

## Teaching Elementary Statistics Concepts Through $k\sigma$ Outliers

2006

**Michael J. Bossé**  
Dept. of Math & Science Education  
College of Education  
East Carolina University  
Greenville, NC 27858  
bossem@ecu.edu  
(252) 328-6219

**Frederick W. Morgan**  
Department of Mathematics  
Indiana University of Pennsylvania  
Indiana, PA 15705  
fwmorgan@iup.edu  
(724) 357-4765

Michael J. Bossé has a PhD in Curriculum and Instruction from the University of Connecticut and is an Associate Professor of Mathematics Education in the Department of Mathematic and Science Education at East Carolina University in Greenville, NC, 27858, USA. His research interests include cognition, learning, pedagogy, distance education and instruction through technology in mathematics.

### ABSTRACT

This paper demonstrates how the application of  $k\sigma$  outliers can assist in the instruction of introductory statistical concepts to high school and undergraduate students. Given a finite set of  $N$  elements of discrete numerical data on a range from  $a$  to  $b$ , this paper answers the questions: (1) Of the  $N$  elements, what is the largest number of elements,  $l$ , which can exceed  $k$  standard deviations from the mean and what is the smallest  $N$  for which  $l$  element(s) exceed(s)  $k$  standard deviations from the mean? (2) What is the relation of these findings with Chebyshev's inequality:  $P[|x - \mu| > k\sigma] \leq \frac{1}{k^2}$ . (3) Utilizing the initial bounds  $a$  and  $b$ , can  $l$  elements of  $c$  and  $d$  be generated ( $d < a < b < c$ ) which would be guaranteed to exceed  $k$  standard deviations from the mean of the new data set on the range from  $d$  to  $c$ ?

## Teaching Elementary Statistics Concepts Through $k\sigma$ Outliers

During the construction of conceptual understanding within elementary statistics, simple data sets may lead more efficiently to learning than overly complex data sets. As is well understood by professional educators, to ensure the learning of mathematical concepts, introductory instructional examples may differ significantly from more complex real-world investigations. In the process of teaching elementary statistical concepts, contrived and controlled data sets may often be utilized; simplified data sets can assist to control and direct students learning to specific concepts.

The mathematical ideas within this paper were born from simple statistical investigations in a mathematics education course for elementary education teachers. Combining standards prescribed in the *Principles and Standards for School Mathematics* (NCTM, 2000) – number and operations; algebra, geometry; measurement; and data analysis and probability – the instructor wished to provide the students with an intuitive understanding of “how far” from the mean a datum would need to be to exceed the three standard deviations from the mean of the data set. Having previously investigated some student-discovered, real-world data sets, none of the sets contained any elements which exceeded three standard deviations from the mean of the data set. The instructor understood that real-world data sets would only occasionally have such characteristics. Hoping to demonstrate a data set in which one element exceeded three standard deviations from the mean of the data, the instructor decided he would investigate how to construct such sets. After reviewing the pertinent literature, no publications were found which specifically addressed his concerns.

By demonstrating a series of questions which arose in this investigation, this paper discusses

ideational extensions of published papers addressing tangential matters to this investigation. Many of the findings within this investigation are directly applicable to the teaching of introductory statistical concepts and could be utilized in many ways by statistical educators to stimulate students interest in statistics.

The investigation herein followed an evolution of thought depicted through the following series of questions:

1. What is the fewest number of data elements necessary for one element to exceed three standard deviations?;
2. What is the fewest number of data elements necessary for one element to exceed  $k$  standard deviations?; and
3. What is the fewest number of data elements necessary for  $l$  element(s) to exceed  $k$  standard deviations?

### **Background Literature**

Sincich (1986) and Younger (1979) respectively define outliers as those values in a data set which exceed 3 and 4 standard deviations from the mean (Montgomery & Peck, 1982). Others have defined outliers as data values in a set which are outside the outer fences (Tukey, 1977). On small data sets, however, these definitions are problematic, particularly in the event where all data values are equal except for one (Shiffler, 1988). This investigation allows outliers to be defined on any range of standard deviations. Herein, we offer the notion of  $k\sigma$ - and  $ks$ -outliers, where one or more data elements exceed  $k$  standard deviations from the mean of the data set.

For any set of real numbers,  $x_1 \geq x_2 \geq \dots \geq x_n$ , where  $\bar{x} = \frac{\sum x}{n}$  and  $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ ,

Samuelson (1968) and Arnold (1974) have verified the following relationships:

$$x_1 \leq \bar{x} + s\sqrt{n-1} \quad \text{and} \quad x_n \geq \bar{x} - s\sqrt{n-1}.$$

Wolhowicz and Styan (1979) extend this relationship and note that for  $k = 1, 2, \dots, n$ ,

$$\bar{x} - s\sqrt{\frac{k-1}{n-k+1}} \leq x_k \leq \bar{x} + s\sqrt{\frac{n-k}{k}}$$

Connecting this notion to another perspective, Shiffler (1987) refines the findings of Huck, Cross, and Clark (1986) and demonstrates that for any data set of  $n$  elements, the largest possible  $Z$  score is  $Z < \sqrt{n-1}$  approaching the lower bound  $\sqrt{\frac{1}{n}}$ .

Shiffler (1988) makes a number of additional observations. He states that, for an ordered set,  $x_1 \leq x_2 \leq \dots \leq x_{n-1}$ , where an additional value is introduced to the set, with  $x_n$  being the greatest value of the set, the largest possible  $Z$  score for  $x_n$  is  $Z_{(n)} = (x_n - \bar{x})/S_n$ , which is maximized when  $S_{n-1}^2 = 0$  and reduces to  $Z_{(n)} = (n-1)/\sqrt{n}$ . However, a data set in which  $S_{n-1}^2 = 0$  can be recognized as the trivial case in which  $x_1 = x_2 = \dots = x_{n-1}$ . Shiffler then proceeds to use his findings to tabulate the maximum absolute  $Z$ -score for selected values for  $n$  and state that for certain combinations of  $n$  and  $Z$ , outliers cannot exist. Gray and Woodall (1994) employs the findings of Shiffler and other to demonstrate that consistent “with standardized  $Z$  scores in univariate data, the sizes of the standardized residuals and the internally Studentized residuals in regression are bounded.” (p. 113)

Notably, the preceding references are more concerned with bounds for values within the set and  $Z$  scores. This paper has asked a slightly different question regarding the number of elements in a set required to create Shiffler’s conditions. Students of introductory statistical concepts find beginning these investigations from the perspective of  $Z$  scores to be overly esoteric; simultaneously, they find questioning of the number of elements within a set to be more understandable.

## Selecting One $k\sigma$ Outlier

### Preliminary Statistics.

For a data set on a range  $[a, b]$ , we seek a new data element  $c$  such that  $a \leq b < c$  on the newly defined range  $[(a, b), c]$ . For simplicity, we will denote the following:

$$\mu_{[a, b]} = \mu; \quad \sigma_{[a, b]} = \sigma; \quad s_{[a, b]} = s; \quad \mu_{[(a, b), c]} = \mu'; \quad \sigma_{[(a, b), c]} = \sigma'; \quad \text{and} \quad s_{[(a, b), c]} = s'.$$

Since the different opinions exist regarding the definition for outliers, this paper will refrain from utilizing the commonly requisite 3 or 4 standard deviation from the mean and opt for more flexibility. Herein,  $c$  will be considered a “ $k\sigma$  outlier” if for any selected number of standard deviations,  $k$ ,  $|c - \mu'| > k\sigma'$ .

Thus, we seek  $c$  such that  $c > \mu' + k\sigma'$  or  $c > \mu' + ks'$ . In order to accomplish this task, we must determine the largest possible mean and standard deviation on a set of data on a range of  $[a, c]$ . The following begins this task.

**Maximum Mean.** Clearly, on a data set with range  $[a, b]$  the mean is maximized if all values are  $b$ . However, this would allow the trivial case where  $a = b$ . Disallowing this case, the data set on the range  $[a, b]$  which would have the greatest mean would be defined as form  $x_1 = x_2 = \dots = x_{n-1} = b$  and  $x_n = a$ . Therefore,  $\max[\mu] = \max[\bar{x}] = \frac{a + (n-1)b}{n}$ .

**Maximum Standard Deviation.** The data set with the greatest standard deviation on the range  $[a, b]$  will be one in which all elements are  $as$  and  $bs$ . We begin with the data set defined as:

$$x_1 = x_2 = \dots = x_m = a \quad \text{and} \quad x_{m+1} = x_{m+2} = \dots = x_{n-1} = x_n = b.$$

We can quickly determine that  $\sum x_i = ma + (n-m)b$  and  $\bar{x} = \frac{ma + (n-m)b}{n}$ .

Furthermore, the sample and population variances and standard deviations for this data set can be

determined.

$$\begin{aligned}
s^2 &= \frac{1}{n-1} \left[ m \left( a - \frac{ma + (n-m)b}{n} \right)^2 + (n-m) \left( b - \frac{ma + (n-m)b}{n} \right)^2 \right] \\
&= \frac{1}{n-1} \left[ m \left( \frac{na - ma - (n-m)b}{n} \right)^2 + (n-m) \left( \frac{nb - ma - (n-m)b}{n} \right)^2 \right] \\
&= \frac{1}{n-1} \left[ m \left( \frac{a(n-m) - (n-m)b}{n} \right)^2 + (n-m) \left( \frac{nb - ma - nb + bm}{n} \right)^2 \right] \\
&= \frac{1}{n-1} \left[ m \left( \frac{(n-m)(a-b)}{n} \right)^2 + (n-m) \left( \frac{-ma + bm}{n} \right)^2 \right] \\
&= \frac{1}{n-1} \left[ m \left( \frac{(n-m)(a-b)}{n} \right)^2 + (n-m) \left( \frac{-m(a-b)}{n} \right)^2 \right] \\
&= \frac{1}{(n-1)n^2} [m(n-m)^2(a-b)^2 + (n-m)m^2(a-b)^2] \\
&= \frac{1}{(n-1)n^2} [m(n-m)(a-b)^2[(n-m) + m]] \\
&= \frac{1}{(n-1)n^2} [m(n-m)(a-b)^2n] \\
&= \frac{m(n-m)(a-b)^2}{(n-1)n}
\end{aligned}$$

Therefore,  $s = \sqrt{\frac{m(n-m)(a-b)^2}{(n-1)n}} = (b-a)\sqrt{\frac{m(n-m)}{(n-1)n}}$  and

$$\sigma = \sqrt{\frac{m(n-m)(a-b)^2}{n^2}} = \frac{(b-a)\sqrt{m(n-m)}}{n}$$

The variance on this data set is greatest when there are an equal, or near equal number of  $as$  and  $bs$ , or when  $\begin{cases} 2m = n; & \text{if } n \text{ is even} \\ 2m + 1 = n; & \text{if } n \text{ is odd} \end{cases}$ . Therefore,  $\max[s]$  and  $\max[\sigma]$  will be investigated in two cases: where  $n$  is even, and where  $n$  is odd.

**Case I. If  $n = 2m$ :**

$$\begin{aligned}
 \max\{s\} &= (b-a)\sqrt{\frac{m(n-m)}{(n-1)n}} \\
 &= (b-a)\sqrt{\frac{m^2}{(n-1)n}} \\
 &= (b-a)\sqrt{\frac{\left(\frac{n}{2}\right)^2}{(n-1)n}} \\
 &= \frac{n(b-a)}{2\sqrt{(n-1)n}} \\
 &= \frac{(b-a)}{2}\sqrt{\frac{n}{n-1}}
 \end{aligned}$$

$$\max\{\sigma\} = \frac{n(a-b)}{2\sqrt{n^2}} = \frac{a-b}{2}$$

**Case II. If  $n = 2m + 1$ :**

$$\begin{aligned}
 \max\{s\} &= (b-a)\sqrt{\frac{m(n-m)}{(n-1)n}} \\
 &= (b-a)\sqrt{\frac{m(2m+1-m)}{(n-1)n}} \\
 &= (b-a)\sqrt{\frac{m(m+1)}{(n-1)n}} \\
 &= (b-a)\sqrt{\frac{\left(\frac{n-1}{2}\right)\left(\frac{n-1+2}{2}\right)}{(n-1)n}} \\
 &= (b-a)\sqrt{\frac{\left(\frac{n-1}{2}\right)\left(\frac{n+1}{2}\right)}{(n-1)n}} \\
 &= (b-a)\sqrt{\frac{(n-1)(n+1)}{4(n-1)n}} \\
 &= (b-a)\sqrt{\frac{n+1}{4n}} \\
 &= \frac{(b-a)\sqrt{n+1}}{2\sqrt{n}} \\
 &= \frac{(b-a)}{2}\sqrt{\frac{n+1}{n}}
 \end{aligned}$$

$$\begin{aligned}
 \max\{\sigma\} &= (b-a)\sqrt{\frac{m(n-m)}{n^2}} \\
 &= \frac{(b-a)}{n}\sqrt{m(2m+1-m)} \\
 &= \frac{(b-a)}{n}\sqrt{m(m+1)} \\
 &= \frac{(b-a)}{n}\sqrt{\left(\frac{n-1}{2}\right)\left(\frac{n-1+2}{2}\right)} \\
 &= \frac{(b-a)\sqrt{n^2-1}}{2n} \\
 &= \frac{(b-a)}{2}\frac{\sqrt{n^2-1}}{n}
 \end{aligned}$$

However, for  $n > 1$ , since

$$\sqrt{\frac{n}{n-1}} > \sqrt{\frac{n+1}{n}} \quad | \quad 1 > \sqrt{\frac{n^2-1}{n^2}}$$

$\max[s]$  and  $\max[\sigma]$  from a data set with an even number of elements will be greater than  $\max[s]$  and  $\max[\sigma]$  on a data set with an odd number of elements. Summarily, given a range  $[a, b]$  with  $n$  elements:  $\max[\mu] = \max[\bar{x}] = \frac{a + (n-1)b}{n}$ ;  $\max[s] = \frac{(b-a)}{2} \sqrt{\frac{n}{n-1}}$ ; and  $\max[\sigma] = \frac{a-b}{2}$ .

### Finding $c$ .

We can now restate our previous quest: For a data set on a range  $[a, b]$ , we seek  $c$  such that  $a \leq b < c$  and  $c > \mu' + k\sigma'$  or  $c > \mu' + ks'$ .

Let  $x_1, x_2, \dots, x_n$  be a data set on the interval  $[a, b]$ , where  $a < b$ . Thus, at least one of the  $n$  values differs from the others. Let  $\mu = \frac{\sum x_i}{n}$  and  $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$ . Introducing a new element,  $c$ , to the data, such that  $a < b < c$ , produces the new data set  $x_1, x_2, \dots, x_n, x_{n+1} = c$ , with

$$\mu' = \frac{x_1 + x_2 + \dots + x_n + c}{n+1} = \frac{n\mu + c}{n+1} \quad \text{and} \quad \sigma' = \sqrt{\frac{\sum (x_i - \mu')^2 + (c - \mu')^2}{n+1}}.$$

Clearly,  $c - \mu' = c - \frac{n\mu + c}{n+1} = \frac{nc + c - n\mu - c}{n+1} = \frac{n}{n+1}(c - \mu)$ . Therefore, if  $c$  is to be more than  $k$  standard deviations above  $\mu'$ , then  $c > \mu' + k\sigma' \sqrt{\frac{n+1}{n-k^2}}$ . (This result follows from

the proof for selecting  $l$  elements given below.) This carries the understanding that  $n \geq k^2 + 1$ .

Therefore, in order to find an appropriate  $c$  fulfilling the previous conditions, the original data set must have at least  $k^2 + 1$  data elements (This result is consistent with Shiffler (1987, 1988) and Gray & Woodall (1994).).

### Applications for Teaching.

The preceding discussions demonstrate a number of distinct findings which are applicable to teachers of introductory statistics throughout the K-14 grades:

(1.) Notably,  $c$  can be selected through a simple relationship between  $a$  and  $b$ , the original data limits. Using the conservative value of  $\mu = b$  and  $\sigma = \frac{b-a}{n}$  yields:

$$\begin{aligned}c &> b + k \left( \frac{b-a}{2} \right) \sqrt{\frac{n+1}{n-k^2}} \\c &> b + (k^2 + 1) \left( \frac{b-a}{2} \right) > b + k\sqrt{k^2 + 2} \left( \frac{b-a}{2} \right) \\c &> \frac{bk^2 + 3b - ak^2 - a}{2} = \frac{b(k^2 + 3) - a(k^2 + 1)}{2}\end{aligned}$$

Three observations resulting from these and previous relationships are of most importance to K-16 statistics educators. First, the inequality  $c > b + k \left( \frac{b-a}{2} \right) \sqrt{\frac{n+1}{n-k^2}}$  connotes that in order for a data set to have  $k\sigma$  one outlier,  $n \geq k^2 + 1$ . Therefore, the new data set, with the addition of the new datum must have at least  $k^2 + 2$  elements. Thus, it is impossible for a data set of only 36 elements to have a  $6k$  outlier (Compare to Shiffler (1987, 1988) and Gray & Woodall (1994).). See examples below.

Second, calculating the mean of either the old or new data set is not necessary to create  $k\sigma$  outliers! An inequality constructed from a simple relationship between lower and upper bound of the original data set can produce a new element for the extended data set which will be a  $k\sigma$  outlier. Furthermore, the final inequality,  $c > \frac{b(k^2 + 3) - a(k^2 + 1)}{2}$  or  $c > \frac{a(k^2 + 3) - b(k^2 + 1)}{2}$ , allows for instructors to arithmetically and extemporaneously construct values for  $c$  while observing or creating a data set within classroom instruction. This can open new dynamics in teaching statistical concepts and in providing robust classroom examples.

Third, introductory students of elementary statistics often have difficulty distinguishing between the needs and uses of either population or sample statistics over the other. In respect to this discussion, findings of (1.) above vary slightly: Given a set of data on a range from  $a$  to  $b$  ( $a < b$ ), in order for it to be possible to choose any  $c$  which is a  $k\sigma$  or  $ks$  outlier, the number of initial data elements prior to the introduction of the new value must be greater than or equal to  $\begin{cases} k^2 + 1; & \text{for population statistics} \\ k^2 + 2; & \text{for sample statistics} \end{cases}$  (Compare to Shiffler (1987, 1988) and Gray & Woodall (1994).). Introducing this insight to the classroom may lead students to investigate the nuances of population verses sample statistics in more detail.

*Examples:* For  $k = 3$ :  $c > \mu + 3\sigma\sqrt{\frac{n+1}{n-9}}$ ;  $n \geq 10$ ;  $c > 6b - 5a$

For  $k = 2$ :  $c > \mu + 2\sigma\sqrt{\frac{n+1}{n-4}}$ ;  $n \geq 5$ ;  $c > \frac{7b - 5a}{2}$

For  $k = 1$ :  $c > \mu + \sigma\sqrt{\frac{n+1}{n-1}}$ ;  $n \geq 2$ ;  $c > 2b - a$

For  $k = 6$ :  $c > \mu + 6\sigma\sqrt{\frac{n+1}{n-36}}$ ;  $n \geq 37$ ;  $c > \frac{39b - 37a}{2}$ .

(2.) Utilizing the preceding findings, examples and further discussions will increase student curiosity in this discussion. First, it is valuable to demonstrate how easily a teacher can extemporaneously create an example of a  $3\sigma$  outlier. Let us begin with the data set  $\{4, 5, 7, 7, 8, 9, 9, 10, 10, 10, 12, 13, 15, 15, 15, 15\}$ . The 16 elements within this set fulfill the condition that  $n \geq 10$ . Therefore, the above findings demonstrate that any  $c > 6b - 5b = 6(15) - 5(4) = 90 - 20 = 70$ , would exceed three standard deviations from the mean of the new data set which includes  $c$ .

Second, as demonstrated by Shiffler (1988), some small data set cannot have a  $k\sigma$  or  $ks$

outlier. For instance,  $x_i = 1,000,000$  within the data set  $\{0, 0, 0, 0, 0, 0, 0, 1,000,000\}$  cannot exceed three standard deviations from the mean, since  $n = 8$ .

Third, when  $n \geq 11$ ,  $x_n$  within any data set in the form  $x_1 = x_2 = \dots = x_{n-1} = x$  and  $x_n = x + y$ ,  $y > 0$  will exceed three standard deviations from the mean. For instance, on the data set  $\{10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10\}$ , any additional  $x_i > 10$  will be in the form  $x_i > \mu + 3\sigma$ . Notably, this includes  $x_i = 11$ ,  $10.1$ , or even  $10.0000001$ .

### ***l* Elements Exceeding $k\sigma$ from $\mu$**

Both the instructor and the student may be quick to now inquire regarding the possibility of having more than one  $k\sigma$  outlier. Therefore, we now extend the data set by  $l$  more elements of value  $c$ , making the cardinality of the new data set  $N = n + l$ .

#### **One Sided.**

For a finite, discrete data set on a range  $[a, b]$ , we seek  $l$  number of elements  $c$  such that  $|c - \mu'| > k\sigma'$ . (Initially, we consider  $c$  such that  $a < b < c$ . The more general case where  $c < a$  or  $b < c$  follows.)

Let  $x_1, x_2, \dots, x_n$  be a data set on the interval  $[a, b]$ , where  $a < b$ ,  $\mu = \frac{\sum x_i}{n}$ , and  $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$ . Introducing  $l$  new elements of equal value,  $c$ , to the data, such that  $a < b < c$ , produces the new data set of size  $N = n + l$ , (namely,  $x_1, x_2, \dots, x_n, c_1, c_2, \dots, c_l$ ) with  $\mu' = \frac{n\mu + lc}{n + l}$ ,  $\sigma' = \sqrt{\frac{\sum (x_i - \mu')^2 + l(c - \mu')^2}{n + l}}$ , and  $c - \mu' = \frac{n(c - \mu)}{n + l}$ . Therefore, if  $c$  is to be more than  $k$  standard deviations above  $\mu'$ , then

$$\begin{aligned}
c &> \mu' + k\sigma' \\
c - \mu' &> k\sigma' \\
c - \mu' &> k\sqrt{\frac{\sum(x_i - \mu')^2 + l(c - \mu')^2}{n + l}} \\
\frac{(c - \mu')^2(n + l)}{k^2} &> \sum(x_i - \mu')^2 + l(c - \mu')^2 \\
(c - \mu')^2 \left[ \frac{n + l - lk^2}{k^2} \right] &> \sum(x_i - \mu')^2 \\
(c - \mu')^2 \left[ \frac{n + l - lk^2}{k^2} \right] &> \sum(x_i - \mu')^2 \\
(c - \mu')^2 \left[ \frac{n + l - lk^2}{k^2} \right] &> \sum(x_i - \mu + \mu - \mu')^2 \\
(c - \mu')^2 \left[ \frac{n + l - lk^2}{k^2} \right] &> \sum(x_i - \mu)^2 + 2\sum(x_i - \mu)(\mu - \mu') + \sum(\mu - \mu')^2 \\
(c - \mu')^2 \left[ \frac{n + l - lk^2}{k^2} \right] &> n\sigma^2 + n(\mu - \mu')^2 \\
\left( c - \frac{n\mu + lc}{n + l} \right)^2 \left( \frac{n - lk^2 + l}{k^2} \right) &> n\sigma^2 + n \left( \mu - \frac{n\mu + lc}{n + l} \right)^2 \\
\frac{n^2(c - \mu)^2}{(n + l)^2} \left( \frac{n - lk^2 + l}{k^2} \right) &> n\sigma^2 + \frac{l^2 n(c - \mu)^2}{(n + l)^2} \\
\frac{n(c - \mu)^2}{(n + l)^2} \left( \frac{n^2 + nl - nlk^2 - lk^2}{k^2} \right) &> n\sigma^2 \\
\frac{(c - \mu)^2}{(n + l)^2} \left( \frac{(n - lk^2)(n + l)}{k^2} \right) &> \sigma^2 \\
\left( \frac{n - lk^2}{k^2} \right) \frac{(c - \mu)^2}{(n + l)} &> \sigma^2 \\
c - \mu &> k\sigma \sqrt{\frac{n + l}{n - lk^2}} \\
c &> \mu + k\sigma \sqrt{\frac{n + l}{n - lk^2}}
\end{aligned}$$

Therefore, in the more general one-sided case,  $|c - \mu'| > k\sigma'$ ,

$$c > \mu + k\sigma\sqrt{\frac{n+l}{n-lk^2}} \quad \text{and} \quad c < \mu - k\sigma\sqrt{\frac{n+l}{n-lk^2}}.$$

In summary, in order to find  $l$  appropriate  $c$ 's outside  $\mu \pm k\sigma$ , the original data set must have at least  $lk^2 + 1$  data elements. Hence,  $n \geq lk^2 + 1$  and  $N \geq lk^2 + 1 + l$ . Notably,  $l$  is maximized in respect to  $n$  and  $N$  when  $c_1 = c_2 = \dots = c_r$ .

### Two-Sided.

For a finite set of discrete data on a range  $[a, b]$ , we seek  $l-r$  number of elements  $d$  and  $r$  number of elements  $c$  such that  $d < \mu' - k\sigma' < \mu' + k\sigma' < c$ . This would produce the new data set  $d_1, d_2, \dots, d_{l-r}, x_1, x_2, \dots, x_n, c_1, c_2, \dots, c_r$  where  $d_1 = d_2 = \dots = d_{l-r}$  and  $c_1 = c_2 = \dots = c_r$ . Again, these conditions will maximize  $l$  in respect to  $n$  and  $N$ .

Repeating the previous analysis using  $r = \frac{l}{2}$  produces

$$d < \mu - k\sigma\sqrt{\frac{n+l}{n-lk^2-1+l}} \quad \text{and} \quad c > \mu + k\sigma\sqrt{\frac{n+l}{n-lk^2-1+l}}.$$

Therefore, in order to seek  $l-r$  number of elements  $d$  and  $r$  number of elements  $c$  such that  $d < \mu' - k\sigma' < \mu' + k\sigma' < c$ ,  $n \geq lk^2 + 1 - l$  and  $N \geq lk^2 + 1$ .

### Application for Teaching.

For many students the mathematical arguments which immediately preceded may be overly esoteric. Nevertheless, applications of these findings make it again relatively easy for teachers to develop data sets with any number of  $k\sigma$  outliers on one or both sides of the bounds of the data set. This can provide the instructor with valuable opportunities for the investigation of data sets which demonstrate concepts which the instructor wishes to present.

From the preceding, for a finite, discrete data set of  $n$  elements on a range  $[a, b]$ , we seek  $l$  number of elements  $c$  such that  $|c - \mu'| > k\sigma'$ , where  $\mu$  and  $\sigma$  are determined from the existing set of  $n$  elements and where  $\mu'$  and  $\sigma'$  are determined from the newly constructed set of  $n + l$  elements. Finding appropriate values for  $c$  can be simplified to the following. Notably,  $c$  and/or  $d$  can be selected through a simple relationship between  $a$  and  $b$ , the original data limits, where  $\frac{b-a}{2}$  is the maximum  $\sigma$  for a finite population on the interval  $[a, b]$ .

**One-Sided.**

$$\begin{aligned} c < \mu - k\sigma\sqrt{\frac{n+l}{n-lk^2}} & \quad \text{and} \quad c > \mu + k\sigma\sqrt{\frac{n+l}{n-lk^2}} \\ c < a - k\left(\frac{b-a}{2}\right)\sqrt{\frac{n+l}{n-lk^2}} & \quad c > b + k\left(\frac{b-a}{2}\right)\sqrt{\frac{n+l}{n-lk^2}} \end{aligned}$$

**Two-Sided.**

$$\begin{aligned} d < \mu - k\sigma\sqrt{\frac{n+l}{n-lk^2-1+l}} & \quad \text{and} \quad c > \mu + k\sigma\sqrt{\frac{n+l}{n-lk^2-1+l}} \\ d < a - k\left(\frac{b-a}{2}\right)\sqrt{\frac{n+l}{n-lk^2-1+l}} & \quad c > b + k\left(\frac{b-a}{2}\right)\sqrt{\frac{n+l}{n-lk^2-1+l}} \end{aligned}$$

**Connection to Chebyshev's Inequality**

History has recorded the significance of Chebyshev's inequality. Applications of Chebyshev's inequality reach down into even elementary statistical studies. Proofs of Chebyshev's inequality, however, are necessarily reserved for more advanced mathematical and statistical studies. Employing the previous findings, the following discussions demonstrate an empirical demonstration of the veracity of Chebychev's inequality which may be understood by stronger high school and undergraduate students who do not have a calculus background.

With  $l$  maximized in respect to  $n$  and  $N$ , the following relationships have been developed:

$$\begin{cases} \text{One Sided: } n \geq lk^2 + 1; & N \geq lk^2 + 1 + l \\ \text{Two Sided: } n \geq lk^2 + 1 - l; & N \geq lk^2 + 1 \end{cases}$$

Furthermore, given element  $c$  from a finite, discrete distribution,

$$P[|c - \mu'| > k\sigma'] \leq \frac{l}{N} \quad \text{and} \quad P[|c - \mu'| < k\sigma'] \geq 1 - \frac{l}{N}.$$

### Empirical Discussion.

The adjacent table empirically demonstrates the findings above.

The  $l$  column represents the number of elements  $c$  in a finite set of discrete data in which  $|c - \mu| > k\sigma$ . The table demonstrates data for both the one-sided and two-sided cases. The table demonstrates that for 10 elements of a data set to exceed 2 standard deviations from the mean on one side, the data set must have at least 51 elements; and for 12 elements of a data set to be outside of the range  $[\mu - 3\sigma, \mu + 3\sigma]$ , at least 109 elements are needed in the set.

$l$	$N$					
	$1\sigma$		$2\sigma$		$3\sigma$	
	1-S	2-S	1-S	2-S	1-S	2-S
1	3		6		11	
2	5	3	11	9	21	19
3	7	4	16	13	31	28
4	9	5	21	17	41	37
5	11	6	26	21	51	46
6	13	7	31	25	61	55
7	15	8	36	29	71	64
8	17	9	41	33	81	73
9	19	10	46	37	91	82
10	21	11	51	41	101	91
11	23	12	56	45	111	100
12	25	13	61	49	121	109
13	27	14	66	53	131	118
14	29	15	71	57	141	127
15	31	16	76	61	151	136
16	33	17	81	65	161	145
17	35	18	86	69	171	154
18	37	19	91	73	181	163
19	39	20	96	77	191	172
20	41	21	101	81	201	181
21	43	22	106	85	211	190
22	45	23	111	89	221	199
23	47	24	116	93	231	208
24	49	25	121	97	241	217
25	51	26	126	101	251	226
26	53	27	131	105	261	235
27	55	28	136	109	271	244
28	57	29	141	113	281	253
29	59	30	146	117	291	262
30	61	31	151	121	301	271

The next table reverses the observations. In a set of discrete data, for a given number of data elements  $N$ , what is the largest number  $l$  of elements  $c$  for which  $|c - \mu| > k\sigma$ ? The table depicts that for a set of 256 elements no more than 3 elements  $c$  can fulfill  $|c - \mu| > 6\sigma$  on one side of the mean; and for a set of 1024 elements no more than 112 elements can exceed  $\pm 3\sigma$  from  $\mu$  on two sides of the mean.

N	k											
	1		2		3		4		5		6	
	1-S	2-S	1-S	2-S	1-S	2-S	1-S	2-S	1-S	2-S	1-S	2-S
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
3	1	2	0	0	0	0	0	0	0	0	0	0
4	1	3	0	0	0	0	0	0	0	0	0	0
5	2	4	0	0	0	0	0	0	0	0	0	0
6	2	5	1	0	0	0	0	0	0	0	0	0
7	3	6	1	0	0	0	0	0	0	0	0	0
8	3	7	1	0	0	0	0	0	0	0	0	0
9	4	8	1	2	0	0	0	0	0	0	0	0
10	4	9	1	2	0	0	0	0	0	0	0	0
11	5	10	2	2	1	0	0	0	0	0	0	0
12	5	11	2	2	1	0	0	0	0	0	0	0
13	6	12	2	3	1	0	0	0	0	0	0	0
14	6	13	2	3	1	0	0	0	0	0	0	0
15	7	14	2	3	1	0	0	0	0	0	0	0
16	7	15	3	3	1	0	0	0	0	0	0	0
17	8	16	3	4	1	0	1	0	0	0	0	0
18	8	17	3	4	1	0	1	0	0	0	0	0
19	9	18	3	4	1	2	1	0	0	0	0	0
20	9	19	3	4	1	2	1	0	0	0	0	0
21	10	20	4	5	2	2	1	0	0	0	0	0
22	10	21	4	5	2	2	1	0	0	0	0	0
23	11	22	4	5	2	2	1	0	0	0	0	0
24	11	23	4	5	2	2	1	0	0	0	0	0
25	12	24	4	6	2	2	1	0	0	0	0	0
26	12	25	5	6	2	2	1	0	0	0	0	0
27	13	26	5	6	2	2	1	0	0	0	0	0
28	13	27	5	6	2	3	1	0	0	0	0	0
29	14	28	5	7	2	3	1	0	0	0	0	0
30	14	29	5	7	2	3	1	0	0	0	0	0
31	15	30	6	7	3	3	1	0	0	0	0	0
32	15	31	6	7	3	3	1	2	0	0	0	0
33	16	32	6	8	3	3	2	2	1	0	0	0
63	31	62	12	15	6	6	3	4	1	2	0	0
64	31	63	12	15	6	7	3	4	1	2	0	0
65	32	64	12	16	6	7	4	4	2	2	1	0
127	63	126	25	31	11	13	7	7	3	4	1	2
128	63	127	25	31	11	13	7	7	3	4	1	2
129	64	128	25	32	11	13	8	7	4	4	2	2
255	127	254	50	63	24	27	15	16	7	9	3	6
256	127	255	51	63	24	27	15	17	7	9	3	6
257	128	256	51	64	24	27	16	17	8	9	4	6
1023	511	1022	204	255	101	112	63	67	31	40	15	27
1024	511	1023	204	255	101	112	63	67	31	40	15	27
1025	512	1024	204	256	101	112	64	67	32	40	16	27

The accompanying table provides the results of  $P[X > \mu + k\sigma] = \frac{l}{N}$  and  $P[|X - \mu| > k\sigma] = \frac{l}{N}$  for selected values of  $N$ ,  $1 \leq N \leq 1025$ . Through observation, one may recognize that each column approaches a limit and that the 2-sided scenario approaches this limit more rapidly. The following section demonstrates that these probabilities approach the probabilities derived by Chebyshev's inequality. Furthermore, each probability within the table is a tighter approximation than provided by Chebyshev's inequality.

N	$\frac{l}{N}$					
	1		2		3	
	1-Sided	2-Sided	1-Sided	2-Sided	1-Sided	2-Sided
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0.3333	0.6667	0	0	0	0
4	0.25	0.75	0	0	0	0
5	0.4	0.8	0	0	0	0
6	0.3333	0.8333	0.1667	0	0	0
7	0.4286	0.8571	0.1429	0	0	0
8	0.375	0.875	0.125	0	0	0
9	0.4444	0.8889	0.1111	0.2222	0	0
10	0.4	0.9	0.1	0.2	0	0
11	0.4545	0.9091	0.1818	0.1818	0.0909	0
12	0.4167	0.9167	0.1667	0.1667	0.0833	0
13	0.4615	0.9231	0.1538	0.2308	0.0769	0
14	0.4286	0.9286	0.1429	0.2143	0.0714	0
15	0.4667	0.9333	0.1333	0.2	0.0667	0
16	0.4375	0.9375	0.1875	0.1875	0.0625	0
17	0.4706	0.9412	0.1765	0.2353	0.0588	0
18	0.4444	0.9444	0.1667	0.2222	0.0556	0
19	0.4737	0.9474	0.1579	0.2105	0.0526	0.1052
20	0.45	0.95	0.15	0.2	0.05	0.1
21	0.4762	0.9524	0.1905	0.2381	0.0952	0.0952
22	0.4545	0.9545	0.1818	0.2273	0.0909	0.0909
23	0.4783	0.9565	0.1739	0.2174	0.0870	0.0870
24	0.4583	0.9583	0.1667	0.2083	0.0833	0.0833
25	0.48	0.96	0.16	0.24	0.08	0.08
26	0.4615	0.9615	0.1923	0.2308	0.0769	0.0769
27	0.4815	0.9630	0.1852	0.2222	0.0741	0.0741
28	0.4643	0.9643	0.1786	0.2143	0.0714	0.1071
29	0.4828	0.9655	0.1724	0.2414	0.0690	0.1034
30	0.4667	0.9667	0.1667	0.2333	0.0667	0.1
31	0.4839	0.9677	0.1935	0.2258	0.0968	0.0967
32	0.4688	0.9688	0.1875	0.2188	0.0936	0.0938
33	0.4848	0.9697	0.1818	0.2424	0.0909	0.0909
63	0.4921	0.9841	0.1905	0.2381	0.0952	0.0952
64	0.4844	0.9844	0.1875	0.2344	0.0936	0.1094
65	0.4923	0.9846	0.1846	0.2462	0.0923	0.1077
127	0.4961	0.9921	0.1969	0.2441	0.0866	0.1024
128	0.4922	0.9922	0.1953	0.2422	0.0860	0.1016
129	0.4961	0.9922	0.1938	0.2481	0.0853	0.1008
255	0.4980	0.9961	0.1961	0.2471	0.0941	0.1059
256	0.4961	0.9961	0.1992	0.2461	0.0938	0.1055
257	0.4981	0.9961	0.1984	0.2490	0.0934	0.1051
1023	0.4995	0.9990	0.1994	0.2493	0.0987	0.1095
1024	0.4990	0.9990	0.1992	0.2490	0.0986	0.1094
1025	0.4995	0.9990	0.1990	0.2498	0.0985	0.1093

**Theoretical Discussion.**

**One-Sided.** From the results above, for a population of size  $N$  to have  $l$  observations more than  $k$  standard deviations above (or below) the mean,  $N \geq lk^2 + l + 1$ . It then follows that  $N - 1 \geq l(k^2 + 1)$  and  $\frac{N - 1}{k^2 + 1} \geq l$ . Therefore, for a randomly selected value in the population,  $X$ ,

$$P[X > \mu + k\sigma] \leq \frac{l}{N} \leq \frac{\frac{N - 1}{k^2 + 1}}{N} = \frac{1}{k^2 + 1} \left(1 - \frac{1}{N}\right).$$

As the population size increases without bound, this probability approaches the one-sided Chebyshev inequality,

$$P[X \geq \mu + k\sigma] \leq \frac{1}{k^2 + 1}.$$

**Two-Sided.** Similarly, for a population of size  $N$  to have  $l$  observation(s) more than  $k$  standard deviations above or below the mean (two-side),  $N \geq lk^2 + 1$  and  $\frac{N - 1}{k^2} \geq l$ . Thus,

$$P[|X - \mu| > k\sigma] \leq \frac{l}{N} \leq \frac{\frac{N - 1}{k^2}}{N} = \frac{1}{k^2} \left(1 - \frac{1}{N}\right).$$

Again, as  $N$  increases without bound, this probability also approaches Chebyshev's inequality,

$$P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}.$$

**Summary.** While both the one-sided and two-sided probabilities approach Chebyshev's inequality, for finite populations, both versions provide tighter probabilities than does Chebyshev's inequality. This can be denoted by the following inequality.

$$P[|X - \mu| > k\sigma] \leq \underbrace{\frac{1}{k^2 + 1} \left(1 - \frac{1}{N}\right)}_{1\text{-Sided}} \leq \underbrace{\frac{1}{k^2} \left(1 - \frac{1}{N}\right)}_{2\text{-Sided}} \leq \underbrace{\frac{1}{k^2}}_{\text{Chebyshev}}$$

## Conclusion

The preceding results stipulate the minimum size of a finite population in order for a specified number of values in the population to exceed  $k$  standard deviations from the mean. The conditions are almost identical for samples. Statistical analysis identifying “outliers” as exceeding  $k$  standard deviations from the mean should be aware that, for certain sample sizes and choices of  $k$ , “outliers” are impossible. For teachers who wish to demonstrate values exceeding  $k\sigma$  (or  $ks$ ) from the mean, the results above are useful for determining values that can be added to an original population (or sample) that meet the  $\pm k\sigma$  criterion.

The initial results reduce to modestly tighter Chebyshev bounds for finite populations. As the size of the population increases without bound, the standard Chebyshev inequalities are obtained.

## References

- Arnold, Barry C. (1974). Schwarz, Regression and Extreme Diviance. *The American Statistician*, 28, pp. 22-23
- Gray, J. Brian & Woodall, William, H. (1994). The Maximum Size of Standardized and Internally Studentized Residuals in Regression Analysis. *The American Statistician*. Vol. 48, No. 2, pp. 111-113.
- Huck, S. W., Cross, T. L. & Clark, S. B. (1986). Overcoming Misconceptions About Z-Scores. *Teaching Statistics*, 8, pp. 38-40
- Montgomery, D. C. & Peck, E. A. (1982). *Introduction to Linear Regression Analysis*. New York: John Wiley

- National Council of Teachers of Mathematics [NCTM] (2000). *Principles and standards for school mathematics*. Reston, VA: The National Council of Teachers of Mathematics.
- Shiffler, Ronald E. (1988). Maximum Z Scores and Outliers. *The American Statistician*, Vol. 42, No. 1, pp. 79-80
- Shiffler, Ronald E. (1987). Bounds on the Maximum Z-Score. *Teaching Statistics*, Vol. 9, pp. 80-81
- Samuelson, Paul A. (1968). How Deviant Can You Be? *Journal of the American Statistical Association*, 63, pp. 1522-1525
- Sincich, T. (1986). *Business Statistics by Example* (2nd ed.). San Francisco: Dellen
- Tukey, J. W. (1977). *Explanatory Data Analysis*. Reading, MA: Addison-Wesley
- Wolkowicz, Henry & Styan, George, P. H. (1979). Extensions of Samuelson's Inequality. *The American Statistician*, Vol. 33, No. 3, pp. 143-144
- Younger, M. S. (1979). *A Handbook for Linear Regression*. North Scituate, MA: Duxbury Press