

### 3.4 Measures of Variability

#### Measures of Spread

1. Range
2. IQR
3. Variance and Standard Deviation

**Definition:** The Range = Highest Value - Lowest Value

**Example:** You are going to buy a car. You have narrowed your choice to two different models. You like them both equally well so you will choose based on fuel economy. To arrive at the average mpg of each model the car manufacturer randomly chose 8 cars of each model and drove them to find their fuel economy. The number on the sticker is the average of the 8 cars.

If the raw data is as follows:

Model #1: 25,10,35,12,28,36,24,30 (mpg)

Model #2: 25,25,25,25,25,25,25,25 (mpg)

Then both models would have the same sticker value of 25 mpg (the average of the eight cars). It would appear the two models have the same fuel economy, but note the fuel economy of model #1 is very inconsistent. With model #1 you could get a very efficient car or a very inefficient car. Model #2 on the other hand is very consistent. The average is not giving the buyer enough information to make an informed choice. If you calculate the range you will find,

Range Model #1=  $36-12=24$  mpg

Range Model #2= $25-25=0$  mpg

Now it is clear that model #1 is more inconsistent than model #2. But what if the data were as follows:

Model #1: 25,10,35,12,28,36,24,30 (mpg)

Model #2: 25,25,25,25,25,25,45,5 (mpg)

Now they still both have an average fuel economy of 25 mpg, but

Range Model #1=  $36-12=24$  mpg

Range Model #2= $45-5=40$  mpg

This gives the FALSE impression that model #2 is less consistent. One disadvantage in using the range is that it is based on the two most extreme data values. Because of this it can often be misleading.

**Definition:** The first quartile  $q_1$  is the median of the lower half of the data.

The third quartile  $q_3$  is the median of the upper half of the data.

The interquartile range  $IQR = q_3 - q_1$

**Example:** Find the IQR for the data:

Model #1: 25,10,35,12,28,36,24,30 (mpg)

Model #2: 25,25,25,25,25,25,45,5 (mpg)

First sort the data

Model #1: 10,12,24,25,28,30,35,36 (mpg)

Model #2: 5,25,25,25,25,25,45 (mpg)

$$\begin{array}{cccc} 10, 12, & & 24, 25, & & 28, 30, & & 35, 36 \\ & \uparrow & & \uparrow & & \uparrow & \\ q_1 = \frac{12 + 24}{2} = 18 & & q_2 = \frac{25 + 28}{2} = 26.5 & & q_3 = \frac{30 + 35}{2} = 32.5 \end{array}$$

$$\begin{array}{cccc} 5, 25, & & 25, 25, & & 25, 25, & & 25, 45 \\ & \uparrow & & \uparrow & & \uparrow & \\ q_1 = \frac{25 + 25}{2} = 25 & & q_2 = \frac{25 + 25}{2} = 25 & & q_3 = \frac{25 + 25}{2} = 25 \end{array}$$

For model #1  $IQR = 32.5 - 18 = 14.5$

For model #2  $IQR = 25 - 25 = 0$

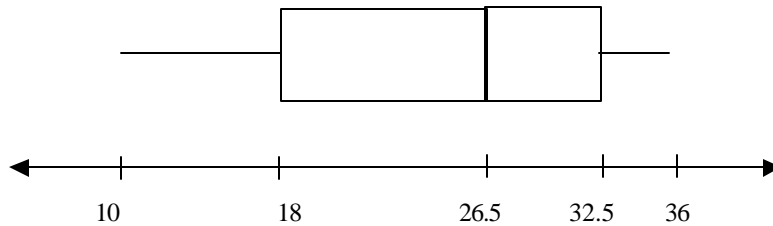
In this case the IQR gives a more accurate feel for the spread than the Range did. The IQR measures the range of the middle half of the data. It has the advantage that a few extreme data points do not affect it. It has the disadvantage that it ignores half of the data (the top and bottom quartile).

**Definition:** The set  $\{s, q_1, m, q_3, L\}$  where  $s$  is the smallest data value,  $m$  is the median, and  $L$  is the largest value, is called the 5-Number Summary.

**Example:** The 5-number summary for model #1 in the example above is  $\{10, 18, 26.5, 32.5, 36\}$

Definition: A Box and Whisker Plot is a graphical display of the 5 number summary.

**Example:** The Box and Whisker plot representing the 5-number summary above is:



### Standard Deviation

Definition: The deviation from the mean is the difference between the data point and the mean.

**Example:** Find the deviation of each number in the data set  $\{2, 5, 10, 20, 23\}$

Note  $\bar{x} = 12$

Data Point	Deviation
2	$2 - 12 = -10$
5	$5 - 12 = -7$
10	$10 - 12 = -2$
20	$20 - 12 = 8$
23	$23 - 12 = 11$
<b>Total</b>	<b>0</b>

The sum of the deviations from the mean is always zero, so the average of the deviations from the mean is always zero. To get a better estimate of the spread we might try squaring all the deviations from the mean to avoid the cancellation we are seeing above.

Definition: Given a sample of  $n$  measurements  $x_1, x_2, \dots, x_n$  with sample mean  $\bar{x}$ , the

$$\text{sample variance } s^2 \text{ is } s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

The sample variance is roughly speaking the average of the squared deviations from the mean. We divide by  $n - 1$  instead of  $n$  for reasons that are beyond the scope of this course.

**Example:**

Data Point	Deviation	(Deviation) <sup>2</sup>
2	2 - 12 = -10	100
5	5 - 12 = -7	49
10	10 - 12 = -2	4
20	20 - 12 = 8	64
23	23 - 12 = 11	121
<b>Total</b>	<b>0</b>	<b>338</b>

$$s^2 = \frac{338}{5-1} = \frac{338}{4} = 84.5$$

Definition: Given a sample of  $n$  measurements  $x_1, x_2, \dots, x_n$  with sample mean  $\bar{x}$ , the sample standard deviation  $s$  is given by

$$s = \sqrt{s^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$$

**Example:** For the example above:  $s = \sqrt{84.5} \approx 9.19$

Definition: Suppose the set of all  $N$  measurements  $x_1, x_2, \dots, x_N$  in a population of  $N$  elements is given. If the population mean is  $m$ , then the population variance  $s^2$  is  $s^2 = \frac{(x_1 - m)^2 + (x_2 - m)^2 + \dots + (x_n - m)^2}{N}$  and the population standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{(x_1 - m)^2 + (x_2 - m)^2 + \dots + (x_n - m)^2}{N}}$$

**Example:** For the example above:  $s^2 = \frac{338}{5} = 67.6$  and  $s = \sqrt{67.6} \approx 8.22$

**Example:** Use a calculator to find  $s, s$  for  $\{2, 3, 3, 5, 5, 5, 9\}$ .

Answer:  $s_x = 2.128, \bar{x} = 4.57$

The notes above are for Math 108, Math for the Modern World using *Mathematics in Life, Society and the World 2<sup>nd</sup> edition* by Parks, Musser, Burton, and Siebler. Prentice Hall 2000.