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

How large are families today?
Student Guide

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".

1. What is the mean?

Let's consider question 1 in two parts:

2. How do you calculate the mean?

3. How do you interpret the mean?

Level A—How is the mean interpreted, and how is variability measured from the mean?

Formulate Statistical Questions

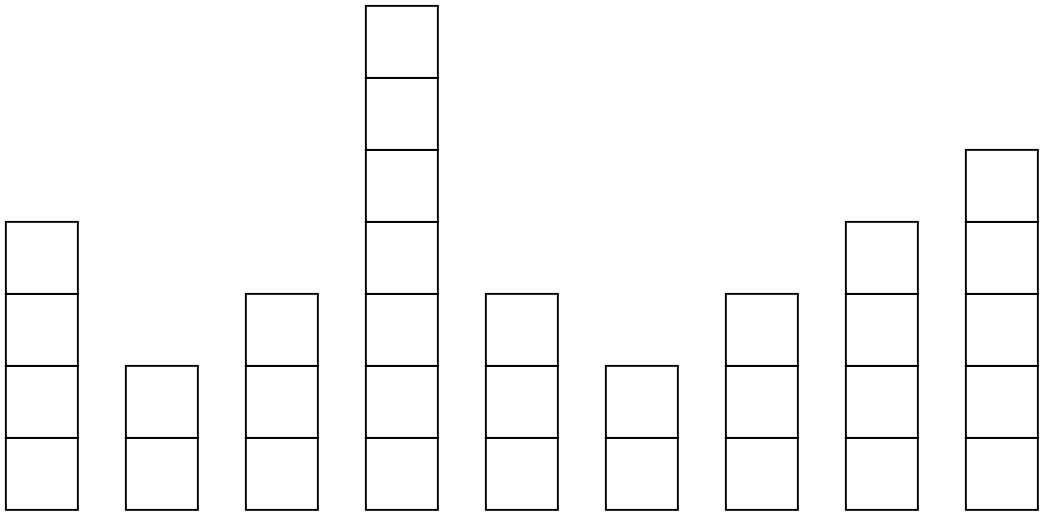
A local school is interested in knowing how many people live in the household of each student. A statistical question they might ask is: *How do the number of people in a student's household at this local school vary?*

4. What are three characteristics of this question that lead to classifying the question as a statistical question?

Collect Appropriate Data

Understanding the time and effort involved with taking a census of all student households at the school, the principal decides to 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 you live for most of year?" This is called the survey question to help answer the statistical or investigation question posed earlier.

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

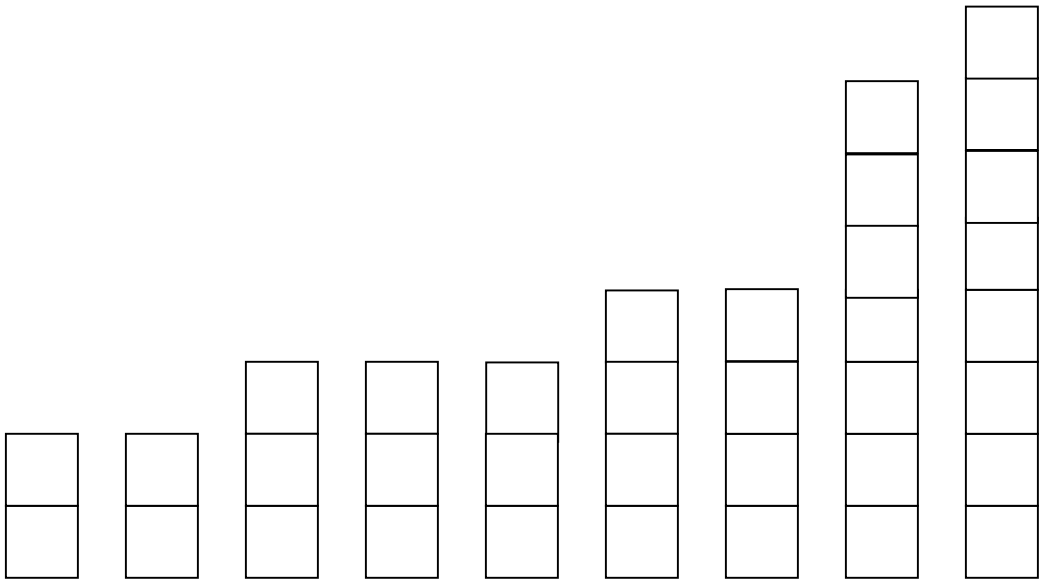


Analyze the Data

6. How might we examine the data as represented by the snap cube stacks for these nine children?

   If we put the original 9 stacks in order, we have the snap cube representation that is below after number 7.

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? This is called the "fair share" value.



Let's consider a new problem

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?

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 share value for each group is 6, we cannot distinguish the two groups based on the fair share value. A question we might ask is:

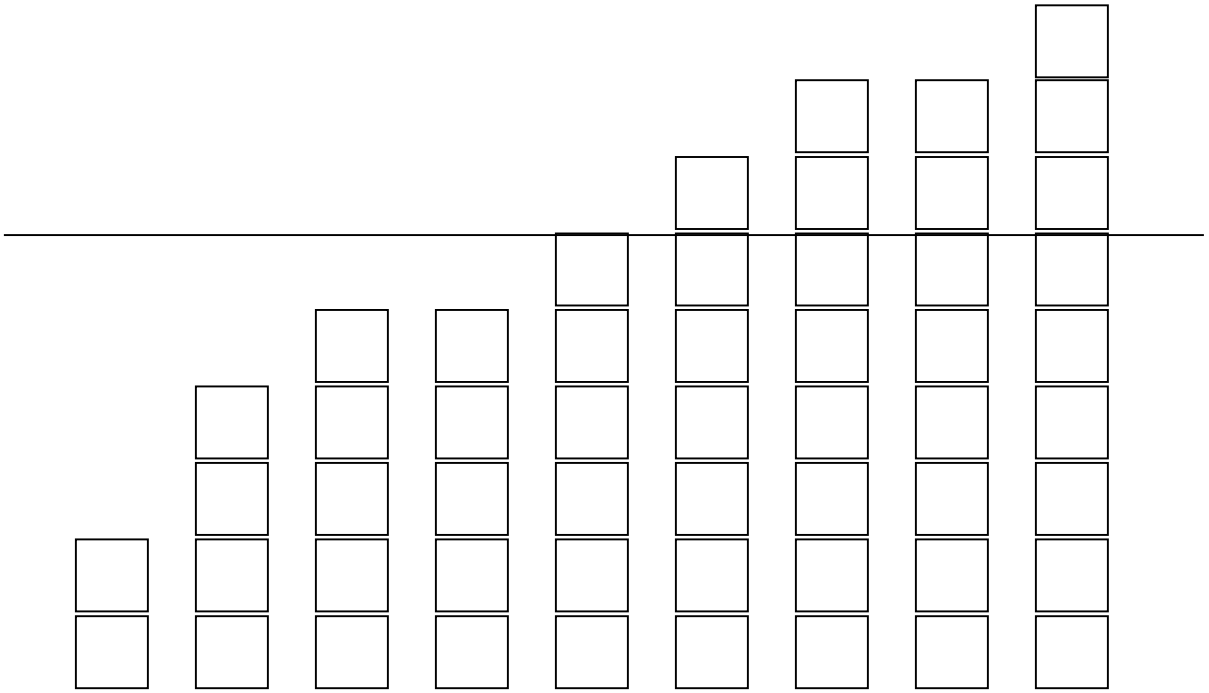
*Which group is **closer** to being fair?*

8. How might we decide "how close" a group of family sizes is to being fair?

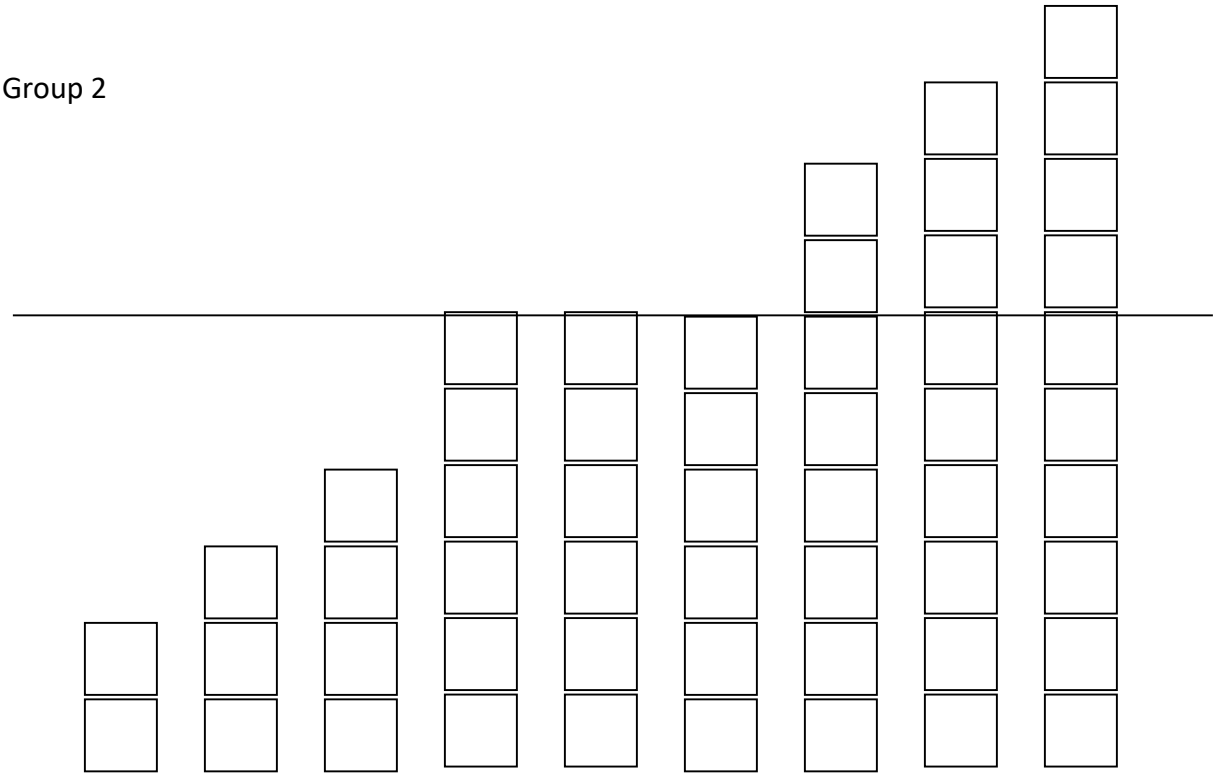
One way is to count the number of 'steps to fair' in moving snap cubes required to make the group fair to where the stacks are all the same size (or nearly the same unless there is a remainder). Fewer steps indicates closer to being fair and less variability.

How many steps are required for Group 1 and Group 2 to make the groups fair? Which group has less variability?

Group 1



Group 2



9. One of the two groups is symmetric. Is it Group 1 or Group 2? Explain your choice.

Let's create other groups of family sizes for nine children where the fair share value is 6. Your teacher will provide you with snap cubes and certain conditions that must be met when creating the stacks of snap cubes for the 9 children.

Interpret the results in the Context of the Original Question

Once all the new groups of families are created, determine the numbers of steps to fair for each group. Arrange the groups in order from the least amount of variability from the fair share to the greatest amount of variability from the fair share.

10. For each of the sampled groups, interpret the results to answer the question, "How do the number of people in a student's household at this local school vary?"

Summary of Level A

After completing this Level A activity, you 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—How is the mean interpreted, and how is variability measured from the mean?

We learned how to use snap cube representations for numerical data at Level A. We 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."

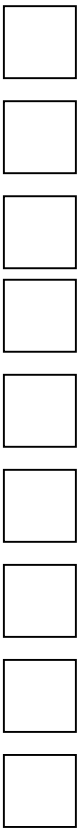
At Level B, we will learn an alternative interpretation of the mean and another way to quantity the degree of variability from the mean in the data. These two ideas will be developed using the dotplot representation for data.

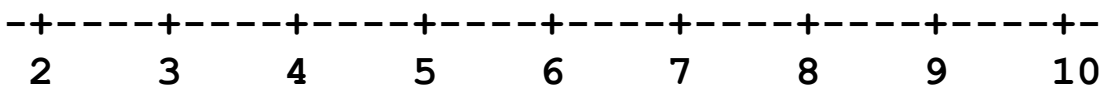
Formulate Statistical Questions

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?

Collect Appropriate Data

As in Level A, a pilot sample of nine students were selected to find out more about the size of their households. Each student was asked "How many people are in the household you live for most of year?" We will represent his/her family size with a post it note on a 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.

Analyze the Data

Let's investigate what other distributions of the family size might look like for nine students.

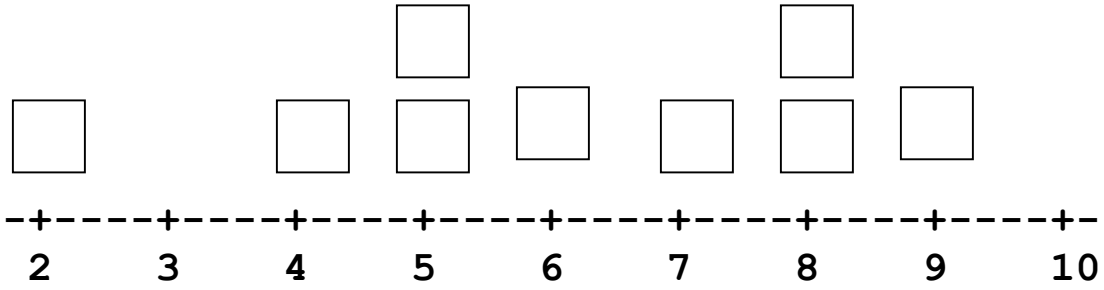
11. Working in groups, create a new dotplot on poster board representing the distribution of nine families with a mean family size of 6. Use 9 Post it notes to represent each family size. No family size can be below 2 and no family sizes can be above 10. Your teacher might provide you with additional conditions that must be taken into consideration.

    Your group should be prepared to explain how the distribution was developed.

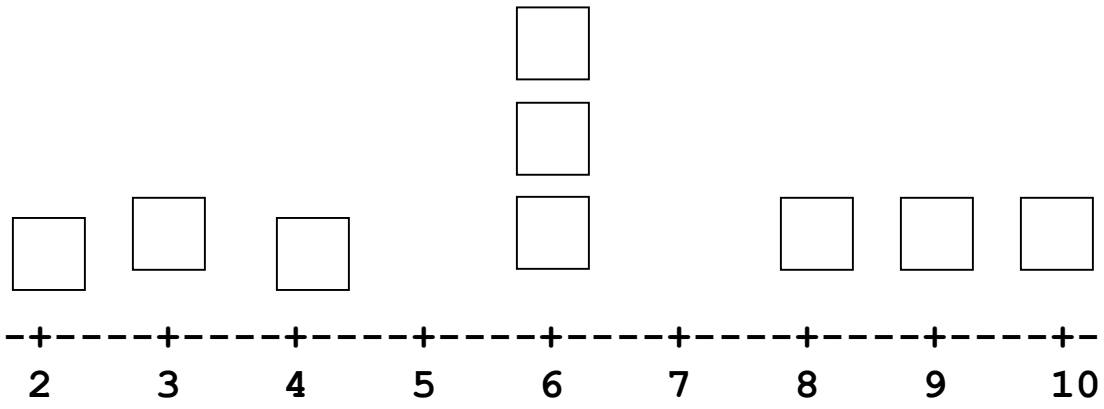
12. Once all the groups have explained how they created their distributions, observe that each of the distributions has different variability. As a class arrange the distributions represented on the poster boards in order of the least amount of variability from the balance point or mean value of 6 to the greatest amount of variability from the balance point. Explain how you quantified the variability for each distribution.

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

Group 1



Group 2

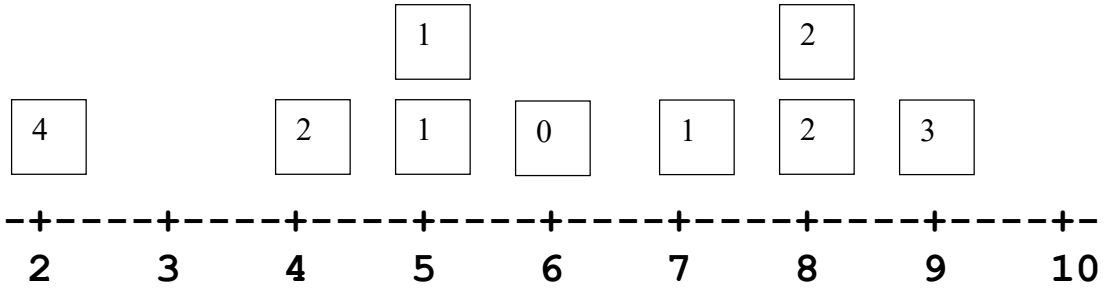


13. How do these two distributions compare? What is similar? What is different?

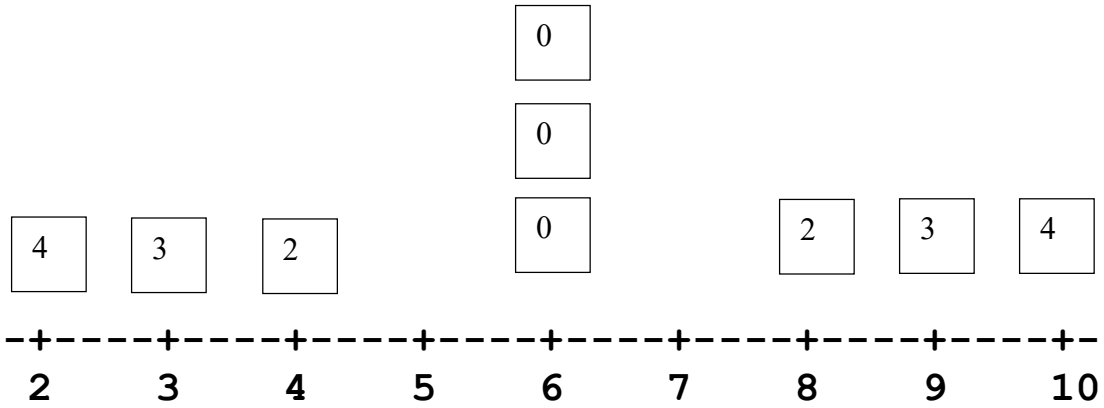
14. How might we quantify the amount of variability from the mean value of 6?

Recall the "number of steps" from Level A. We can think of a step as the unit distance each individual data value is from the mean.

Group 1



Group 2



We are interested in "overall" which group has more variability from the mean. One indicator is the Sum of these individual distances.

What is the sum or total for Group 1 and for Group 2? Which group has less variability from the mean? Explain.

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

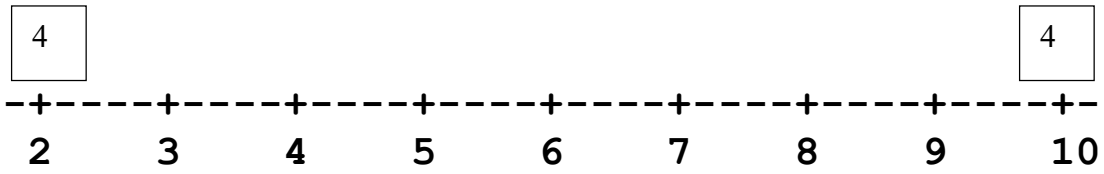
SAD = 2*(Number of Steps to Fair Share).

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.

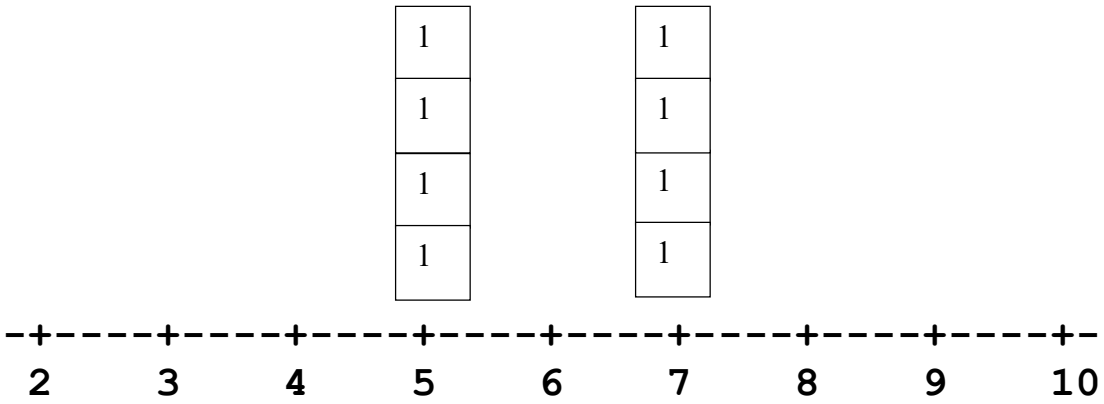
Adjusting the SAD for Group Size

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?

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?

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 mean in each distribution?

Interpret the results in the Context of the Original Question

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?

Summary of Level B

After completing this Level B activity, you 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.