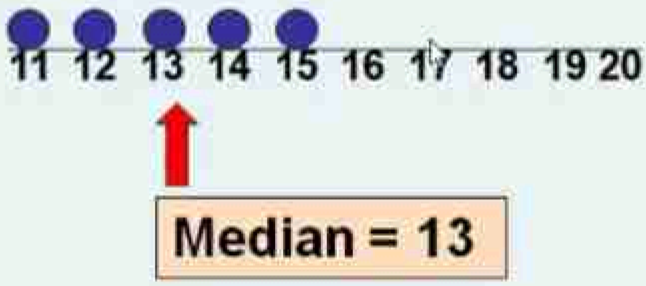
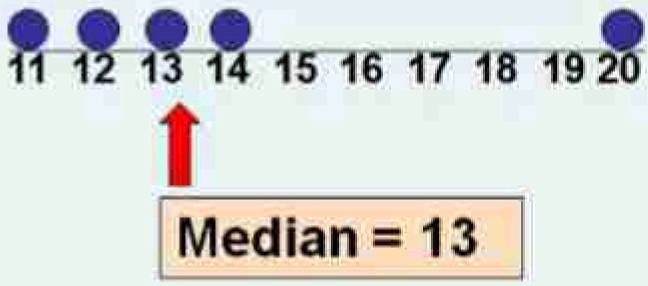
Measures of Central Tendency: The Median

- In an ordered array, the median is the “middle” number (50% above, 50% below)





- Not affected by extreme values

Measures of Central Tendency: Locating the Median

- The location of the median when the values are in numerical order (smallest to largest):

$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

- If the number of values is odd, the median is the middle number
- If the number of values is even, the median is the average of the two middle numbers

Note that $\frac{n+1}{2}$ is not the *value* of the median, only the *position* of the median in the ranked data

Median for Grouped Data

Formula for Median is given by

$$\text{Median} = \boxed{L + \frac{(n/2) - m}{f} \times c}$$

Where

L = Lower limit of the median class

n = Total number of observations = $\boxed{\sum f(x)}$

m = Cumulative frequency preceding the median class

f = Frequency of the median class

c = Class interval of the median class

Example

Find the median for the following continuous frequency distribution:

Class	0-1	1-2	2-3	3-4	4-5	5-6
Frequency	1	4	8	7	3	2

Cont'd…

Class	Frequency	Cumulative Frequency
0-1	1	1
1-2	4	5
2-3	8	13
3-4	7	20
4-5	3	23
5-6	2	25
Total	**25**	

L =Lower limit of the median class
n = Total number of observations
m = Cumulative frequency **preceding** the median class
f = Frequency of the median class
c = Class interval of the median class

Substituting in the formula the relevant values,

Median = $L + \frac{(n/2) - m}{f} \times c$ we have Median = $2 + \frac{(25/2) - 5}{8} \times 1$

= 2.9375

Example

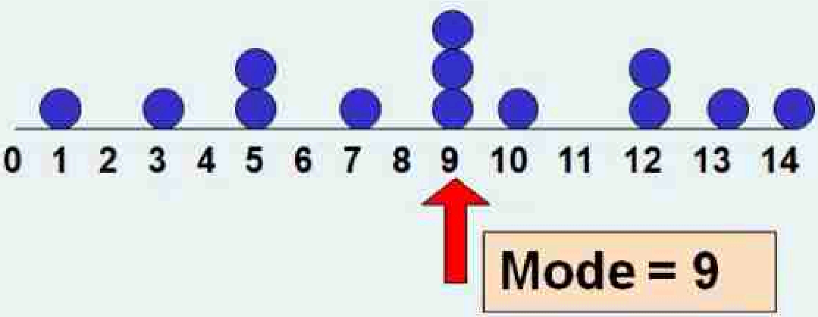
Class interval		f	Cum f
0	49.99	78	78
50	99.99	123	201
100	149.99	187	388
150	199.99	82	
200	249.99	51	
250	299.99	47	
300	349.99	13	
350	399.99	9	
400	449.99	6	
450	499.99	4	
		600	

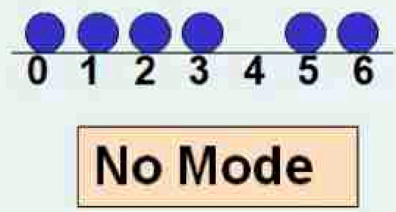
L = Lower limit of the median class
n = Total number of observations
m = Cumulative frequency preceding the median class
f = Frequency of the median class
c = Class interval of the median class

$$L + \frac{(n/2) - m}{f} \times c = 126.47$$

Measures of Central Tendency: The Mode

- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical (nominal) data
- There may be no mode
- There may be several modes





Mode for Grouped Data

Mode = $L + \dfrac{d_1}{d_1 + d_2} \times c$

Where L =Lower limit of the modal class

$d_1 = f_1 - f_0$ $d_2 = f_1 - f_2$

f_1 = Frequency of the **modal class**

f_0 = Frequency **preceding** the modal class

f_2 = Frequency **succeeding** the modal class. C = **Class Interval** of the modal class

Advantages and Disadvantages

Advantages:
Not affected extreme values
Can be computed in case of open class, if median is not in open class
Can be computed in case categorical variable

DisAd: Arraying of the data is time consuming.
To estimate population parameter, mean is easier.

Solution

Class	Frequency
0-1	1
1-2	4
2-3	8
3-4	7
4-5	3
5-6	2
Total	**25**

Mode = $L + \frac{d_1}{d_1 + d_2} \times c$

L = 2

$d_1 = f_1 - f_0$ = 8-4 = 4

$d_2 = f_1 - f_2$ = 8-7 = 1

C = 1 Hence Mode = $2 + \frac{4}{5} \times 1$

= 2.8

Measures of Central Tendency: Review Example

House Prices:
$2,000,000
$ 500,000
$ 300,000
$ 100,000
$ 100,000
Sum $ 3,000,000

- **Mean:** ($3,000,000/5)

  = **$600,000**
- **Median:** middle value of ranked data

  = **$300,000**
- **Mode:** most frequent value

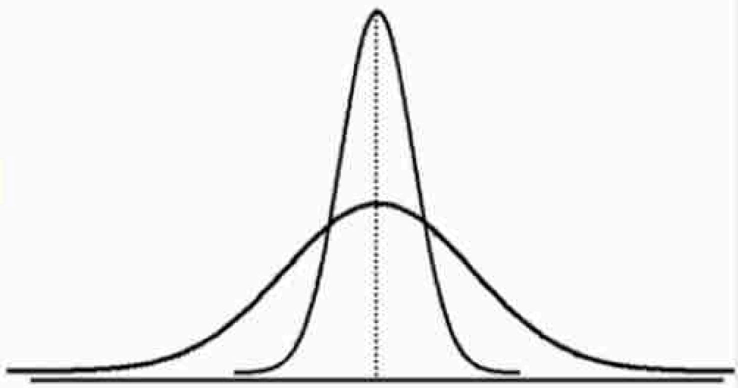
  = **$100,000**

Measures of Variation

Variation

- **Range**
- **Variance**
- **Standard Deviation**
- *Coefficient of Variation*

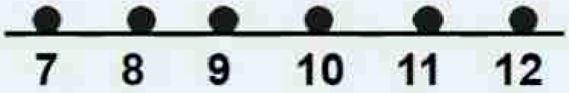
- Measures of variation give information on the **spread** or **variability** or **dispersion** of the data values.



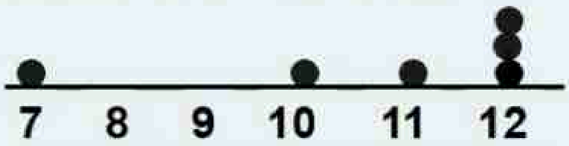
Same center, different variation

Measures of Variation: Why The Range Can Be Misleading

- Ignores the way in which data are distributed



Range = 12 - 7 = 5



Range = 12 - 7 = 5

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**5**

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**120**

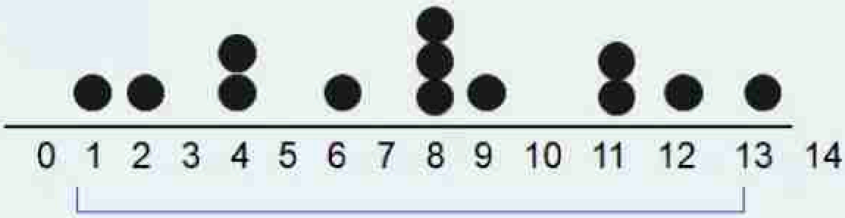
Range = 120 - 1 = 119

Measures of Variation: The Range

- Simplest measure of variation
- Difference between the largest and the smallest values:

$$\text{Range} = X_{largest} - X_{smallest}$$

Example:



Range = 13 - 1 = 12

Interfractile range Median is 0.5 fractile.

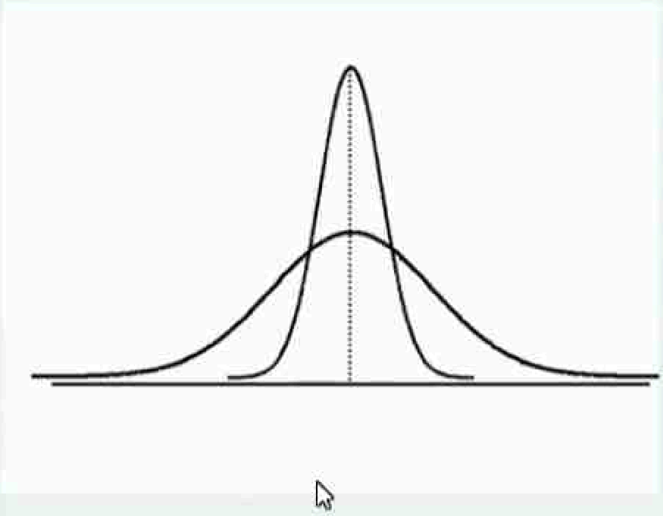
First third	Second third	Last third
863		1138
	1698	
903		1204
	1745	
957		1354
	1802	
1041		1624
	1883	

1/3 fractile

2/3 fractile

If we divide data in??????? deciles , quartile and percentile

Measures of Variation: The Sample Variance



Low variation: more points close to the mean
High variation: more points far from the mean

So, measure the distance to the mean

Measures of Variation: The Sample Variance

- Average (approximately) of squared deviations of values from the mean
  - Sample variance:

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

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

Suppose we draw n independent observations from a population with mean μ and variance σ^2.

Usually unknown

Usually unknown

The sample mean $\bar{x}$ estimates the population mean μ.

The sample variance s^2 estimates the population variance σ^2.

Ideally we would estimate σ^2 with:

$$\frac{\sum(x_i - \mu)^2}{n}$$

This is the average squared distance from the true mean

Problem: μ is unknown!

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

On average, this estimator equals the population variance σ^2.

We could try:

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

Tends to underestimate σ^2