for quantifying roughly on average how much numerical data deviates from the mean. After students understand how to use and interpret the MAD, they will find the interpretation of the standard deviation to be similar.

References

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., and Scheaffer, R. (2007), *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K–12 Curriculum Framework,* Alexandria, VA: American Statistical Association.

Franklin, C. and Mewborn, D, (2008), "Statistics in the Elementary Grades: Exploring Distributions of Data," *Teaching Children Mathematics,* 15(1), 10-16.

Kader, G. (1999), "Means and Mads," *Mathematics Teaching in the Middle School,* 4(6), 398-403.