Beware of outliers: Teacher notes

http://topdrawer.aamt.edu.au/Statistics/Misunderstandings/Misunderstandings-of-averages/Outliers

Overview

This short lesson is intended as a review of outliers and their potential effect on the mean but their lack of influence on the median. A data set is provided that contains the salt content of 30 food products measured in mg/100 gm. The data set contains an extreme value, which although it is genuine (not an error), acts as an outlier. Students are asked to plot the data – this may be done 'by hand' or by entering the data into a software package such as TinkerPlots or Excel. They then calculate the mean and median. After removing the outlier they repeat the steps above, discussing the differences. In replotting the data without the outlier it is important to consider the scale on the horizontal axis. (The product names are not necessary for the analysis but provide interest and context for classroom discussion.)

Objectives

- Review meaning of outlier.
- Appreciate the difference in influence an outlier has on the mean and median.
- Realise that the median is often the better measure of middle to use in a set with an outlying value that is not removed.

Materials

- Paper, pencils, rulers, calculators.
- Software for creating graphs (optional).

Lesson plan

Introduction

1. Decide in advance how the data provided on the student worksheet is to be handled. Students may plot it from the worksheet or enter it themselves into a software package (such as TinkerPlots). Or you may have already entered it into a package for the students to use in a computer lab or on their laptops.
2. Begin a discussion from one of two perspectives: the context of salt content in food and issues for health or the mathematical context of defining an outlier and looking for real-world examples where outliers occur. Sometimes outliers appear in data sets because of an error in measurement or recording and they can be deleted. Sometimes, as in this data set, the value is correct and the user has to have another reason to delete the value from the data set.

Exploration

3. Students create the plots and answer the questions on the worksheet. Reinforce the questions by asking "Why?" as the students work.

Wrap-up

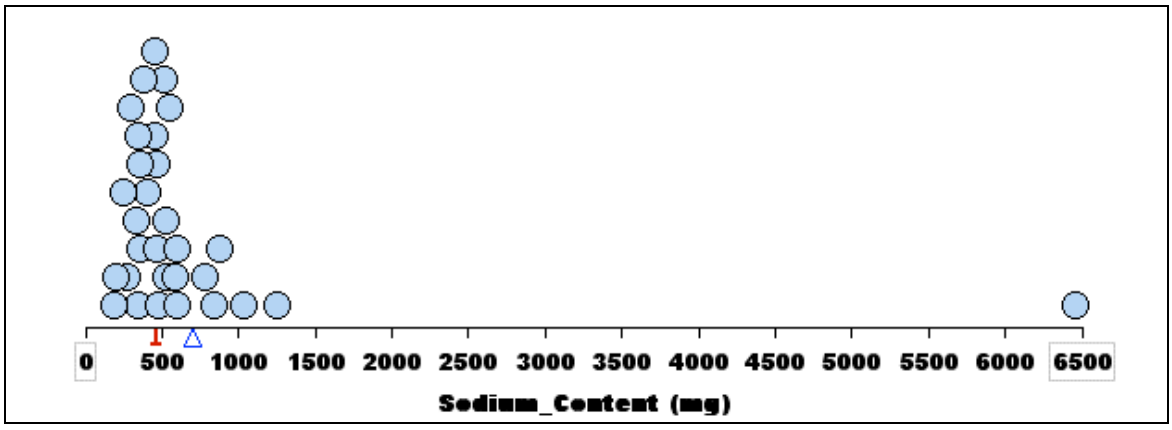
4. Ask students to report their findings, reinforcing the change in value of the mean and the reason for it. Remind them of the definition of mean as *balancing* every value in the data set. On the other hand by summarising the *ordered* data, the median does not take into account the actual values, only their order.

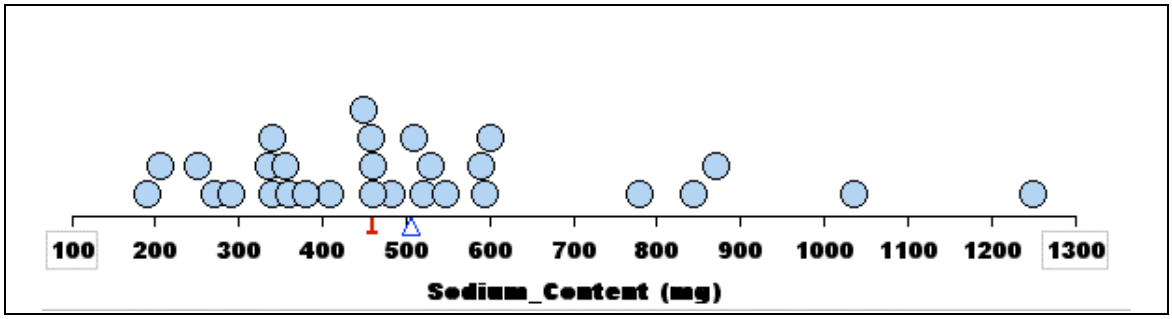
5. Discuss which is the most stable statistic to use in data sets where genuine outliers may present.

Answers

Here are the plots for the data sets with and without soy sauce. The plots show how different the data appear with the scale reduced without the outlier. The ⊥ symbol shows the median and the Δ shows the mean. Notice that the gap between ⊥ and Δ looks similar on the two plots but because the scales are so different, it hides the significance of the difference.

With soy sauce:	mean	= 706.5 mg
	median	= 460 mg
	difference	= 246.5 mg
Without soy sauce:	mean	= 507.2 mg
	median	= 460 mg
	difference	= 47.2 mg





Replotting the data without soy sauce on the same scale as the first plot, shows the much closer relationship of the mean and median.