

## Properties of the Standard Deviation that are Rarely Mentioned in Classrooms

Mohammad Fraiwan Al-Saleh<sup>1</sup> and Adil Eltayeb Yousif<sup>2</sup>

<sup>1</sup> Department of Mathematics, University of Sharjah, UAE

<sup>2</sup> Department of Mathematics and Physics, Qatar University, Qatar

**Abstract:** Unlike the mean, the standard deviation  $\sigma$  is a vague concept. In this paper, several properties of  $\sigma$  are highlighted. These properties include the minimum and the maximum of  $\sigma$ , its relationship to the mean absolute deviation and the range of the data, its role in Chebyshev's inequality and the coefficient of variation. The hidden information in the formula itself is extracted. The confusion about the denominator of the sample variance being  $n - 1$  is also addressed. Some properties of the sample mean and variance of normal data are carefully explained. Pointing out these and other properties in classrooms may have significant effects on the understanding and the retention of the concept.

**Zusammenfassung:** Anders als beim Mittel unterliegt die Standardabweichung  $\sigma$  einem vagen Konzept. In diesem Aufsatz werden verschiedene Eigenschaften von  $\sigma$  hervorgehoben. Diese Eigenschaften inkludieren das Minimum und das Maximum von  $\sigma$ , ihre Beziehung zur Mean Absolute Deviation und der Spannweite der Daten, ihre Rolle in der Chebyshev Ungleichung und dem Variationskoeffizient. Die versteckte Information in der Formel selbst ist extrahiert. Die Konfusion über den Nenner der Stichprobenvarianz ( $n - 1$ ) wird auch angesprochen. Einige Eigenschaften des Stichprobenmittels und der Stichprobenvarianz von normalverteilten Daten sind sorgfältig erklärt. Weist man auf diese und andere Eigenschaften in der Klasse hin, so kann dies maßgebliche Effekte auf das Verstehen und gegen Vorbehalte dieses Konzepts haben.

**Keywords:** Coefficient of Variation, Standard Deviation, Mean Absolute Error, Chebyshev's Inequality.

### 1 Introduction

Unlike other summary quantities of the data, the standard deviation is a concept that is not fully understood by students. The majority of students in elementary statistics courses, though it can calculate the standard deviation for a set of data, does not understand the meaning of its value and its importance. The purpose of this short paper is to highlight some of the known properties of the standard deviation that are usually not discussed in courses.

Assume first that we have a finite population that consists of  $N$  units. Denote these elements by  $u_1, u_2, \dots, u_N$ . Let  $u_{(1)}, u_{(2)}, \dots, u_{(N)}$  be the ordered data. In later sections,

we will consider the case of random samples taken from a population. The most commonly used quantities to summarize or describe the population elements are the mean

$$\mu = \frac{1}{N} \sum_{i=1}^N u_i,$$

the median

$$\text{med} = \frac{1}{2} (u_{(\lfloor (N+1)/2 \rfloor)} + u_{(\lfloor N/2 \rfloor + 1)}) ,$$

the variance

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (u_i - \mu)^2,$$

and the standard deviation

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (u_i - \mu)^2}$$

as also the mean absolute deviation

$$\text{MAD} = \frac{1}{N} \sum_{i=1}^N |u_i - \mu|$$

and the range

$$\text{range} = \max - \min = u_{(N)} - u_{(1)} .$$

## 2 Some Basic Properties of the Standard Deviation

Clearly,  $\sum_{i=1}^N (u_i - \mu) = 0$ . Thus,  $\mu$  is the only number that makes the sum of deviations zero. A measure of variability of the data can be based on  $(u_i - \mu)^2$  or on  $|u_i - \mu|$ . Thus, one can use  $\sigma^2$  or MAD. The unit of  $\sigma^2$  is the square of the unit of the data, while  $\sigma$  and MAD have the same unit as the data. Now since  $\left(\sum_{i=1}^N (u_i - \mu)^2\right)^{1/2} \leq \sum_{i=1}^N |u_i - \mu|$ , we have the following inequality that relates  $\sigma$  to MAD:

$$\sigma \leq \sqrt{N} \text{MAD} .$$

Using some simple calculus, it can easily be shown that

$$\min_t \left[ \sum_{i=1}^N (u_i - t)^2 \right] = \sum_{i=1}^N (u_i - \mu)^2,$$

while

$$\min_t \sum_{i=1}^N |u_i - t| = \sum_{i=1}^N |u_i - \text{med}| .$$

Thus,  $\mu$  is the reference point that minimizes the sum of squared distances, while in the case of absolute distances, the reference point is the median. Therefore, we may raise the following question: Don't you think that in defining the MAD,

$$\frac{1}{N} \sum_{i=1}^N |u_i - \text{med}|$$

should be used instead of

$$\frac{1}{N} \sum_{i=1}^N |u_i - \mu| ?$$

### 3 Average of Squares and the Square of the Average

Since  $\sum_{i=1}^N (u_i - \mu)^2 = \sum_{i=1}^N u_i^2 - n\mu^2 > 0$ , assuming not all the  $u_i$  are equal, we have (a special case of the Cauchy-Schwarz inequality)

$$\frac{1}{n} \sum_{i=1}^N u_i^2 > \left( \frac{1}{n} \sum_{i=1}^N u_i \right)^2,$$

i.e.

$$\overline{u^2} > \bar{u}^2.$$

The inequality says that the average of squares is larger than the square of the average. The difference between the left and the right side of the inequality is the variance! This can neatly be put as: the variance is the difference between the average of the squares and the square of the average.

### 4 Minimum and Maximum Value of the Standard Deviation

For  $u_1, u_2, \dots, u_N$  the minimum possible value of  $\sigma$  is zero, which occurs when  $u_1 = \dots = u_N$ . It can be shown that if  $a \leq u_i \leq b$  ( $\forall i$ ), then the maximum possible value of  $\sigma$  occurs when half of the values are equal to  $a$  and half of them are equal to  $b$ , provided that  $N$  is even (this also holds with a slight modification, if  $N$  is odd):

$$\max_{a \leq (u_1, u_2, \dots, u_N) \leq b} \left[ \frac{1}{N} \sum_{i=1}^N (u_i - \mu)^2 \right] = \frac{(b-a)^2}{4},$$

i.e.

$$\sigma \leq \frac{b-a}{2} = \frac{1}{2} \text{ range}.$$

This can be seen easily if our data is binary with proportion  $p$  of one kind and  $1-p$  of another kind. In this case,  $\sigma^2 = p(1-p)$ , with the maximum value of  $1/4$  occurring at  $p = 1/2$ .

A large value of  $\sigma$  is an indication that the data points are far from their mean (or far from each other), while a small value indicates that they are clustered closely around their mean. For example, when dealing with students marks  $u$ ,  $0 \leq u \leq 100$  say,

$$0 \leq \sigma \leq \frac{100 - 0}{2} = 50.$$

A very small value of  $\sigma$  means that all students have obtained almost the same grade. So the test was unable to distinguish between students' abilities, putting them in one very homogeneous group. An extreme value of  $\sigma$  means that the test was too strict and classifies the students into two very different groups (strata). Hence, the value of  $\sigma$  can be used as a guide on how many different strata (a stratum is a set of relatively homogeneous measurements) a set of measurements can have. Thus,  $\sigma$  can be used as a measure of discrimination. If the marks have a normal curve, then about 95% of the values are within two standard deviations from the mean. Thus, the range is about four standard deviations, i.e.  $\sigma \approx 100/4 = 25$ .

In repeated measurements on the same object using the same scale,  $\sigma$  can be regarded as a measure of precision. A very small value of  $\sigma$  indicates that the scale is very precise and a large value indicates that the scale is imprecise. Note that:

precise scale  $\nrightarrow$  accurate scale!

If an object has an actual weight of 60 kg, but a weighing scale gives in five trials the measurements 70.04, 70.05, 70.06, 70.03, and 70.05 kg, then the scale can be regarded as being very precise, because the values are very close to each other (very small standard deviation) but not accurate because they are far away from the actual value (60 kg). Hence, a precise scale does not have to be an accurate one. Thus,  $\sigma$  can be used as a measure of precision.

## 5 Jointly considering Mean and Standard Deviation

The following inequality gives a further indication of the importance of the standard deviation:

*Chebyshev's or Chebyshev-Bienaymé's Inequality:* For any  $k > 0$

$$\Pr(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

Thus, for any set (the population) we can say that at least  $(1 - \frac{1}{k^2})100\%$  of the elements belong to the interval

$$(\mu - k\sigma, \mu + k\sigma).$$

This percentage is at least 75.0%, 88.9%, 93.8%, and 96.0% for  $k = 2, 3, 4, 5$ , respectively. For a normal population the corresponding percentages are: 95.4%, 99.7%, 99.99%, and 100.0% while for the standard double-exponential distribution, these percentages are 94.4%, 98.0%, 99.6%, and 99.9%.

In particular, at least 50% of the data set is in the interval

$$(\mu - \sqrt{2}\sigma, \mu + \sqrt{2}\sigma).$$

The Chebyshev-Bienaymé inequality is sharp and equality holds if the population consists of three values,  $c$ ,  $0$ , and  $-c$  say, such that  $\frac{50}{k^2}\%$  are  $c$ ,  $\frac{50}{k^2}\%$  are  $-c$ , and the remaining  $100(1 - \frac{1}{k^2})\%$  are  $0$ .

If  $X_1, X_2, \dots, X_n$  is a random sample from a population with mean  $\mu$  and variance  $\sigma^2$ , then it may not be true that at least  $(1 - \frac{1}{k^2})100\%$  of the data points belong to the interval  $(\mu - k\sigma, \mu + k\sigma)$ . What is still true is that at least  $(1 - \frac{1}{k^2})100\%$  of the data points belong to the interval  $(\bar{x} - ks, \bar{x} + ks)$ , where  $\bar{x}$  and  $s$  are the sample mean and sample standard deviation.

## 6 Comparison of Two Data Sets

The standard deviation  $\sigma$  has the same unit as the measurements. This makes it not a suitable measure to use in order to compare data sets with different units of measurements (for example temperature in  $F^\circ$  or in  $C^\circ$ , weight in kg or in pounds, etc.). Even when the units of measurements are the same, it may also not be a suitable measure to compare the variability in two or more data sets. Note that, for any constant  $c$ , the standard deviation of the data set  $\{u_i : i = 1, \dots, N\}$  is the same as the standard deviation of  $\{u_i + c : i = 1, \dots, N\}$ . For example the two sets  $\{10, 20, 30, 40, 50\}$  and  $\{10^6 + 10, 10^6 + 20, 10^6 + 30, 10^6 + 40, 10^6 + 50\}$  have the same value of  $\sigma = 10\sqrt{2} = 14.14$ . However, it is very clear that the variability in the second set is negligible when compared to that of the first set.

A normalized measure of variation that can be used to compare two data sets is called the coefficient of variation (CV):

$$CV = \frac{\sigma}{|\mu|}, \quad \mu \neq 0.$$

The CV is unit free. For the above two sets, it is 0.47 for the first set and almost zero for the second one.

### 6.1 Comparison using the Difference or the Ratio

To compare two quantities, one may obtain their difference or their ratio. If the two quantities are very close to zero, then their ratio is more informative than their difference. For example, if the probability of getting a disease in one place is  $p_1 = 0.0004$  and in another place is  $p_2 = 0.0001$ , then the difference is negligible and, thus, misleading while the ratio is  $\frac{p_1}{p_2} = 4$ , which is more informative. A similar argument applies if the two quantities are very close to 1. In general, if two quantities are extremely large or extremely small then they should be compared using their ratio rather than their difference, because the difference can be misleading. Now consider the following argument:

$$\frac{1}{n} \sum_{i=1}^N u_i^2 > \left( \frac{1}{n} \sum_{i=1}^N u_i \right)^2 = \mu^2$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^N u_i^2 - \mu^2 > 0$$

$$CV^2 = \frac{\frac{1}{n} \sum_{i=1}^N u_i^2}{\mu^2} - 1 > 0.$$

Thus, it is interesting to note that  $\sigma$  is a measure that is obtained based on the difference between the average of squares and the square of the average of the data, while CV is a measure that is obtained based on the ratio of the same two quantities. Based on this observation, one may conclude when to use  $\sigma$  and when to use CV as a measure of variation.

## 6.2 Problems with the CV

- It cannot be used with a population that has zero mean. Even when  $\mu$  is close to zero, the CV is sensitive to small changes in the mean.
- The data  $\{100cm, 200cm, \dots, 1000cm\}$  has the same CV as  $\{1m, 2m, \dots, 10m\}$ , while the CV of  $\{10C^\circ, 15C^\circ, \dots, 50C^\circ\}$  is 0.46 but the CV of  $\{50F^\circ, 59F^\circ, \dots, 122F^\circ\}$  is 0.29. Thus in order to use the CV in comparing data sets, the base of the units should be the same.

Thus, the CV does not truly remove the effect of the unit of measurements.

The standard deviation can be used to standardize the score to allow the comparison between different sets  $Z = (X - \mu)/\sigma$ , the comparison of relative standards of level can be done using standard scores (Bluman, 2007). Other use of the standard deviation is in the standardized moment which is given by  $\mu_k/\sigma^k$ , where  $\mu_k$  is the  $k$ th moment about the mean. The first standard moment is zero, the second standard moment is one, the third standard moment is called the skewness, and the fourth standard moment is called the kurtosis.

## 7 The Sample Mean and Variance

Assume that we have a random sample  $X_1, X_2, \dots, X_n$  from a population with mean  $\mu$  and variance  $\sigma^2$ . The sample mean and variance are defined as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Since dividing by  $n$  is more natural and seems to be logical, many students wonder why the formula calls for dividing by  $n - 1$ . We face some difficulties in convincing students of reasons behind this specific divisor.

## 7.1 Supporting Arguments for the $n - 1$ Divisor

For more details on this aspect see Al-Saleh (2005).

**Unbiasedness of  $S^2$ :** When the underlying population that we sample from is infinite (i.e.  $N = \infty$ ), this divisor is the only divisor that makes  $S^2$  an unbiased estimator of the population variance  $\sigma^2$ , i.e.  $E(S^2) = \sigma^2$ . Now,

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Therefore, dividing by  $n$  or any other divisor greater than  $n - 1$  will give, on the average, values of  $S^2$  that are smaller than the population variance  $\sigma^2$ . Hence, in using  $S^2$  to estimate  $\sigma^2$ , the divisor  $n - 1$  provides the appropriate correction. In other words, the above argument says that the  $(n - 1)$  is used because it is the only choice that makes  $S^2$  an unbiased estimator of  $\sigma^2$ .

But, what about  $S$ , the sample standard deviation. Actually, by applying Jensen's inequality we get  $E^2(S) < E(S^2) = \sigma^2$  and hence  $E(S) < \sigma$  which means that  $S$  is biased. So why not choose a quantity that makes  $S$  unbiased? Also, if  $N$  is finite then  $S^2$  is biased because

$$E(S^2) = \frac{N}{N-1} \sigma^2.$$

Thus

$$\frac{N-1}{N} \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

is the unbiased estimator of  $\sigma^2$  and not  $S^2$ . That makes this argument about the division by  $(n - 1)$  even harder.

**The Indefinite of 0/0 Argument:** The sample variability is used to assess the variability in the population. Clearly, we can not assess the variability of a population based on a sample of size one. We should have at least two data points from the population to get an idea about the variability in the population. Defining  $S^2$  with  $n - 1$  as divisor reflects this intuitive argument. If  $n = 1$  then the formula of  $S$  breaks down giving the indefinite value  $0/0$ , which reflects the impossibility of estimating the variability of the population elements based on one observation (Samuels and Witmer, 1999; Bland, 1995).

**The Degrees of Freedom Argument:** This argument is based mainly on the fact that

$$\sum_{i=1}^n (X_i - \bar{X}) = 0.$$

Hence, once the first  $n - 1$  deviations have been obtained the last becomes known: We have the freedom to choose only  $(n - 1)$  deviations. So, we are not actually averaging  $n$  unrelated numbers, but only  $n - 1$  unrelated pieces of information. Hence the divisor should be  $n - 1$ . Most statisticians refer to this number as the degree of freedom (Heath, 1995; Moore and McCabe, 1993; Samuels and Witmer, 1999; Siegal and Morgan, 1996).

Clever students respond to this argument by saying, although we have only  $n - 1$  unrelated numbers or pieces of information, we are still adding  $n$  not  $n - 1$  numbers. Furthermore, in the case of a finite population,  $\{u_1, u_2, \dots, u_N\}$ , say, we also have  $\sum_{i=1}^N (u_i - \mu) = 0$ , but we call for  $N$  as divisor, not  $N - 1$ , in defining the formula of  $\sigma^2$ . The same thing is noticed in defining the formula of  $\text{MAD} = \frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}|$ . Thus, clever students can easily notice our inconsistency here.

Therefore, a more convenient way of introducing students to the concept of standard deviation or variance is through the separation of descriptive and inferential statistics.

- Assume that we have a random sample of  $n$  points from a population with mean  $\mu$  and variance  $\sigma^2$ . The variability in the data set is measured by the standard deviation, which is naturally defined as the positive square root of the average squared deviations. Thus, using  $n$  as divisor instead of  $n - 1$  gives

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

If  $n = 1$ , then  $S = 0$ , a acceptable value for a sample of size one.

- Now, when moving from descriptive to inferential statistics, we want to use the data set to estimate the unknown parameters  $\mu$  and  $\sigma^2$ . For this we can use different available methods of estimating the parameters of the population. For example, if we assume that the population is infinite then  $\bar{X}$  is an unbiased estimator of  $\mu$  and  $\frac{n}{n-1}S^2$  is an unbiased estimator of  $\sigma^2$ . If the population is of size  $N$ , then again  $\bar{X}$  is an unbiased estimator of  $\mu$  while  $\frac{(N-1)n}{N(n-1)}S^2$  is an unbiased estimator of  $\sigma^2$ .

## 8 Normal Population

Assume that the data  $X_1, X_2, \dots, X_n$  is a random sample from a normal population with mean  $\mu$  and variance  $\sigma^2$ . Then  $(\bar{X}, S^2)$  is a complete sufficient statistic for  $(\mu, \sigma^2)$ .

The following statements are some simple meanings of the above fact that can be easier understood by students:

- $(\bar{X}, S^2)$  being sufficient means that all information in the sample  $X_1, X_2, \dots, X_n$  about  $(\mu, \sigma^2)$  is contained in  $(\bar{X}, S^2)$ . Given the value of  $(\bar{X}, S^2)$  an equivalent data set can be reproduced. The difference between the information in  $(\bar{X}, S^2)$  and  $(X_1, X_2, \dots, X_n)$  is ancillary (of no use).
- $(\bar{X}, S^2)$  is actually minimal sufficient which means that its value can be obtained from any other sufficient statistic. Actually it is the sufficient statistic of the lowest dimension.
- $(\bar{X}, S^2)$  is complete, which means that any nontrivial function of  $(\bar{X}, S^2)$  carries some information about  $(\mu, \sigma^2)$ .

For a random sample from a normal population with mean  $\mu$  and variance  $\sigma^2$ , the statistics  $\bar{X}$  and  $S^2$  are independent. Actually,  $\bar{X}$  and  $S^2$  are independent only if the random sample is from a normal distribution.

- To better see the above fact, a large number of  $\bar{x}$  and  $s^2$  are generated from several distributions. The plots of  $\bar{x}$  vs.  $s^2$  are given in Figure 1. The statistical package Minitab was used to generate samples from five populations: normal, logistic, log-normal, exponential, and binomial. For each sample the mean  $\bar{x}$  was stored in a column and the standard deviation was stored in a column of the Minitab workplace. The contents of the two columns are then plotted.

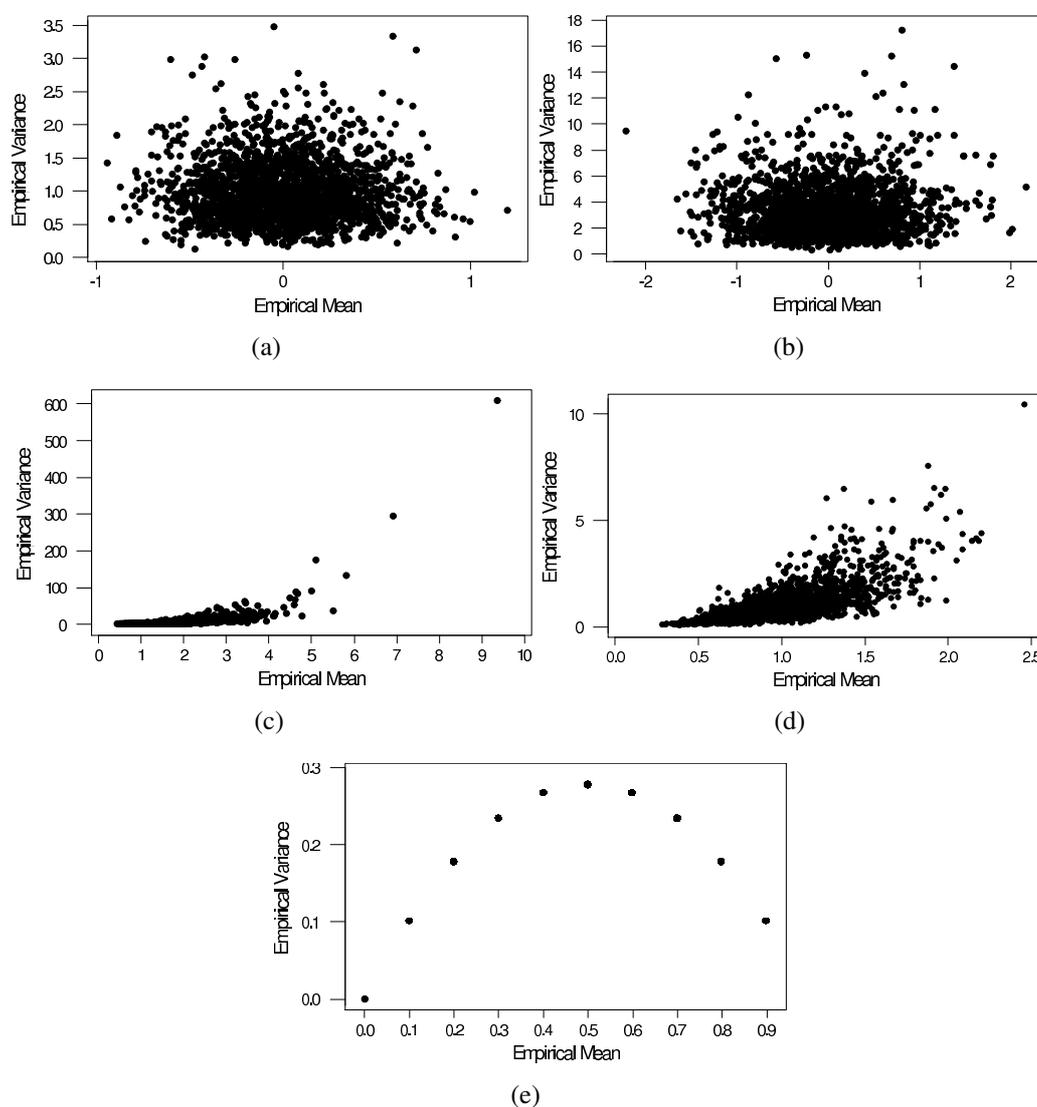


Figure 1: Plots of simulated empirical variances depending on the respective empirical means for data generated from normal (a), logistic (b), lognormal (c), exponential (d), and binomial (e) populations.

## Acknowledgements

We thank the referee for the careful reading of the manuscript and for the very helpful suggestions and comments, which lead to a significant improvement of our paper.

## References

- Al-Saleh, M. F. (2005). On the confusion about the  $(n - 1)$ -divisor of the standard deviation. *Journal of Probability and Statistical Science*, 5, 139-144.
- Bland, M. (1995). *An Introduction to Medical Statistics*. New York: Oxford University Press.
- Bluman, A. G. (2007). *Elementary Statistics: Step by Step Approach* (6th ed.). New York: McGraw Hill.
- Heath, D. (1995). *An Introduction to Experimental Design and Statistics for Biology*. London: CRC Press.
- Moore, D., and McCabe, G. (1993). *Introduction to the Practice of Statistics*. New York: W. H. Freeman and Company.
- Samuels, M., and Witmer, J. (1999). *Statistics for the Life Sciences*. New Jersey: Prentice Hall.
- Siegal, A., and Morgan, C. (1996). *Statistics and Data Analysis: An Introduction*. New York: John Wiley & Sons.

Authors' addresses:

Mohammad Fraiwan Al-Saleh  
Department of Mathematics  
College of Sciences  
University of Sharjah  
P O Box 27272, Sharjah  
United Arab Emirates  
E-Mail: malsaleh@sharjah.ac.ae

Adil Eltayeb Yousif  
Department of Mathematics and Physics  
Qatar University  
Qatar  
E-Mail: aeyousif@qu.edu.qa