Using statistical tools to analyze and make decisions concerning one-variable distributions

Descriptive statistics provides students with a variety of concepts that help them to begin making inferences. In secondary school, students become familiar with tools for processing the data collected, extracting information from that data and exercising critical judgment in order to identify potential sources of bias. The suggested activities involve displaying data using tables or graphs, depending on the type of data used. Students will also be required to interpret data by examining its distribution (e.g. shape, range, centre, groupings) or by comparing distributions. By the end of secondary school, students are aware of the variability of samples as well as the limitations and constraints associated with population sampling.

Analyzing a distribution and making related decisions requires students to choose statistical measures that describe the distribution. Two statistical measures are generally used to describe one-variable distributions: a measure of central tendency and a measure of dispersion. In Secondary III, students learn to use the mode, the median and the weighted mean as well as the interquartile range. In Secondary IV, students in the *Cultural, Social and Technical* option learn to use the mean deviation, and those in the *Technical and Scientific* option learn to use the mean deviation and the standard deviation. Students must choose the most appropriate pair of measures for the distribution they are analyzing. The two most frequently used measures in this regard are the mean and the standard deviation, but the median and the interquartile range are also used.

For example, students can be presented with the following situation:

Various studies have shown the impact of head injuries on reflexes.

To see if he is fit to resume play following an injury, a hockey player must, among other things, undergo a test measuring the quality of his reflexes. His results are then compared with those obtained by the whole team at training camp.

During the last game, Felix suffered a head injury after a violent check. A few days later, he obtained a score of 34 on his reflex test. To decide whether Felix is fit to resume play, the coach must compare his results with those of the rest of the team:

36	36	37	38	39	39	40
40	40	41	41	41	42	42
42	42	42	42	43	43	43
43	44	44	44			

What might the coach decide?

If the coach knew that Felix's score at training camp was 38, do you think he would change his mind?

Comments

To analyze the distribution and provide answers in this situation, students must choose a measure of central tendency and a measure of dispersion. When selecting a measure central tendency for a set of data values, students often automatically calculate its mean. However, the mean is not always appropriate or sufficient to describe the data. Because the mean is greatly affected by extreme values, particularly in cases where the distribution is skewed, it is preferable to use the median, since it is not influenced by extreme values and outliers. In addition, calculating the median can provide some assurance as to the validity of the mean when the values of these two measures are similar.

In Secondary IV, students may choose to use the mean and the mean deviation. Although used less frequently than the standard deviation, the mean deviation can also be used to describe the dispersion of a one-variable distribution. The most common way of calculating it enables students to understand the close relationship between the mean and the mean deviation. Compared with the mean deviation, the standard deviation is less intuitive. However, because of its mathematical properties, it has the advantage of being useful in a great deal of statistical work. Furthermore, since it is based on the square of the deviations, the standard deviation captures the influence of the extreme values of the distribution much more than the mean deviation does. However, because of their close relationship with the mean, standard deviation and mean deviation can be greatly affected if the mean does not accurately represent the central tendency.

On the other hand, even in Secondary IV, students could choose to use the median and the interquartile range. The latter provides a good estimate of the dispersion, and it is much easier to understand and calculate than the standard deviation or the mean deviation. Moreover, since it is concerned with the dispersion of the values around the median, it is not affected by extreme values, making it particularly useful in the case of skewed distributions.

Let's return to the example on the previous page. After performing various calculations, the student may obtain the following measures:

mean: 40.9	mean deviation: 1.9	standard deviation: 2.3
Q1: 39.5	median: 42	Q3: 43

Preliminary analysis of these results, which shows that the values of the mean and the median are significantly different and that the distribution is skewed, could lead students to question whether the mean is a valid measure of central tendency for this distribution. Students in both Secondary III and IV should therefore choose to use the median and the interquartile range.

To predict the coach's decision, students must determine the position of Felix's score of 34 in the distribution. Generally, a data value is considered an outlier if it is less than Q1 – 1.5(Q3 – Q1). Since 34 is less than 34.25, students could conclude that Felix's score is an outlier. Consequently, Felix is not fit to play.

On the other hand, Felix's score at training camp was lower than Q1. Students could argue that the coach might change his mind because 34 is greater than 38 – 1.5(Q3 – Q1).