

## Lecture 2

# Describing Data with Numerical Measures

**Dr Md Rifat Ahmmad Rashid**  
**Assistant Professor**  
**Dept. of CSE**  
**East West University**

# Describing Data with Numerical Measures

- Graphical methods may not always be sufficient for describing data.
- **Numerical measures** can be created for both populations and samples.
  - A **parameter** is a numerical descriptive measure calculated for a population.
  - A **statistic** is a numerical descriptive measure calculated for a sample.

# Central Tendency (*Center*) and Dispersion (*Variability*)

- A **distribution** is an ordered set of numbers showing how many times each occurred, from the lowest to the highest number or the reverse
- **Central tendency**: measures of the degree to which scores are clustered around the mean of a distribution
- **Dispersion**: measures the fluctuations (variability) around the characteristics of central tendency

# Central Tendency and Variability

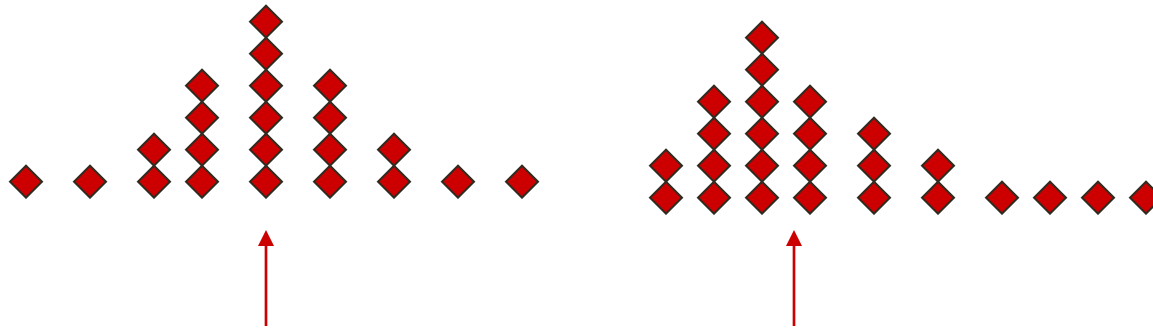
- Central tendency describes the central point of the distribution, and variability describes how the scores are scattered around that central point.
- Together, central tendency and variability are the two primary values that are used to describe a distribution of scores.

# Central tendency

## - Measures of Center

# Measures of Center

- A measure along the horizontal axis of the data distribution that locates the center of the distribution.



# Arithmetic Mean or Average

- The **mean** of a set of measurements is the sum of the measurements divided by the total number of measurements.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

where  $n$  = number of measurements

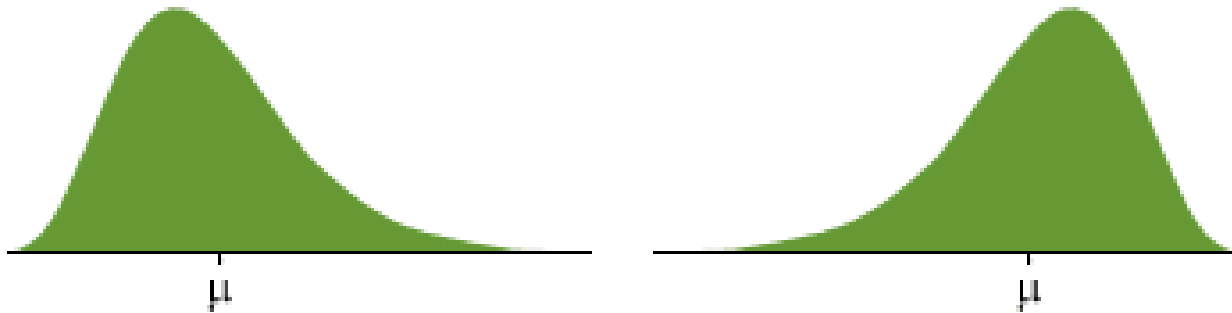
$\sum x_i$  = sum of all the measurements

# Mean, Median, Mode

- mean – also known as the arithmetic mean or average. Calculated by adding the scores and dividing by the number of scores
- median – the number in the middle when the data is arranged in ascending or descending order
- mode – the most frequent. If two numbers occur the same amount of times the set is bimodal. If all the same, more than one mode.



Skewness – data is skewed if it is not symmetric and extends more to one side than the other.



- The one on the left is positively skewed.
- The one on the right is negatively skewed.

# Mean

- $\Sigma$  - denotes *summation* of a set of values
- $x$  – is the *variable* usually used to represent the individual data values
- $n$  – represents the *number of values in a sample*
- $N$  – represents the *number of values in a population*

$$\bar{X} = \frac{\sum x}{n} \text{ is the } \textit{mean of a set of sample values}$$

$$\mu = \frac{\sum x}{N} \text{ is the } \textit{mean of all values in a population}$$

$$\bar{X} = \frac{\sum(f(x))}{\sum f} \text{ mean from a frequency table}$$

# Median

- The method for finding the median is slightly different depending if the total  $f$  is even or odd.
- If Total  $f$  is **odd**, then there is one middle value. To find it, calculate  $(\text{Total } f)/2$  and round up.
- If Total  $f$  is **even**, then there are two middle values and the median is the average of these. To find the two middle values, calculate  $\text{total } f / 2$ . The two middle scores are the ones corresponding to that number and the next. Once you have the two scores, the median is the average of them.

Below is the weight of some of our favorite meals.  
Compute the mean, median, and mode.

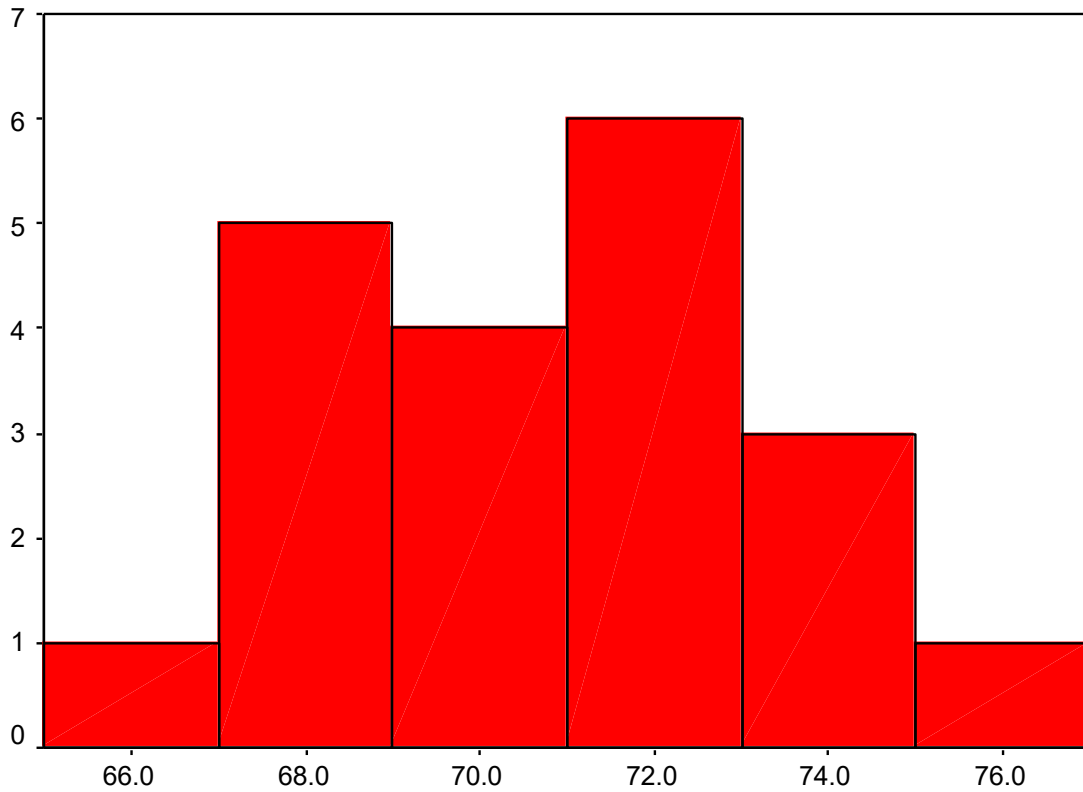
Sandwich	Grams
Whopper and King size fries	70
Wendy's Big Bacon Classic with Biggie Fries	53
Quarter Pounder with Cheese and large fries	56
Big Mac with large fries	58
Arby's Regular Roast Beef and large curly fries	50
Arby's Big Montana and large curly fries	70
Angus Big Bacon & Cheese Steak and King size fries	63

**Mean:  $420/7 = 60$  (average daily intake of fast food for 2000 calorie/day should be 65!)**

**Median: 58**

**Mode: 70**

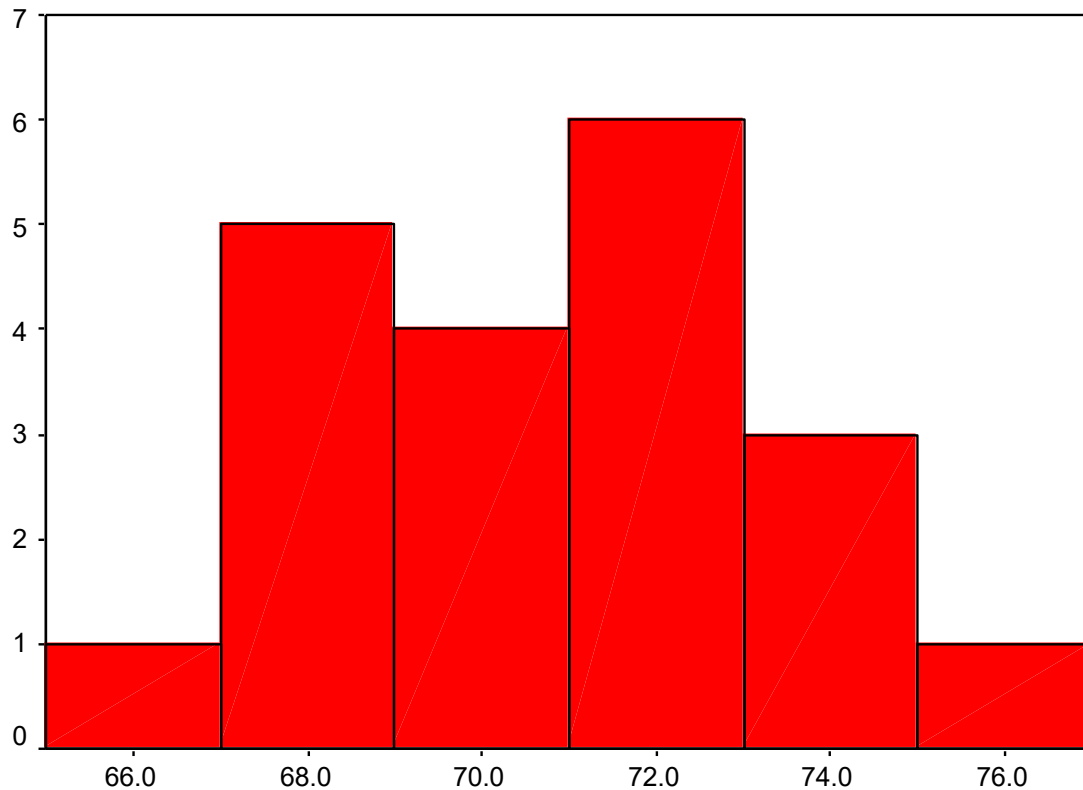
Below you will see a histogram that displays the frequency distribution of a sample of singers' height of voices.



**Find the mean, median, mode.**

[illegible]

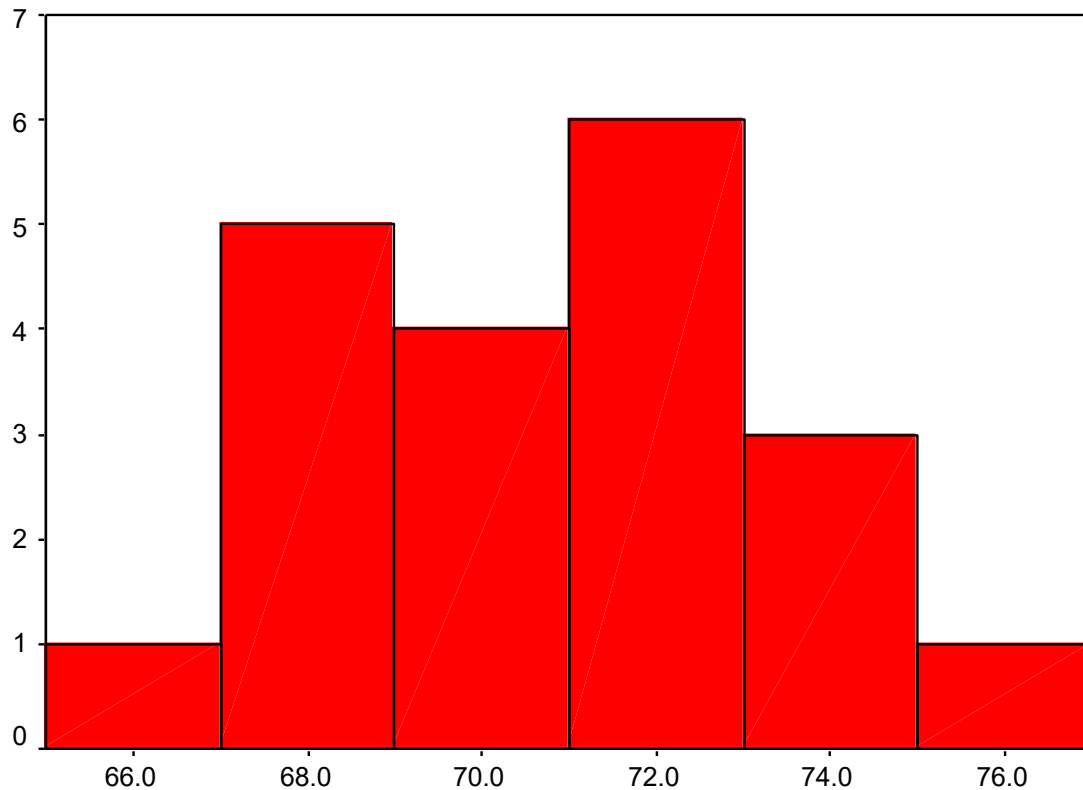
Below you will see a histogram that displays the frequency distribution of a sample of singers' height of voices.



**Find the mean, median, mode.**

Height	Frequency	$xf$
66		
68		
70		
72		
74		
76		

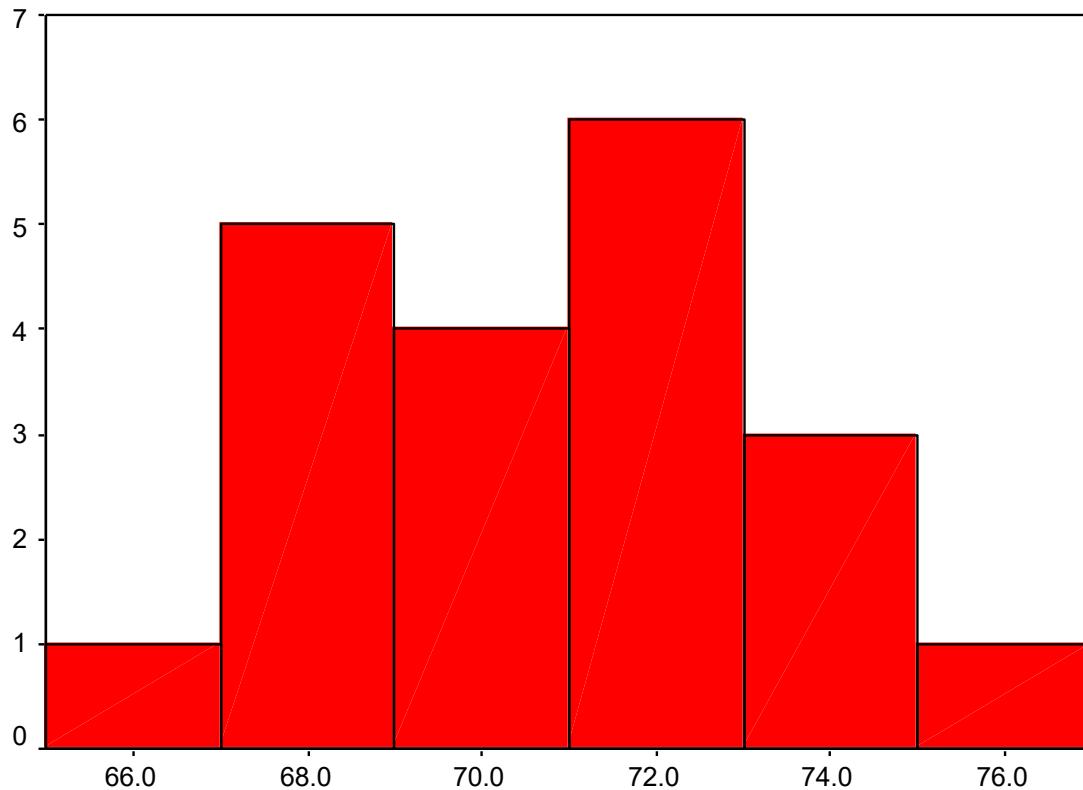
Below you will see a histogram that displays the frequency distribution of a sample of singers' height of voices.



**Find the mean, median, mode.**

Height	Frequency	$xf$
66	1	
68	5	
70	4	
72	6	
74	3	
76	1	

Below you will see a histogram that displays the frequency distribution of a sample of singers' height of voices.

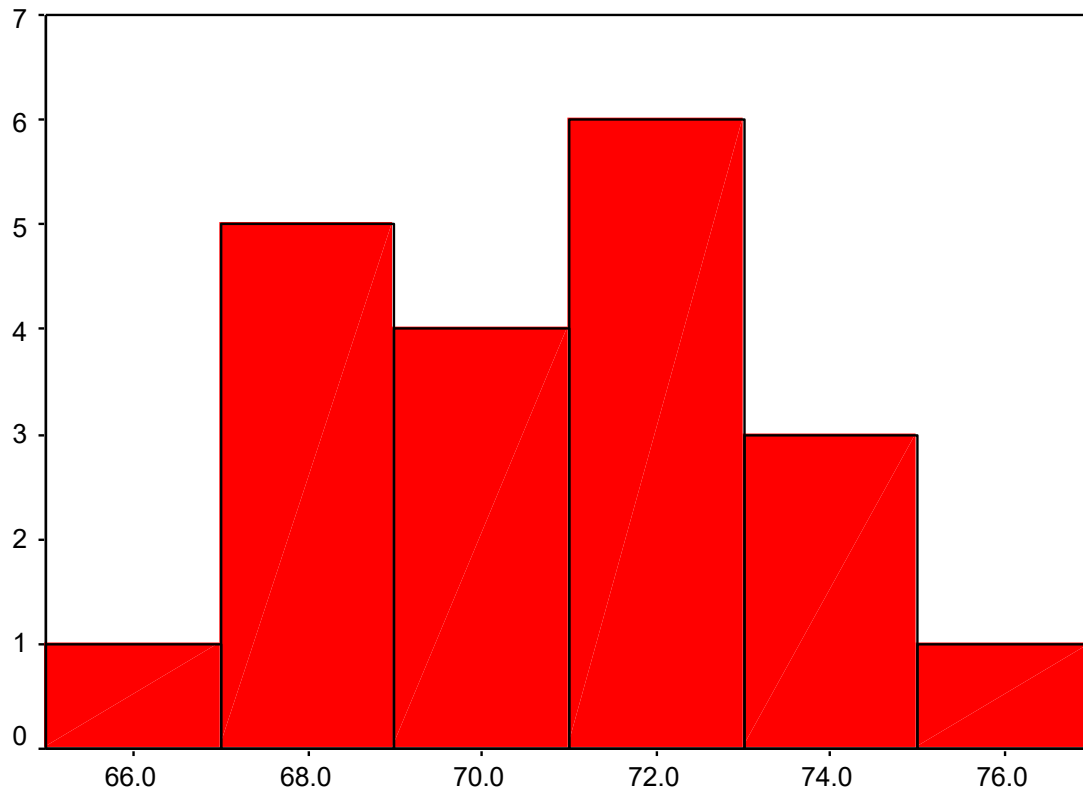


**Find the mean, median, mode.**

Height(x)	Frequency (f)	$xf$
66	1	66
68	5	340
70	4	280
72	6	432
74	3	222
76	1	76



Below you will see a histogram that displays the frequency distribution of a sample of singers' height of voices.



**The mean = 70.8, median = 71,  
mode = 72.**

Height	Frequency	$xf$
66	1	66
68	5	340
70	4	280
72	6	432
74	3	222
76	1	76

# Example

- The set: 2, 9, 11, 5, 6

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2 + 9 + 11 + 5 + 6}{5} = \frac{33}{5} = 6.6$$

If we were able to enumerate the whole population, the **population mean** would be called  $\mu$  (the Greek letter “mu”).

# Median

- The **median** of a set of measurements is the middle measurement when the measurements are ranked from smallest to largest.
- The **position of the median** is

$$0.5(n + 1)$$

once the measurements have been ordered.

# Example

- The set : 2, 4, 9, 8, 6, 5, 3  $n = 7$
- Sort : 2, 3, 4, 5, 6, 8, 9
- Position:  $.5(n + 1) = .5(7 + 1) = 4^{\text{th}}$

Median = 4<sup>th</sup> largest measurement

- The set: 2, 4, 9, 8, 6, 5  $n = 6$
- Sort: 2, 4, 5, 6, 8, 9
- Position:  $.5(n + 1) = .5(6 + 1) = 3.5^{\text{th}}$

Median =  $(5 + 6)/2 = 5.5$  — average of the 3<sup>rd</sup> and 4<sup>th</sup> measurements

# Mode

- The **mode** is the measurement which occurs most frequently.
- The set: 2, 4, 9, 8, 8, 5, 3
  - The mode is **8**, which occurs twice
- The set: 2, 2, 9, 8, 8, 5, 3
  - There are two modes—**8** and **2** (bimodal)
- The set: 2, 4, 9, 8, 5, 3
  - There is **no mode** (each value is unique).

# Example

The number of quarts of milk purchased by 25 households:

0 0 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3  
3 3 3 4 4 4 5

- Mean?

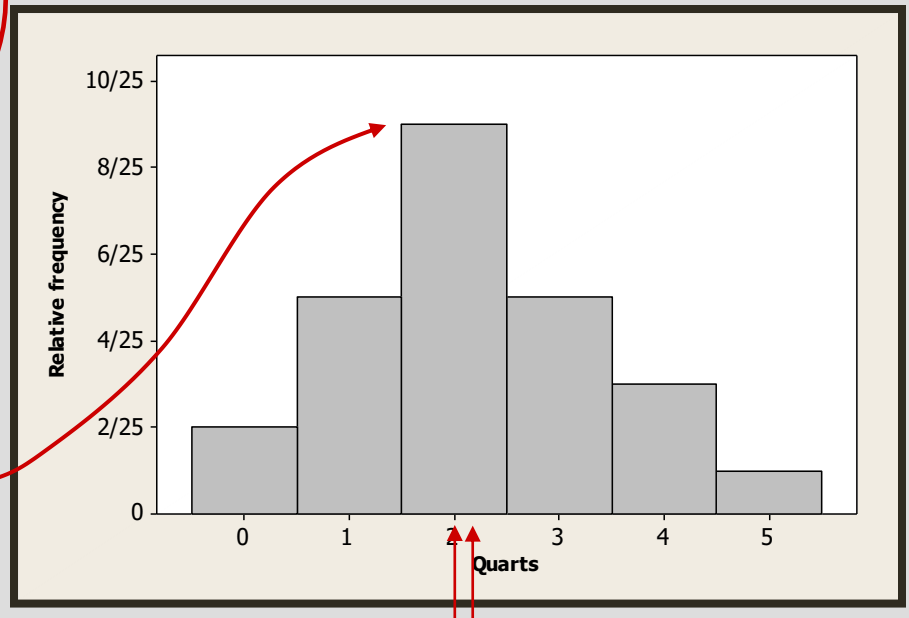
$$\bar{x} = \frac{\sum x_i}{n} = \frac{55}{25} = 2.2$$

- Median?

$$m = 2$$

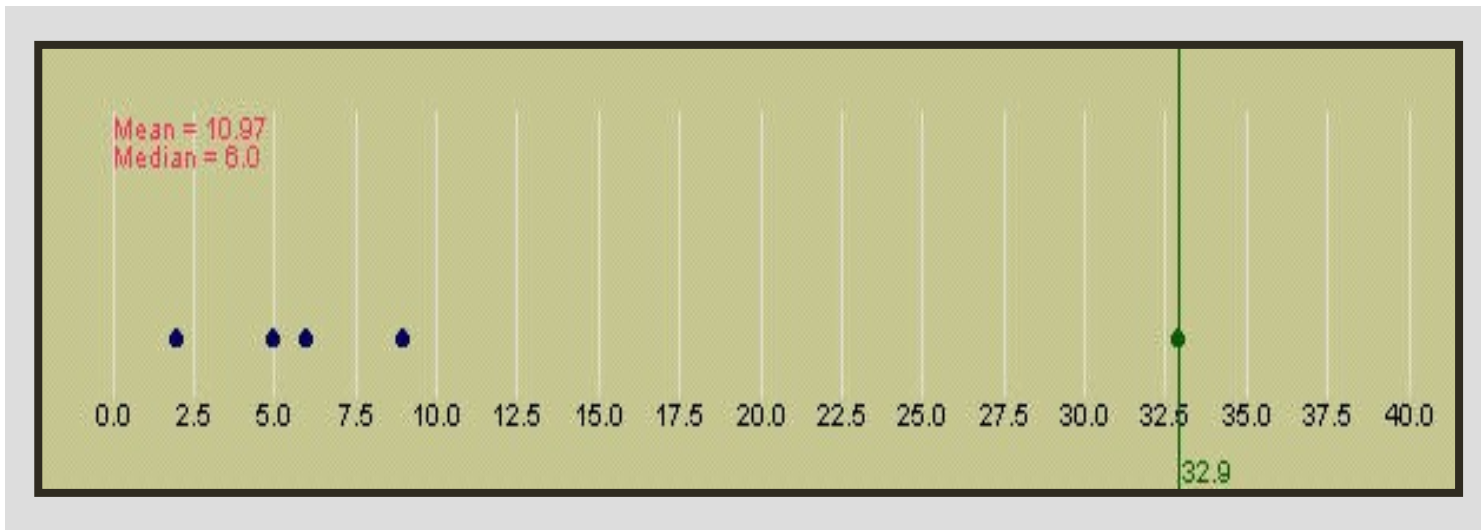
- Mode? (Highest peak)

$$\text{mode} = 2$$



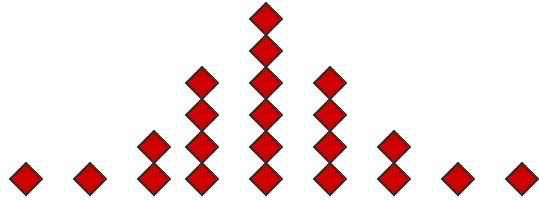
# Extreme Values

- The mean is more easily affected by extremely large or small values than the median.

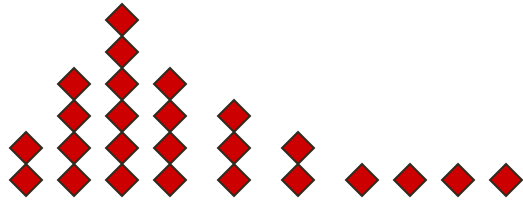


- The median is often used as a measure of center when the distribution is skewed.

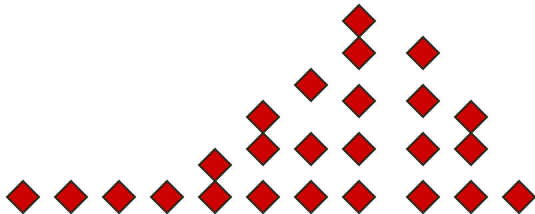
# Extreme Values



**Symmetric: Mean = Median**



**Skewed right: Mean > Median**



**Skewed left: Mean < Median**



# **Dispersion (*Variability*)**

# Variability

- As a descriptive statistic, variability measures the degree to which the scores are spread out or clustered together in a distribution.
- When the population variability is small, all of the scores are clustered close together and any individual score or sample will necessarily provide a good representation of the entire set.
- On the other hand, when variability is large and scores are widely spread, it is easy for one or two extreme scores to give a distorted picture of the general population.

## **Variability**

- describes the degree to which the scores are spread out or clustered together
- Describes distance of the spread of scores or distance of a score from the mean

## **Purposes of Measure of Variability**

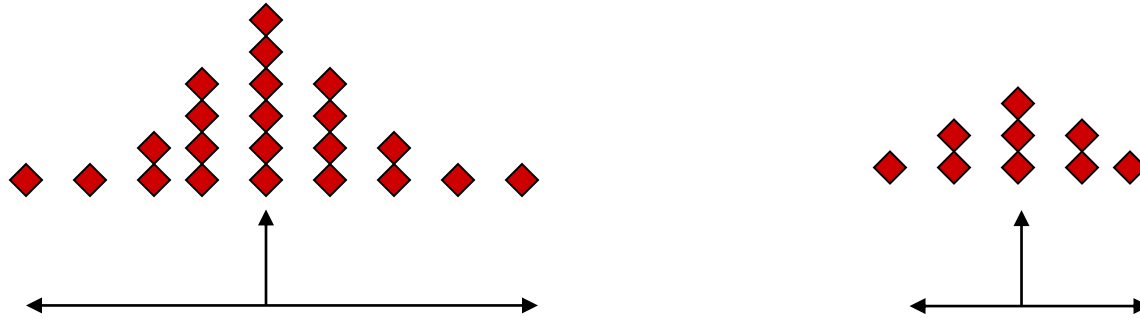
- Describe the amount of variability in the distribution
- Describe how well an individual score or group of scores represents the entire distribution

## **Three Measures of Variability**

- Range
- Interquartile range
- Variance
- Standard deviation

# Measures of Variability

- A measure along the horizontal axis of the data distribution that describes the **spread** of the distribution from the center.



- **Range**

- ✓ Difference between maximum and minimum values

- **Interquartile Range**

- ✓ Difference between third and first quartile ( $Q_3 - Q_1$ )

- **Variance**

- ✓ Average\* of the squared deviations from the mean

- **Standard Deviation**

- ✓ Square root of the variance

# Measures of Variation

- The ***range*** of a set of data is the difference between the greatest and least values.
- The ***interquartile range*** is the difference between the third and first quartiles

# The Range

- The **range,  $R$** , of a set of  $n$  measurements is the difference between the largest and smallest measurements.
- **Example:** A botanist records the number of petals on 5 flowers:

**5, 12, 6, 8, 14**

- The range is

$$R = 14 - 5 = 9.$$

Quick and easy, but only uses 2 of the 5 measurements.

# Interquartile Range ( $IQR = Q_1 - Q_3$ )

The lower and upper quartiles ( $Q_1$  and  $Q_3$ ), can be calculated as follows:

- The **position of  $Q_1$**  is

$$0.25(n + 1)$$

- The **position of  $Q_3$**  is

$$0.75(n + 1)$$

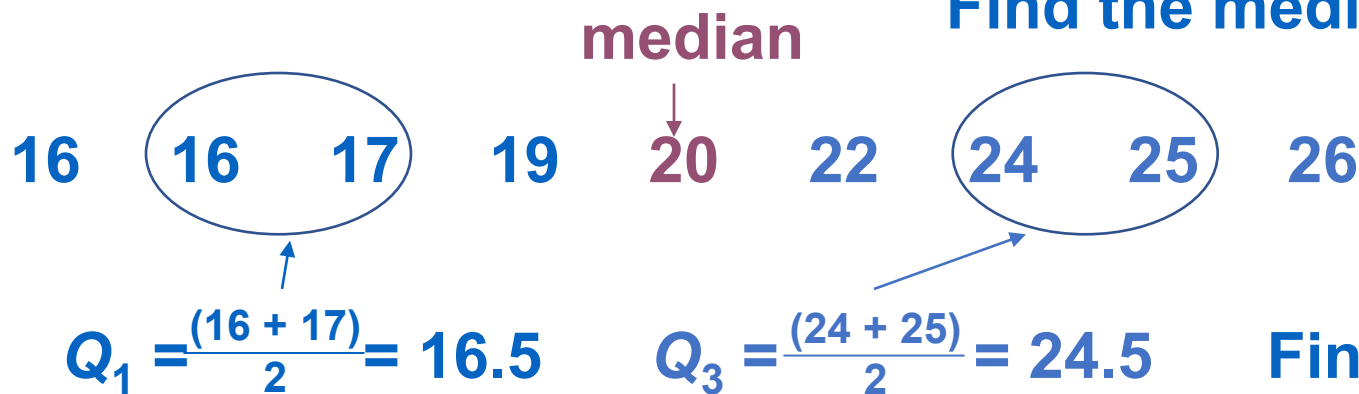
Once the measurements have been ordered. If the positions are not integers, find the quartiles by interpolation.

## Example

There are 9 members of the Community Youth Leadership Board. Find the range and interquartile range of their ages: 22, 16, 24, 17, 16, 25, 20, 19, 26.

greatest value – least value =  $26 - 16 = 10$  Find the range.

Find the median.



$Q_3 - Q_1 = 24.5 - 16.5 = 8$  Find the interquartile range

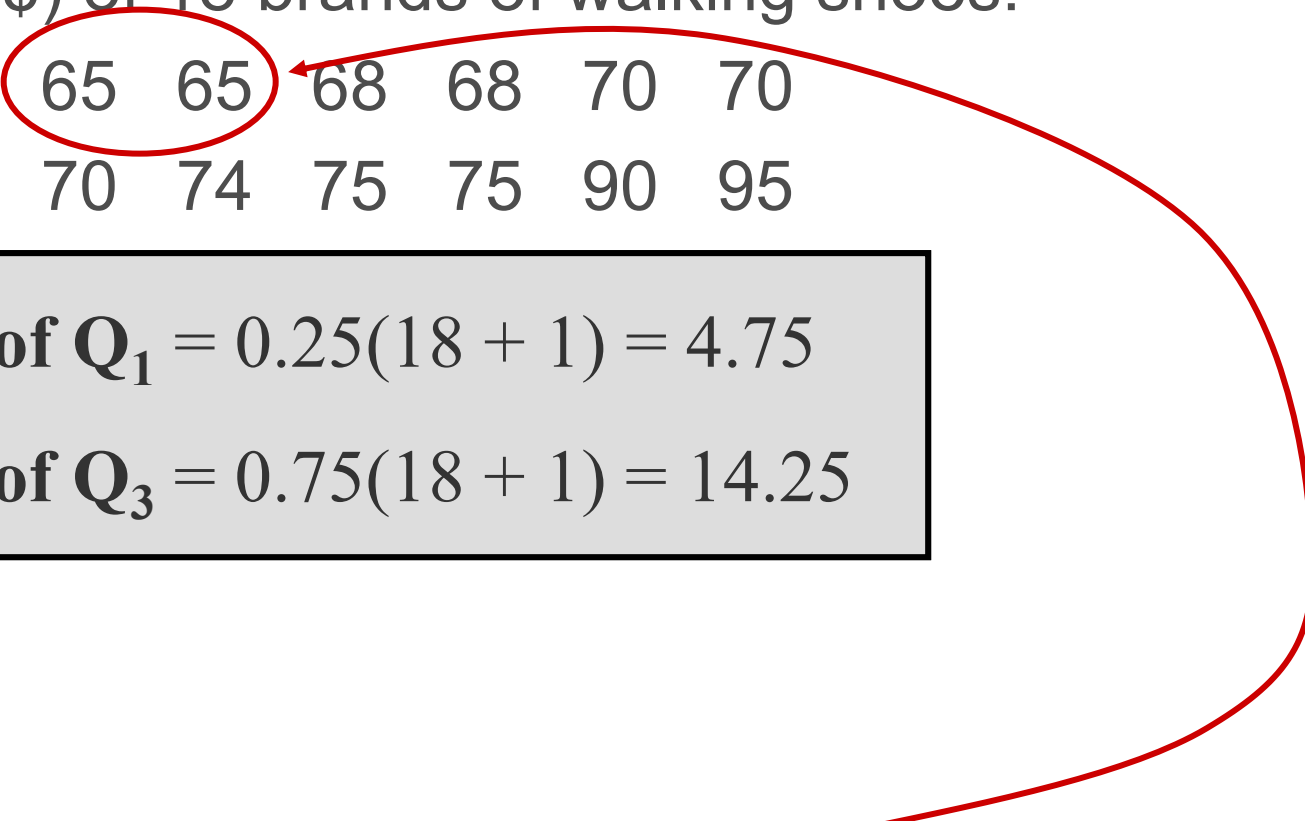
The range is 10 years. The interquartile range is 8 years.



# Example

The prices (\$) of 18 brands of walking shoes:

40 60 65 65 65 68 68 70 70  
70 70 70 70 74 75 75 90 95



$$\text{Position of } Q_1 = 0.25(18 + 1) = 4.75$$

$$\text{Position of } Q_3 = 0.75(18 + 1) = 14.25$$

✓  $Q_1$  is 3/4 of the way between the 4<sup>th</sup> and 5<sup>th</sup> ordered measurements, or  $Q_1 = 65 + 0.75(65 - 65) = 65$ .

# Example

The prices (\$) of 18 brands of walking shoes:

40	60	65	65	65	68	68	70	70
70	70	70	70	74	75	75	90	95

**Position of  $Q_1 = 0.25(18 + 1) = 4.75$**

**Position of  $Q_3 = 0.75(18 + 1) = 14.25$**

✓  $Q_3$  is 1/4 of the way between the 14<sup>th</sup> and 15<sup>th</sup> ordered measurements, or

$$Q_3 = 74 + .25(75 - 74) = 74.25$$

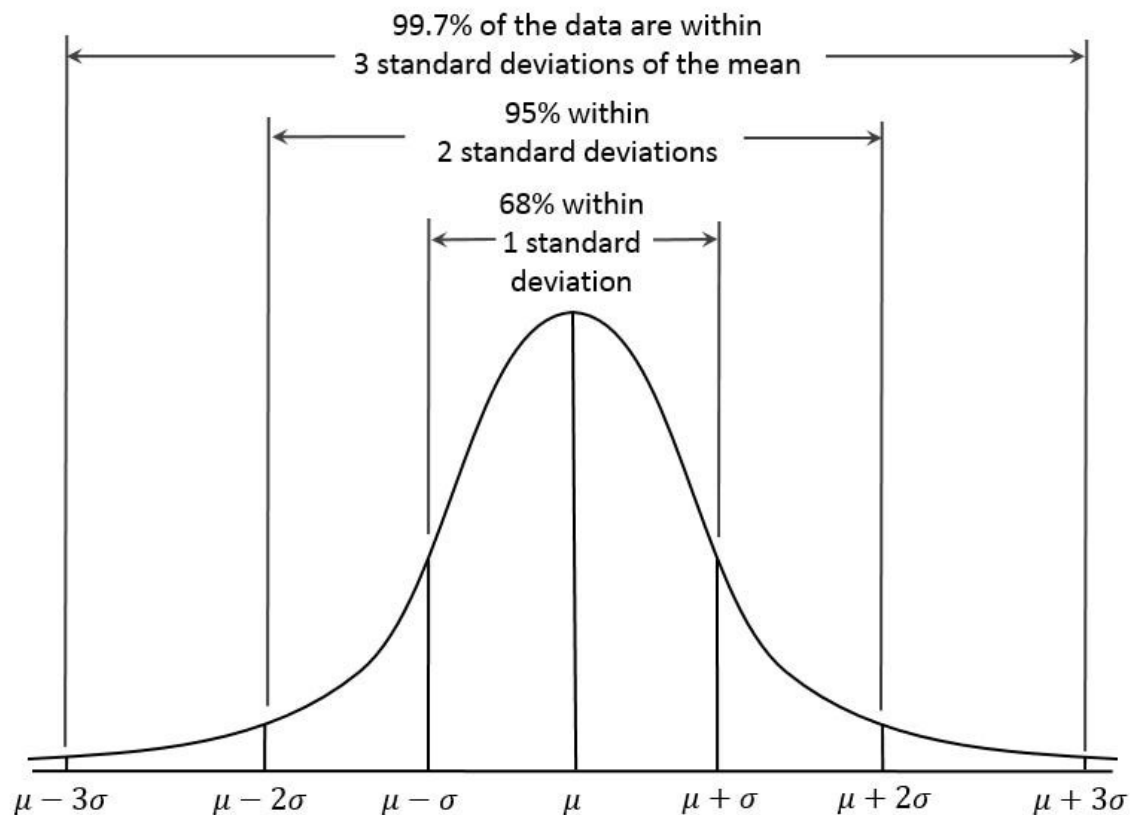
✓ and

$$IQR = Q_3 - Q_1 = 74.25 - 65 = 9.25$$

# More Measures of Variation

- **Standard deviation** is a measure of how each value in a data set varies or deviates from the mean

- A measure of the standard distance from the mean
- Describes whether the scores are clustered closely around the mean or are widely scattered
- "Standard" is defined as the square root of the mean squared deviations
- SD is not the average deviation



# Steps to Finding Standard Deviation

1. Find the mean of the set of data:  $\bar{x}$
2. Find the difference between each value and the mean:  $x - \bar{x}$
3. Square the difference:  $(x - \bar{x})^2$
4. Find the average (mean) of these squares:  $\frac{\sum (x - \bar{x})^2}{n}$
5. Take the square root to find the standard deviation ( $\sigma$ ):

$$\sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

# Standard Deviation

**Find the mean and the standard deviation for the values 78.2, 90.5, 98.1, 93.7, 94.5.**

$$\bar{x} = \frac{(78.2 + 90.5 + 98.1 + 93.7 + 94.5)}{5} = 91$$

**Find the mean**

**Organize the next steps in a table.**

**Find the standard deviation**

$x$	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
78.2	91	-12.8	163.84
90.5	91	-0.5	0.25
98.1	91	7.1	50.41
93.7	91	2.7	7.29
94.5	91	3.5	12.25
		$\sum (x - \bar{x})^2$	234.04

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$\sigma = \sqrt{\frac{234.04}{5}} \approx 6.8$$

**The mean is 91, and the standard deviation is about 6.8**

# Standard Deviation

**Find the mean and the standard deviation for the values 9, 4, 5, 6**

$$\bar{x} = 6$$

**Find the mean**

**Organize the next steps in a table.**

**Find the standard deviation**

$x$	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
9	6	3	9
4	6	-2	4
5	6	-1	1
6	6	0	0
$\sum (x - \bar{x})^2$			14

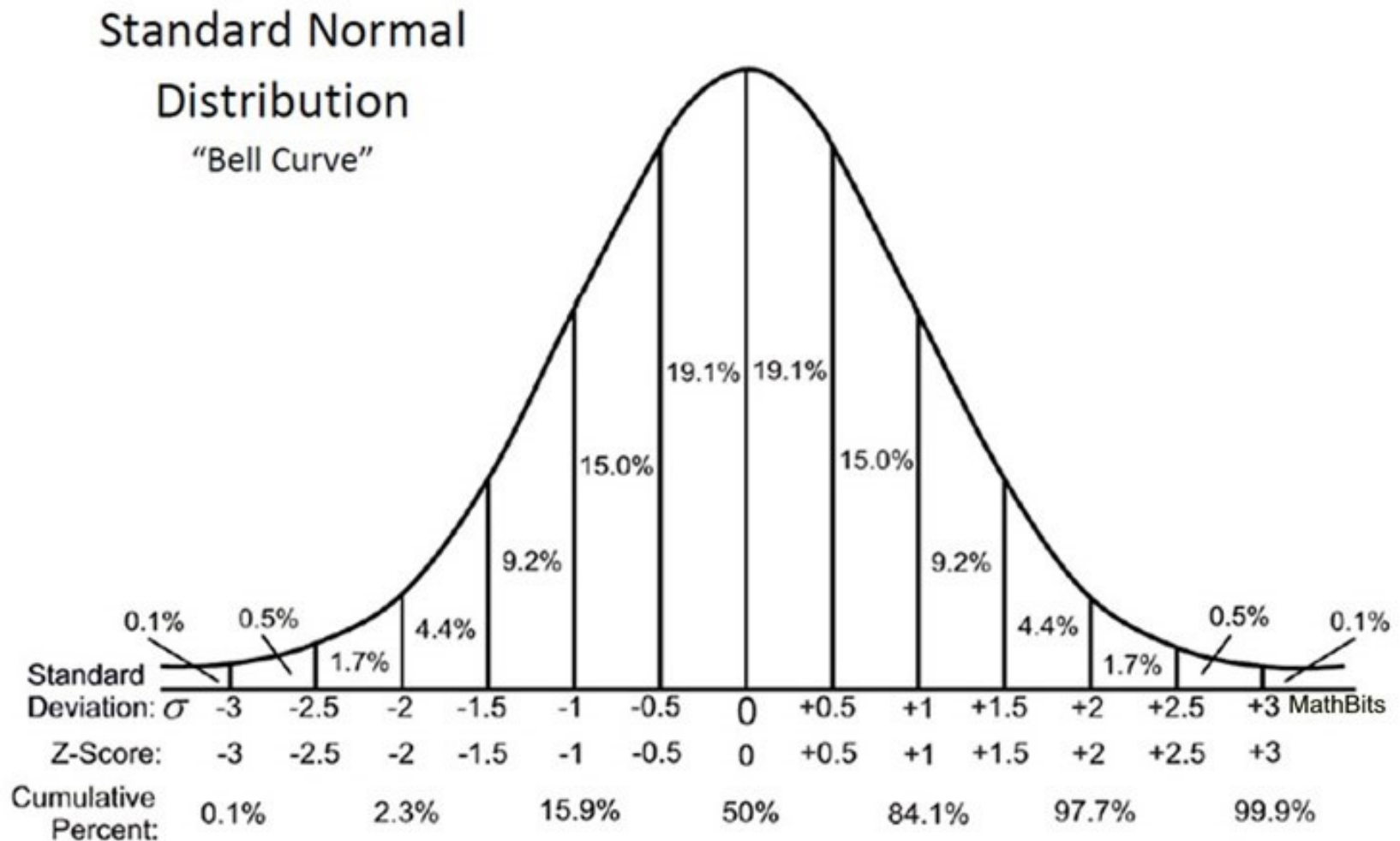
$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$\sigma = 1.87$$

**The mean is 6, and the standard deviation is about 1.87**

# More Measures of Variation

- Z-Score: The Z-Score is the number of standard deviations that a value is from the mean.



# Z-Score

A set of values has a mean of 22 and a standard deviation of 3. Find the z-score for a value of 24.

$$\text{z-score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

$$= \frac{24 - 22}{3}$$

**Substitute.**

$$= \frac{2}{3}$$

**Simplify.**

$$= 0.6$$



# Z-Score

A set of values has a mean of 34 and a standard deviation of 4. Find the z-score for a value of 26.

$$\text{z-score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

$$= \frac{26 - 34}{4} \quad \text{Substitute.}$$

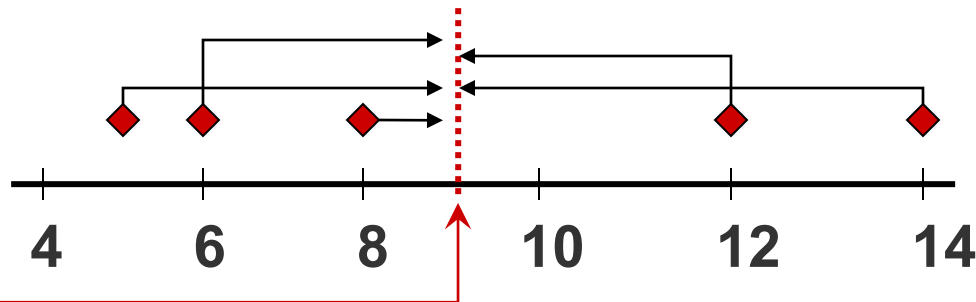
$$= \frac{-8}{4} \quad \text{Simplify.}$$

$$=-2$$

# The Variance

- The **variance** is measure of variability that uses all the measurements. It measures the average deviation of the measurements about their mean.
- **Flower petals: 5, 12, 6, 8, 14**

$$\bar{x} = \frac{45}{5} = 9$$



# The Variance

- The **variance of a population** of  **$N$**  measurements is the average of the squared deviations of the measurements about their mean  $\mu$ .

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

- The **variance of a sample** of  **$n$**  measurements is the sum of the squared deviations of the measurements about their mean, divided by  **$(n - 1)$** .

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

# The Standard Deviation

- In calculating the variance, we squared all of the deviations, and in doing so changed the scale of the measurements.
- To return this measure of variability to the original units of measure, we calculate the **standard deviation**, the positive square root of the variance.

Population standard deviation :  $\sigma = \sqrt{\sigma^2}$

Sample standard deviation :  $s = \sqrt{s^2}$

# Variance: a measure of how data points differ from the mean

- Data Set 1: 3, 5, 7, 10, 10  
Data Set 2: 7, 7, 7, 7, 7

What is the mean and median of the above data set?

Data Set 1: mean = 7, median = 7

Data Set 2: mean = 7, median = 7

But we know that the two data sets are not identical! The **variance** shows how they are different.

We want to find a way to represent these two data set numerically.

# How to Calculate?

- If we conceptualize the spread of a distribution as the extent to which the values in the distribution differ from the mean and from each other, then a reasonable measure of spread might be the average deviation, or difference, of the values from the mean.

$$\frac{\sum(x - \bar{X})}{N}$$

- Although this might seem reasonable, this expression always equals 0, because the negative deviations about the mean always cancel out the positive deviations about the mean.
- We could just drop the negative signs, which is the same mathematically as taking the absolute value, which is known as the mean deviations.
- The average of the squared deviations about the mean is called the variance.

$$\sigma^2 = \frac{\sum (x - \bar{X})^2}{N} \quad \text{For population variance}$$

$$s^2 = \frac{\sum (x - \bar{X})^2}{n - 1} \quad \text{For sample variance}$$

**Find the mean and variance for the values  
3,5,7,10,10**

$x$	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
3			
5			
7			
10			
10			

**The mean is ?**



**Find the mean and variance for the values  
3,5,7,10,10**

$x$	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
3	7		
5	7		
7	7		
10	7		
10	7		

**The mean is =  $35/5 = 7$**

**Find the mean and variance for the values  
3,5,7,10,10**

$x$	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
3	7	3-7=-4	16
5	7	5-7=-2	4
7	7	7-7=0	0
10	7	10-7=3	9
10	7	10-7=3	9
Sum=35			Sum=38

**The mean  
is = 35/5 =7**

$$s^2 = \frac{\sum (x - \bar{X})^2}{n} = \frac{38}{5} = 7.6$$

$x$	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
86	72	14	196
49	72	-23	529
63	72	-9	81
90	72	18	324
82	72	10	100
98	72	26	676
36	72	-36	1296
			Sum=3202

**N=7**

**The mean  
is = 72**

**Variance = 3199/6  
=533.67**

$$s^2 = \frac{\sum (x - \bar{X})^2}{n} = \frac{38}{5} = 7.6$$

$x$	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
180	181.25	-1.25	
313			
101			
255			
202			
198			
109			
183			
181			
113			
171			
165			
318			
145			
131			
145			
226			
113			
268			
108			
			Sum=83605.75

**N=20**

**The mean  
is = 181.25**

**181.25 + 3x66.33**

**181.25 – 3x66.33**

**sd = 66.33**

## Example 2

Dive	Mark	Myrna
1	28	27
2	22	27
3	21	28
4	26	6
5	18	27

Find the mean, median, mode, range?

**mean            23       23**

**median        22       27**

**range          10       22**

What can be said about this data?

**Due to the outlier, the median is more typical of overall performance.**

Which diver was more consistent?

## Example 2

Dive	Mark	Myrna
1	28	27
2	22	27
3	21	28
4	26	6
5	18	27

Find the mean, median, mode, range?

**mean            23        23**

**median        21        28**

**range          10        22**

What can be said about this data?

**Due to the outlier, the median is more typical of overall performance.**

Which diver was more consistent?

Dive	Marks $x$	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
1	28	23	5	25
2	22	23	-1	1
3	21	23	-2	4
4	26	23	3	9
5	18	23	-5	25
	Sum=115		0	Sum=64

Dive	Myrna $x$	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
1	27	23	4	16
2	27	23	4	16
3	28	23	5	25
4	6	23	-17	289
5	27	23	4	16
	Sum=115		0	Sum=362

**Mark's Variance =  $64 / 5 = 12.8$**

**Myrna's Variance =  $362 / 5 = 72.4$**

**Conclusion: Mark has a lower variance therefore he is more consistent.**

# standard deviation - a measure of variation of scores about the mean

- Can think of standard deviation as the average distance to the mean, although that's not numerically accurate, it's conceptually helpful. All ways of saying the same thing: higher standard deviation indicates higher spread, less consistency, and less clustering.

- sample standard deviation: 
$$s = \sqrt{\frac{\sum (x - \bar{X})^2}{n - 1}}$$

- population standard deviation: 
$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$



# Bell shaped curve

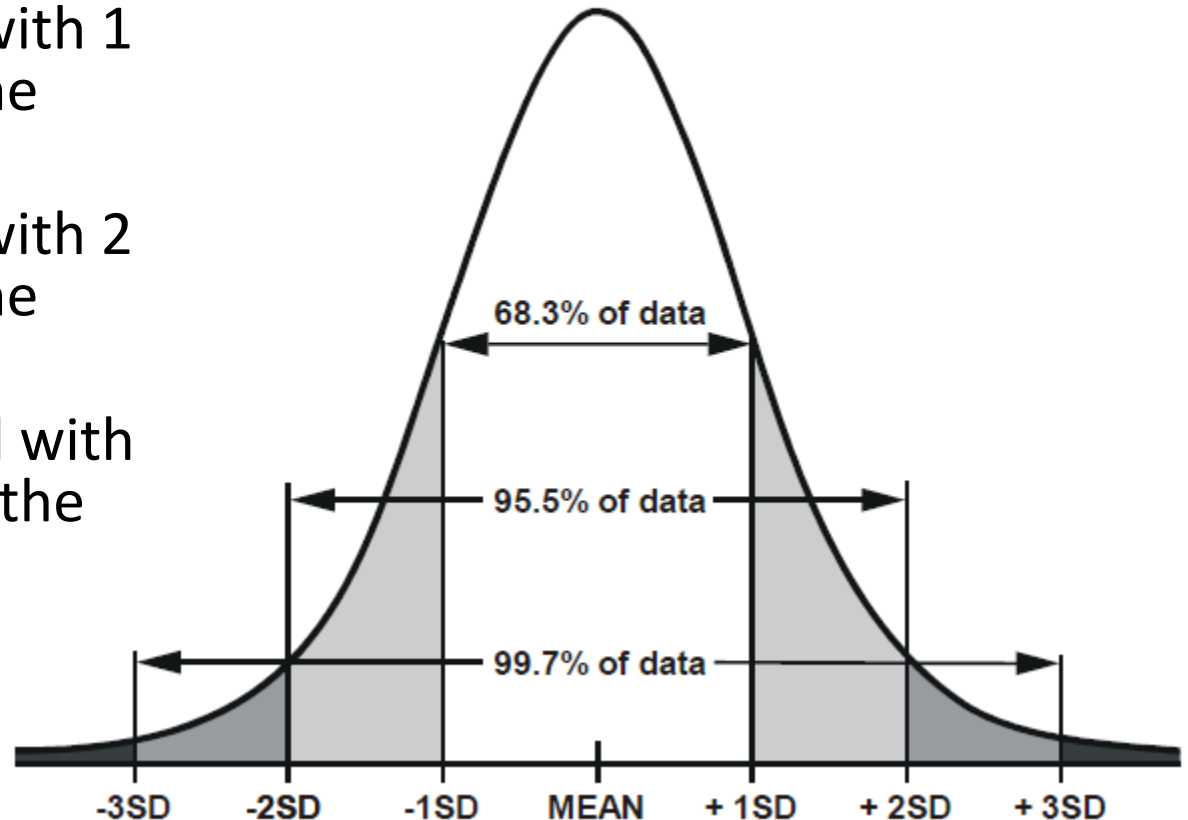
Empirical rule for data (68-95-99) - only applies to a set of data having a distribution that is approximately bell-shaped:

≈ 68% of all scores fall with 1 standard deviation of the mean

≈ 95% of all scores fall with 2 standard deviation of the mean

≈ 99.7% of all scores fall with 3 standard deviation of the mean

Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean



## Variability (cont.)

- When the population variability is small, all of the scores are clustered close together and any individual score or sample will necessarily provide a good representation of the entire set.
- On the other hand, when variability is large and scores are widely spread, it is easy for one or two extreme scores to give a distorted picture of the general population.

# Using Measures of Center and Spread:

## Chebysheff's Theorem

Given a number  $k$  greater than 1 and a set of  $n$  measurements, at least  $1 - (1/k^2)$  of the measurement will lie within  $k$  standard deviations of the mean.

✓ Can be used for either samples ( $\bar{x}$  and  $s$ ) or for a population ( $\mu$  and  $\sigma$ ).

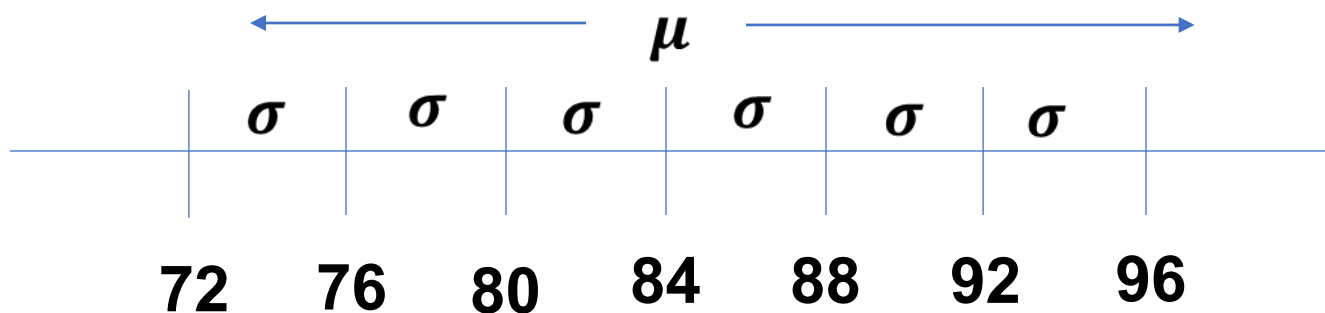
✓ **Important results:**

✓ If  $k = 2$ , **at least**  $1 - 1/2^2 = 3/4$  of the measurements are within 2 standard deviations of the mean.

✓ If  $k = 3$ , **at least**  $1 - 1/3^2 = 8/9$  of the measurements are within 3 standard deviations of the mean.

Suppose that the average score on a math test is an 84 with a standard Deviation of 4 points.

According to chebychev's theorem, at least what percent of the tests have A grade of at least 72 and at most 96 ?



$$K=3 \rightarrow 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} \approx 0.899 \approx 89.9 \%$$

The average check at a local restaurant is \$36.42 with a standard deviation of \$8.15. What is the minimum percentage of checks between \$15.23 and \$57.61?

**K = no. of standard deviations away from the mean**

$$K = \frac{\text{Limit} - \mu}{\sigma} \quad \text{Limit of 15.23 \& 57.61, mean} = 36.42, \text{ sd} = 8.15$$

$$K = \frac{57.61 - 36.42}{8.15} = 2.6$$

$$K = \frac{15.23 - 36.42}{8.15} = -2.6$$

The average check at a local restaurant is \$36.42 with a standard deviation of \$8.15. What is the minimum percentage of checks between \$15.23 and \$57.61?

**K = no. of standard deviations away from the mean**

$$K = \frac{\text{Limit} - \mu}{\sigma} \qquad K = 2.6$$

***at least*  $\left(1 - \frac{1}{K^2}\right) 100\%$  of the data will lie inside the Given interval.**

$$\text{At least } \left(1 - \frac{1}{2.6^2}\right) 100\% = 85.2\%$$

# Chebyshev's Theorem: An example

The arithmetic mean biweekly amount contributed by the Dupree Paint employees to the company's profit-sharing plan is \$51.54, and the standard deviation is \$7.51. At least what percent of the contributions lie within plus 3.5 standard deviations and minus 3.5 standard deviations of the mean?

**CHEBYSHEV'S THEOREM** For any set of observations (sample or population), the proportion of the values that lie within  $k$  standard deviations of the mean is at least  $1 - 1/k^2$ , where  $k$  is any constant greater than 1.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3.5)^2} = 1 - \frac{1}{12.25} = 0.92$$

# Using Measures of Center and Spread:

## The Empirical Rule

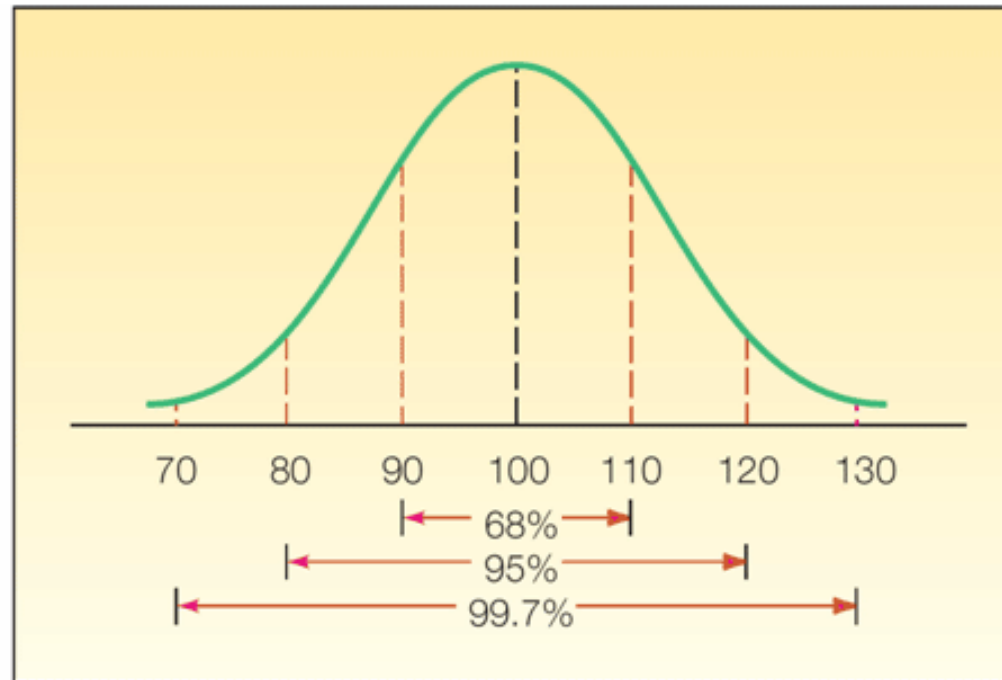
Given a distribution of measurements that is approximately mound-shaped:

- ✓ The interval  $\mu \pm \sigma$  contains approximately 68% of the measurements.
- ✓ The interval  $\mu \pm 2\sigma$  contains approximately 95% of the measurements.
- ✓ The interval  $\mu \pm 3\sigma$  contains approximately 99.7% of the measurements.



# The Empirical Rule: An Example

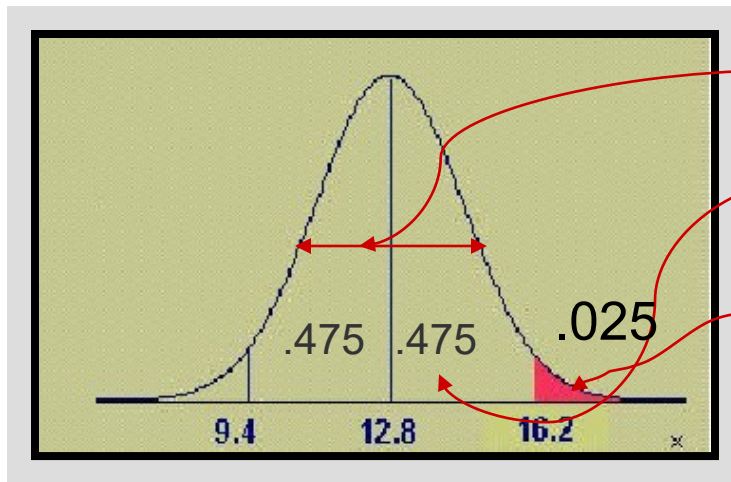
**EMPIRICAL RULE** For a symmetrical, bell-shaped frequency distribution, approximately 68 percent of the observations will lie within plus and minus one standard deviation of the mean; about 95 percent of the observations will lie within plus and minus two standard deviations of the mean; and practically all (99.7 percent) will lie within plus and minus three standard deviations of the mean.



**CHART 3-7** A Symmetrical, Bell-Shaped Curve Showing the Relationships between the Standard Deviation and the Observations

# Example

The length of time for a worker to complete a specified operation averages 12.8 minutes with a standard deviation of 1.7 minutes. If the distribution of times is approximately mound-shaped, what proportion of workers will take longer than 16.2 minutes to complete the task?



**95% between 9.4 and 16.2**

**47.5% between 12.8 and 16.2**

**$(50 - 47.5)\% = 2.5\%$  above 16.2**

# Approximating $s$

- From Chebysheff's Theorem and the Empirical Rule, we know that

$$R \approx 4-6 s$$

- To approximate the standard deviation of a set of measurements, we can use:

$$s \approx R / 4$$

or  $s \approx R / 6$  for a large data set.

# Approximating $s$

The ages of 50 tenured faculty at a state university.

- 34 48 **70** 63 52 52 35 50 37 43 53 43 52 44
- 42 31 36 48 43 **26** 58 62 49 34 48 53 39 45
- 34 59 34 66 40 59 36 41 35 36 62 34 38 28
- 43 50 30 43 32 44 58 53

$$R = 70 - 26 = 44$$

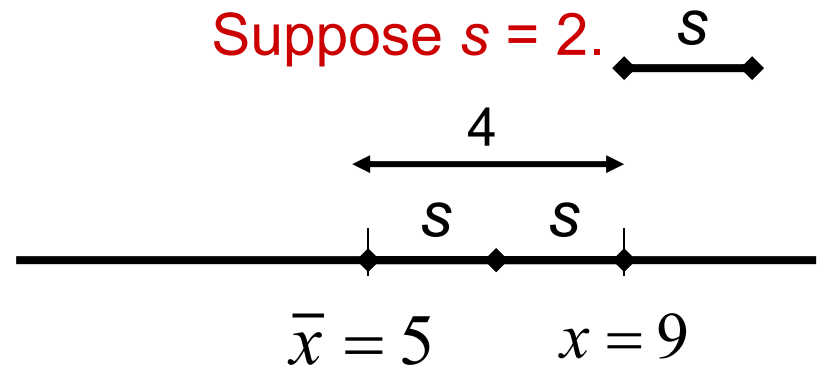
$$s \approx R / 4 = 44 / 4 = 11$$

$$\text{Actual } s = 10.73$$

### 3. Measures of Relative Standing

- Where does one particular measurement stand in relation to the other measurements in the data set?
- How many standard deviations away from the mean does the measurement lie? This is measured by the **z-score**.

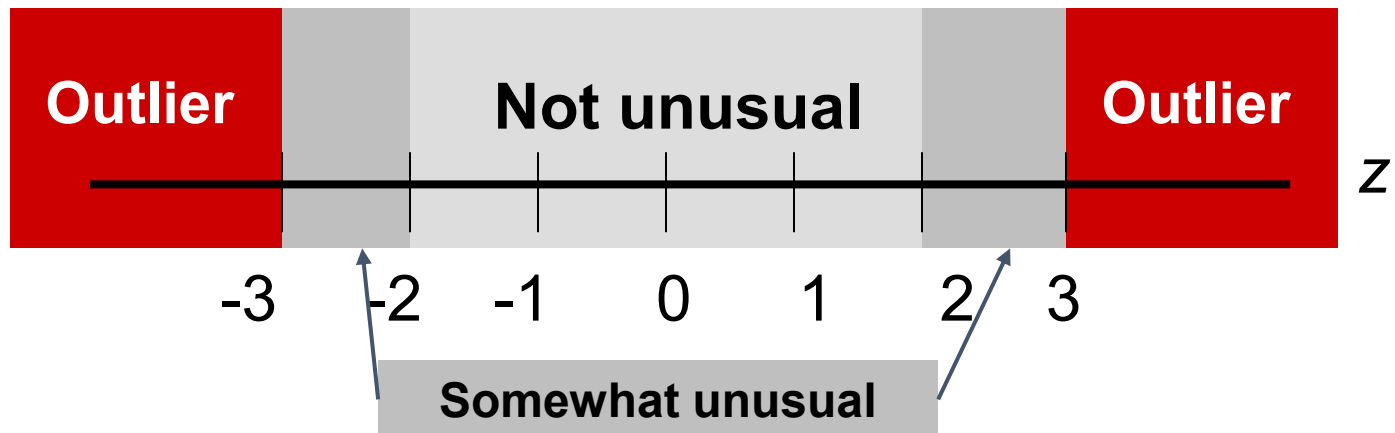
$$z - score = \frac{x - \bar{x}}{s}$$



$x = 9$  lies  $z = 2$  std dev from the mean.

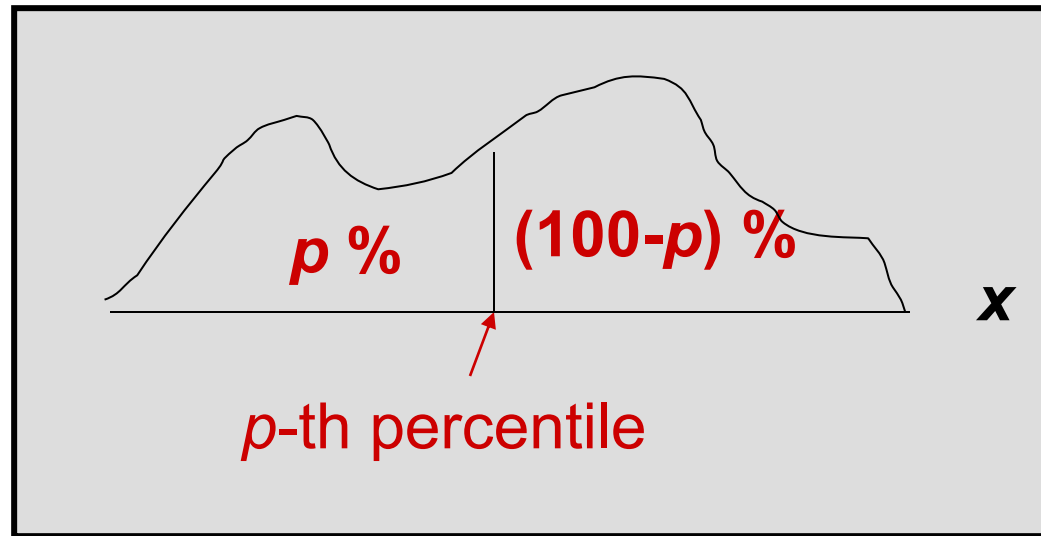
# z-Scores

- From Chebysheff's Theorem and the Empirical Rule
  - At least  $3/4$  and more likely 95% of measurements lie within 2 standard deviations of the mean.
  - At least  $8/9$  and more likely 99.7% of measurements lie within 3 standard deviations of the mean.
- z-scores between  $-2$  and  $2$  are not unusual. z-scores should not be more than 3 in absolute value. z-scores larger than 3 in absolute value would indicate a possible **outlier**.



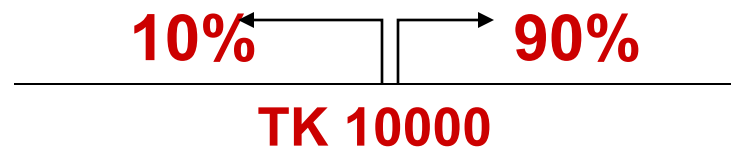
### 3. Measures of Relative Standing (*cont'd*)

- How many measurements lie below the measurement of interest? This is measured by the  $p^{\text{th}}$  percentile.



# Example

- 90% of all men (16 and older) earn more than Tk 10000 per week.



TK 10000 is the 10<sup>th</sup> percentile.

50<sup>th</sup> Percentile  $\equiv$  Median

25<sup>th</sup> Percentile  $\equiv$  Lower Quartile ( $Q_1$ )

75<sup>th</sup> Percentile  $\equiv$  Upper Quartile ( $Q_3$ )



# Using Measures of Center and Spread:

## The Box Plot

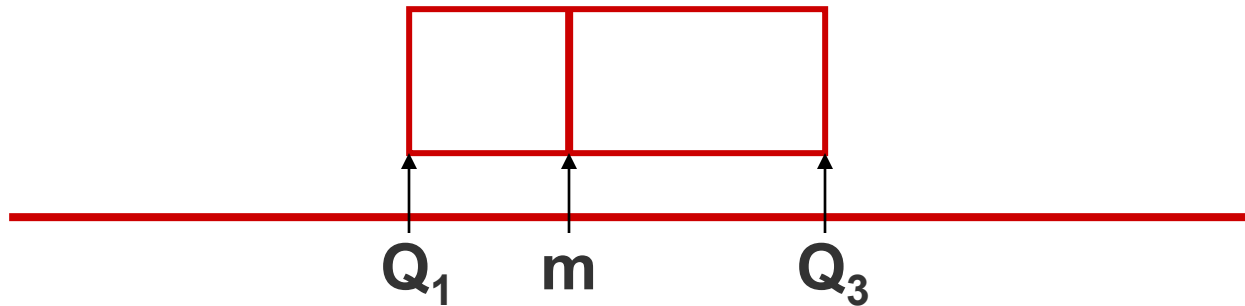
### The Five-Number Summary:

<b>Min</b>	<b><math>Q_1</math></b>	<b>Median</b>	<b><math>Q_3</math></b>	<b>Max</b>
------------	-------------------------	---------------	-------------------------	------------

- Divides the data into 4 sets containing an equal number of measurements.
- A quick summary of the data distribution.
- Use to form a **box plot** to describe the **shape** of the distribution and to detect **outliers**.

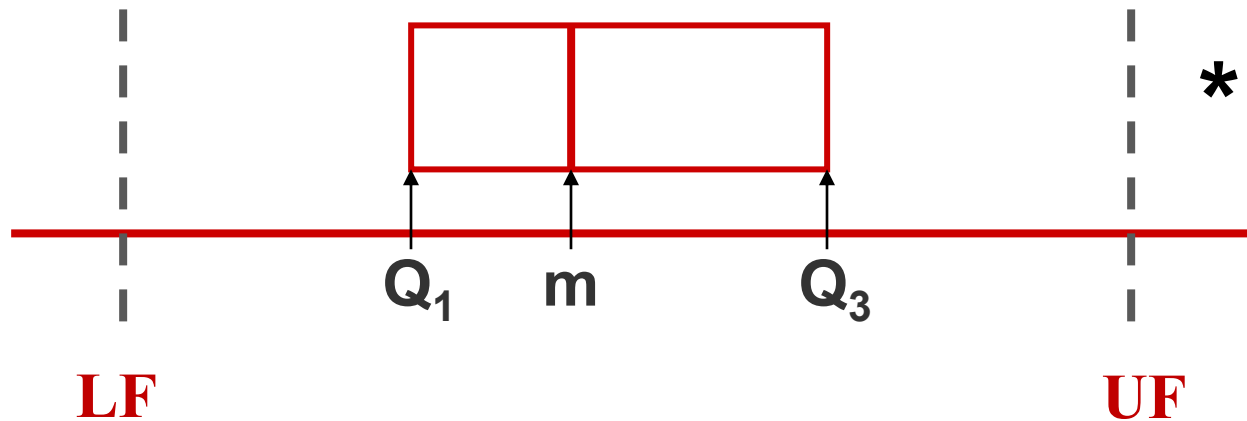
# Constructing a Box Plot

- ✓ Calculate  $Q_1$ , the median,  $Q_3$  and IQR (Interquartile range).
- ✓ Draw a horizontal line to represent the scale of measurement.
- ✓ Draw a box using  $Q_1$ , the median,  $Q_3$ .



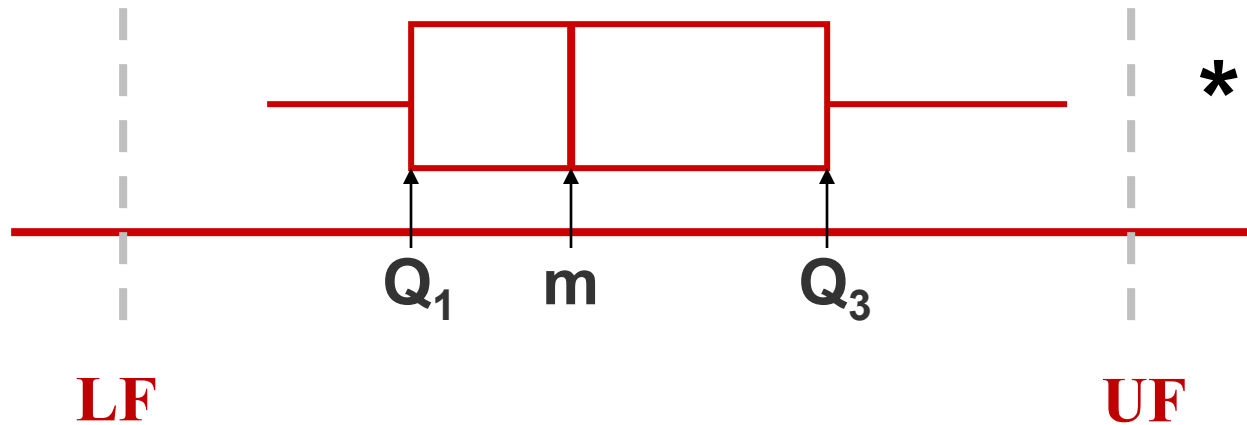
# Constructing a Box Plot

- ✓ Isolate outliers by calculating
  - ✓ Lower fence:  $Q_1 - 1.5 \text{ IQR}$
  - ✓ Upper fence:  $Q_3 + 1.5 \text{ IQR}$
- ✓ Measurements beyond the upper or lower fence is are outliers and are marked (\*).



# Constructing a Box Plot

✓ Draw “**whiskers**” connecting the largest and smallest measurements that are NOT outliers to the box.

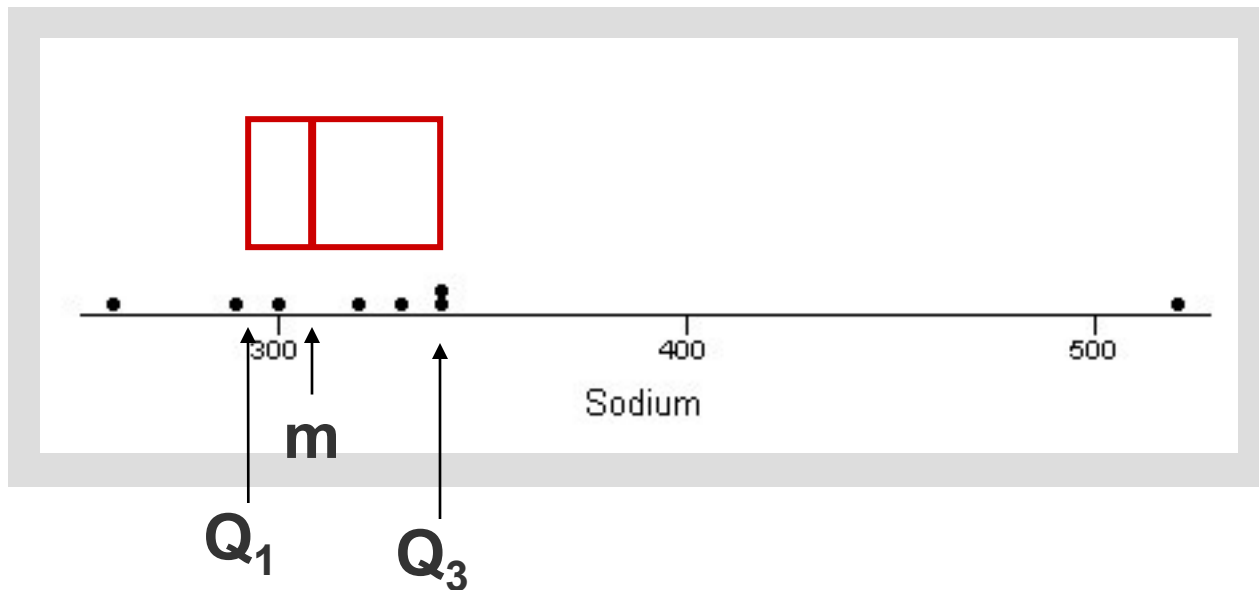


## Example

Amt of sodium in 8 brands of cheese:

260   290   300   320   330   340   340   520

$$Q_1 = 292.5 \quad m = 325 \quad Q_3 = 340$$



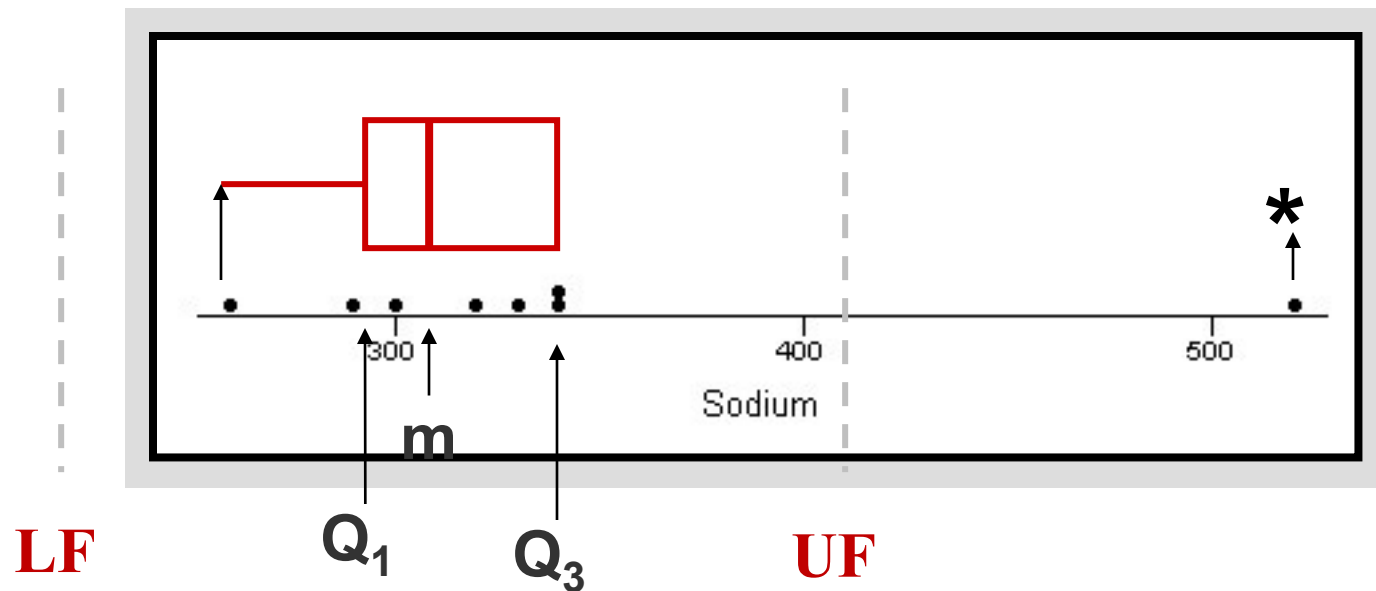
# Example

$$\text{IQR} = 340 - 292.5 = 47.5$$

$$\text{Lower fence} = 292.5 - 1.5(47.5) = 221.25$$

$$\text{Upper fence} = 340 + 1.5(47.5) = 411.25$$

Outlier:  $x = 520$



# Interpreting Box Plots

- ✓ Median line in center of box and whiskers of equal length—symmetric distribution
- ✓ Median line left of center and long right whisker—skewed right
- ✓ Median line right of center and long left whisker—skewed left

