

## Variance and Standard Deviation...More Notes

### I. Different Ways to Express Variance

Given population data  $x_1, x_2, \dots, x_n$ , we can express the population variance  $\sigma^2$  in the following algebraically equivalent ways:

1.  $\sigma^2 = \frac{\sum(x-\bar{x})^2}{n}$
2.  $\sigma^2 = \frac{1}{n} \sum x^2 - \bar{x}^2$
3.  $\sigma^2 = \frac{1}{n^2} \sum_{i < j} (x_i - x_j)^2$

Similarly, given sample data  $x_1, x_2, \dots, x_n$ , we can express the sample variance  $s^2$  in the following algebraically equivalent ways:

1.  $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$
2.  $s^2 = \frac{\sum x^2 - (\sum x)^2/n}{n-1} = \frac{1}{n-1} \sum x^2 - \frac{n}{n-1} \bar{x}^2$
3.  $s^2 = \frac{1}{n(n-1)} \sum_{i < j} (x_i - x_j)^2$   
 $= [\frac{1}{2} \cdot \text{average of the pairwise squared differences}]$

### II. Bounds on Variance and Standard Deviation

• **Theorem.** Let  $\sigma^2$  represent the “population” variance of an ordered data set  $(x_1, x_2, \dots, x_n)$  of size  $n \geq 2$ . Let the range  $r = x_n - x_1$ . Then

$$\sigma^2 \leq \frac{\text{range}^2}{4} \text{ and } \sigma \leq \frac{\text{range}}{2}.$$

• **Theorem.** For positive integer  $n$ , let  $\{x_1, x_2, \dots, x_n\}$  be a finite population with  $x_1 \leq \dots \leq x_n$  and which has variance  $\sigma^2$ . Let  $r$  denote the range of the population, that is,  $r = x_n - x_1$ . Then

$$\sigma^2 \geq \frac{r^2}{2n} \text{ and } \sigma \geq \frac{r}{\sqrt{2n}}$$

• **Lemma 1.**  $\frac{r^2}{2n} \leq \sigma^2 \leq \frac{r^2}{4}$  or  $\frac{r}{\sqrt{2n}} \leq \sigma \leq \frac{r}{2}$   
The above is equivalent to  $2\sigma \leq r \leq \sqrt{2n} \sigma$ .

• **Lemma 2.**  $\frac{r^2}{2(n-1)} \leq s^2 \leq \frac{r^2}{4} \left(\frac{n}{n-1}\right)$  and  $\frac{r}{\sqrt{2(n-1)}} \leq s \leq \frac{r}{2} \sqrt{\frac{n}{n-1}}$ .

Equivalently,  $2\sqrt{\frac{n-1}{n}} s \leq r \leq \sqrt{2(n-1)} s$

**Example.** Suppose a set of sample data of size  $n = 20$  had a minimum of 180 and a maximum of 210. Find bounds on the sample standard deviation  $s$ .

Solution. Since the range is  $r = 210 - 180 = 30$ , we have  $\frac{30}{\sqrt{2(19)}} \leq s \leq \frac{30}{2} \sqrt{\frac{20}{19}}$ . In other words,  $4.86 < s < 15.39$ .

**Example** (of an ill-posed statistics problem ). Suppose a set of sample data of size  $n$  provided the following statistics:  $n = 20$ , minimum = 180, maximum = 210,  $\bar{x} = 200$ , and  $s = 4$ . Find the  $z$ -score for data value 208.

Although one can readily calculate  $z = \frac{208-200}{4} = 2$ , this is an ill-posed problem since there can never exist a data set that satisfies the conditions  $n = 20$ , minimum = 180, maximum = 210,  $\bar{x} = 200$ , and  $s = 4$  because  $s > 4.86$  as shown above.

Also note that, if the problem had used  $s = 16$ , the premise would again be wrong since it was shown above that  $s < 15.39$ .

### III. Chebyshev's Inequality

For **any** data set or finite population, at least  $1 - \frac{1}{k^2}$  of the data fall within  $k$  standard deviations of the mean. [Note : use  $k > 1$ .]

### IV. Empirical Rule for Data with Bell-Shaped Distribution

- (a) Approximately 68% of the data fall within one standard deviation of the mean.
- (b) Approximately 95% of the data fall within two standard deviations of the mean.
- (b) Approximately all of the data fall within three standard deviations of the mean.

### V. Bounds for Finite Populations of Size $n$

- (a) For any data set, all the data fall within  $\sqrt{2n}$  standard deviations of the mean.
- (b) For a data set with a symmetric distribution, all the data fall within  $\sqrt{\frac{n}{2}}$  standard deviations of the mean.