The standard deviation as a descriptive statistic

by
Von Bing Yap*
Department of Statistics and Applied Probability,
National University of Singapore

Introduction

The bulk of statistics essentially deals with the application of probabilistic models to solve practical problems. This sounds easier than it is, for the major issues involved are often obscured by the mathematical technicalities, thus making the subject a big challenge to teach even in a university. A quick glance at the numerous elementary statistics textbooks points to many possible approaches, underlining the complexity of the task.

Thankfully, descriptive statistics is relatively straightforward. In a world increasingly dominated by the analysis of large data sets in professional as well as private situations, descriptive statistics is a vital part of a citizen's education, serving as a solid base for learning complex statistical techniques. It is already an integral part of the secondary school mathematics curriculum in Singapore. The histogram, the dot diagram or the stem-and-leaf diagram provide visual summary of data. The average or the median locate the "centre" of the data distribution, while the mode indicates the peaks, and the percentiles give more details about the shape of the distribution. Moreover, the idea of "spread" in data is quantified by the interquartile range or the standard deviation. However, we often see summary reports containing only the average of a distribution, with no indication of its spread. Perhaps this is due to the relative lack of an attempt to relate the average and the SD to the distribution. The aim of this article is to emphasis this aspect of the standard deviation. The treatment essentially follows the great book *Statistics 4th edition* by Freedman, Pisani and Purves, abbreviated to FPP.

Address: Blk S16, Level 7, Faculty of Science, 6 Science Drive 2, Singapore 117546. Tel: (65) 6516 7143.
FAX: (65) 6872 3919. email: stayapvb@stat.nus.edu.sg

The standard deviation

Let us consider two small lists of numbers:

(a) $1, 3$

(b) $0, 1, 3, 4$

Both data sets are centred at 2, which is the average, or the mean. But it is clear that the second list is more spread out around the centre.

To measure spread, let us first define the **deviations** of a list of numbers as a new list obtained by subtracting the average from the original list. Thus the deviations of our two lists are

(a) $-1, 1$

(b) $-2, -1, 1, 2$

Evidently, the deviations always sum to 0 and tell us how far the numbers in the original list are from the centre. One simple measure of spread is the average of absolute deviations. This gives 1 and $6/4 = 1.5$ respectively for the lists. It is a sensible measure, i.e., the more spread out a list, the larger its average absolute deviations.

However, because of theoretical considerations (essentially related to the normal distribution and the Central Limit Theorem), statisticians have invented a slighly more complicated measure.

The **standard deviation** of a list of numbers is the **root-mean-square** of the deviations.

The calculation of the SD proceeds as follows

(i) Find the deviations, by subtracting the average from every number.

(ii) Square the deviations.

(iii) Find the mean of the squared deviations.

(iv) Take the square-root of the mean squared deviations.

We go through the steps for the first list. (i) The deviations are $-1, 1$. (ii) The squared deviations are $(-1)^2 = 1$, $1^2 = 1$. (iii) The mean squared deviations is $\frac{1+1}{2} = 1$. (iv) The SD is $\sqrt{1} = 1$. Similarly, the SD of the second list is found to be $\sqrt{2.5} \approx 1.6$. Just like the mean absolute deviations, the SD is a reasonable measure of spread. Due to the square root operation, the SD has the same unit as the original list.

The SD tells us the typical size of deviations from the centre of a data set. Thus, most numbers in the list are around 1 SD away (above or below) from the average. Relatively few numbers are beyond 2 or 3 SDs away from the average. These simple observations can be used to guess the SD of a list of numbers.

Example 1 (FPP page 74)

Here is a list of 20 numbers:

0.7	1.6	9.8	3.2	5.4	0.8	7.7	6.3	2.2	4.1
8.1	6.5	3.7	0.6	6.9	9.9	8.8	3.1	5.7	9.1

(a) Without doing any arithmetic, guess whether the average is around 1, 5, or 10.

(b) Without doing any arithmetic, guess whether the SD is around 1, 3, or 6.

Solution

The numbers are centred around 5, so this should be the average. Only three numbers (5.4, 4.1 and 5.7), or 15%, are in the interval (4, 6), so the SD should be more than 1. On the other hand, all numbers lie in the interval $(-1, 11)$, so the SD should be less than 6. It is around 3.

The average and the SD are excellent summary statistics when the data distribution is approximately normal. In this case, around 68% of the numbers are within 1 SD of the average, and around 95% are within 2 SDs. These rules even apply to some non-normal distributions. They come from the fact that the area between -1 and 1 under the standard normal curve is about 0.68, and that between -2 and 2 is about 0.95.

Example 2 (FPP pages 67–69)

In the Health and Nutrition Examination Survey of 1976–1980, a representative sample of the United States' population was studied. There were 6,588 women age 18–74 in the sample. Their average height was about 161 cm and the SD was about 6 cm. Around 67% of the women were in the interval 157 cm to 167 cm, i.e., they differed from the average height by 1 SD. Around 94% differed by 2 SDs from the average. The rules are quite accurate. In fact, the height distribution is approximately normal. Interestingly, the weight distribution of the women is not normal, having a long right-hand tail, i.e., a small number of women had relatively large weights (FPP page 62).

If the data distribution is approximately normal, one can figure out roughly the percentage of numbers within any given range, using only the average and the SD. Just ask how many SDs the end-points of the range are away from the average, and then consult a standard normal table to find the corresponding area under the curve.

Example 3 (partially hypothetical)

The average monthly income of a Singapore household in 2006 is \$5,700, and the SD is \$4,300. Is the household income normally distributed?

Solution

Suppose that the monthly household income in 2006 follow a normal distribution. The income \$0 is

$$\frac{\$0 - \$5,700}{\$4,300} \approx -1.3$$

SDs away from the average, or 1.3 SDs below the average. Since the area under the standard normal curve to the left of -1.3 is about 0.10, we predict that 10% of households have negative income. But the actual percentage is 0%, hence the household income is not normally distributed.

The distribution in example 2 is skewed, having a long right-hand tail. If the degree of skewness is severe, then the average and the SD may not be good summary statistics. They can also be unsatisfactory if the distribution is multi-modal, i.e., it has more than one peak. A famous example is the bi-modal distribution of inter-eruption times of the Old Faithful Geyser in Yellowstone National Park.

Technical notes

Let a list of n numbers be $x_1, \ldots, x_n$. Denote the average by $\bar{x}$. Then

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

With computer softwares, descriptive statistics can be obtained easily by pushing a few buttons. Often, the SD given by a software is not the same as the one defined here, but something slightly larger, which is known as the "sample SD":

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

To find out which one is calculated, enter the list $-1, 1$. If the answer is 1, then it is the SD. If the answer is 1.41..., then it is s. In general, the SD can be computed from s by

$$\text{SD} = \sqrt{\frac{n-1}{n}} \times s$$

In many textbooks, s is used as a summary statistics instead of the SD. Some reasons for not recommending this practice are as follows.

1. The SD definition is simpler. Students will ask why the denominator of s is $n-1$. The reason is quite technical, and well beyond the present scope. One way to phrase it is the following.

   Suppose that we have a large set of numbers with average μ and SD σ, from which a sample random sample of size at least 2 is taken. Then s^2 is an unbiased estimate of σ^2, but the square of the SD of the sample is biased.

   A full understanding of these statements is seldom attained within the first month of undergraduate statistical life.

2. The SD is a better measure of spread. The list $1, 1, 3, 3$ is equivalent to $1, 3$, since the relative frequencies of 1 and 3 are the same. Their SDs are both 1, but the first s is smaller than the second s.

3. The discrimination of "sample SD" and "population SD" begs the question of how to distinguish between a sample and a population. It really depends on the situation: If I take a random sample X from an actual sample Y, which has been taken at random from a population Z, should Y be considered a sample or a population? For the purpose of description, a data set is just "a list of numbers", and there is no need to decide if we are dealing with a sample or a population.

4. The introduction of s in more advanced statistics courses will emphasise the various subtle and delightful issues awaiting every application of probabilistic models in real life. An important one: it is not always reasonable to assume that your data set is like the result of random sampling from some population. Using s for descriptive purposes spoils the fun both in the present and in the future.

Conclusion

I have briefly demonstrated the standard deviation (SD) as a useful descriptive statistics. This is part of a general philosophy of an integrated approach to teaching descriptive statistics in secondary schools. Emphasising the interpretation of various summary statistics, like the histogram, the mean, the SD, the normal distribution, and others, in terms of data sets, will introduce statistics in a more persuasive and fruitful way to our young students.