Descriptive Statistics

upGrad knowledge hut

Learning Objectives

- Describe the key features of a dataset with the measures (mean, median, and mode)
- Explain Skewness
- Discuss Range and IQR
- Differentiate sample and population
- Explain the variance and standard deviation
- Discuss impact of scaling and shifting

Intro

Intro

Descriptive statistics use measures that describe some key features of the data.

These are two different measures of central tendency.

- Average
- Most frequent value

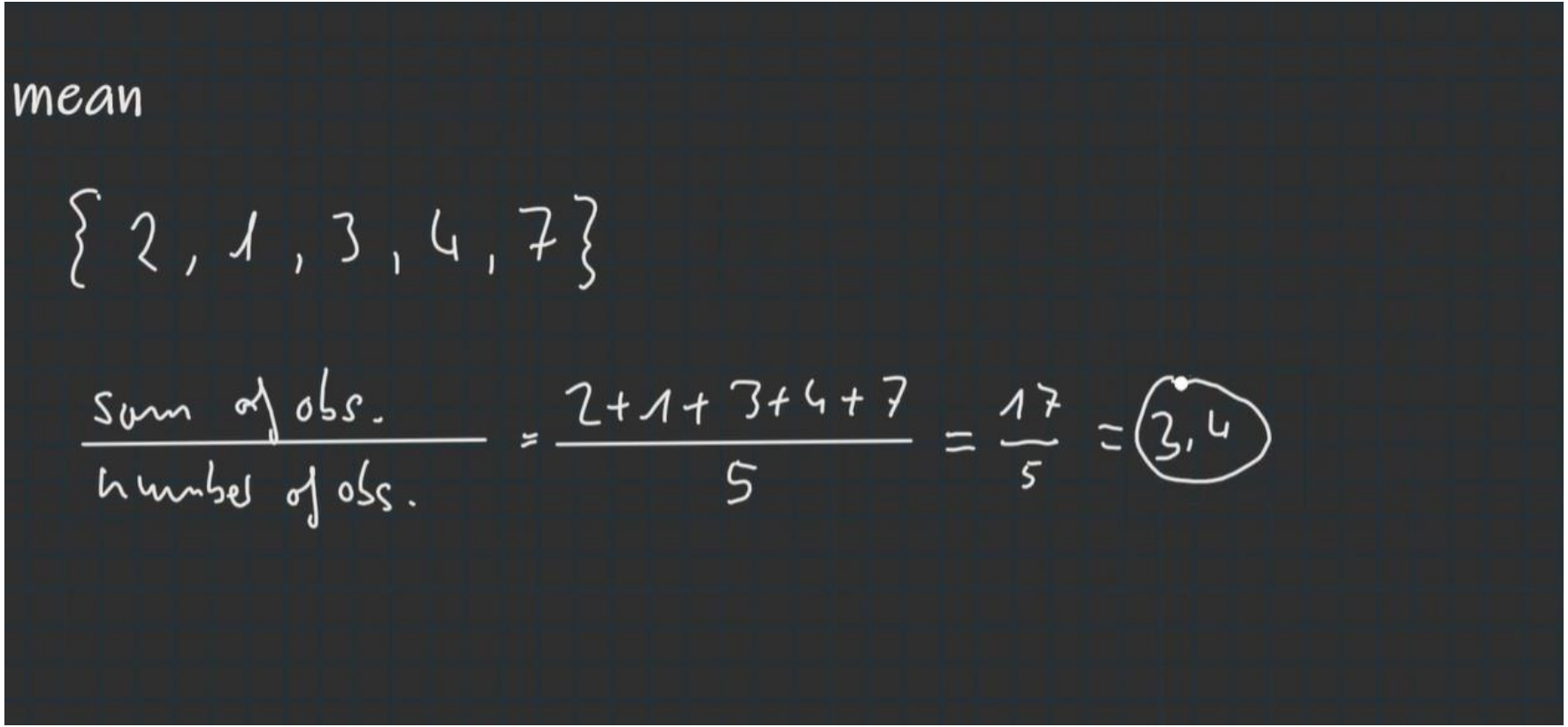
Mean

Mean

- Mean in common language is called average, and mean is just a mathematical term.
- The mean is calculated by the sum of observations and divided by the number of observations.
- Formula in mathematical equation form

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

Mean - Example



Ms. Excel

Syntax

AVERAGE(number1, [number2], ...)

The AVERAGE function syntax has the following arguments:

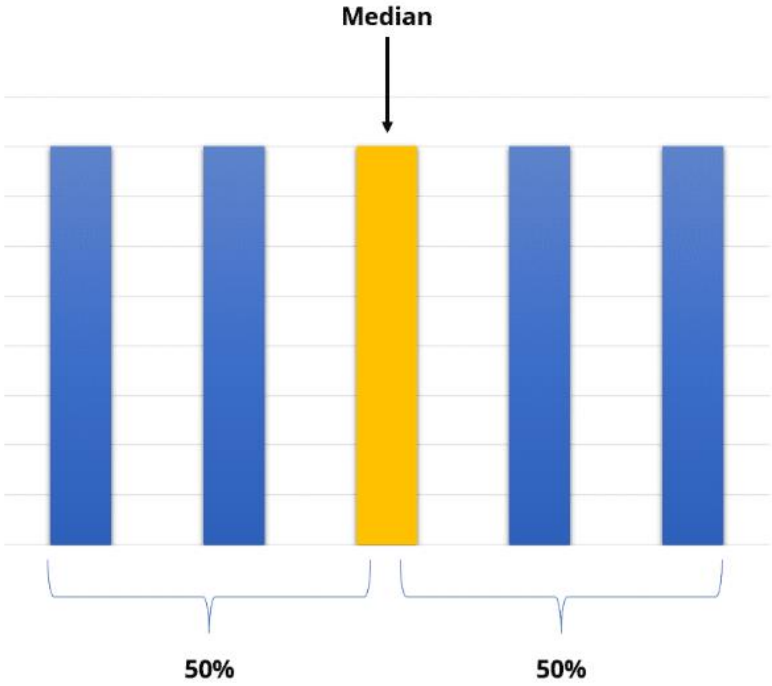
- **Number1** Required. The first number, cell reference, or range for which you want the average.
- **Number2, ...** Optional. Additional numbers, cell references or ranges for which you want the average, up to a maximum of 255.

Median

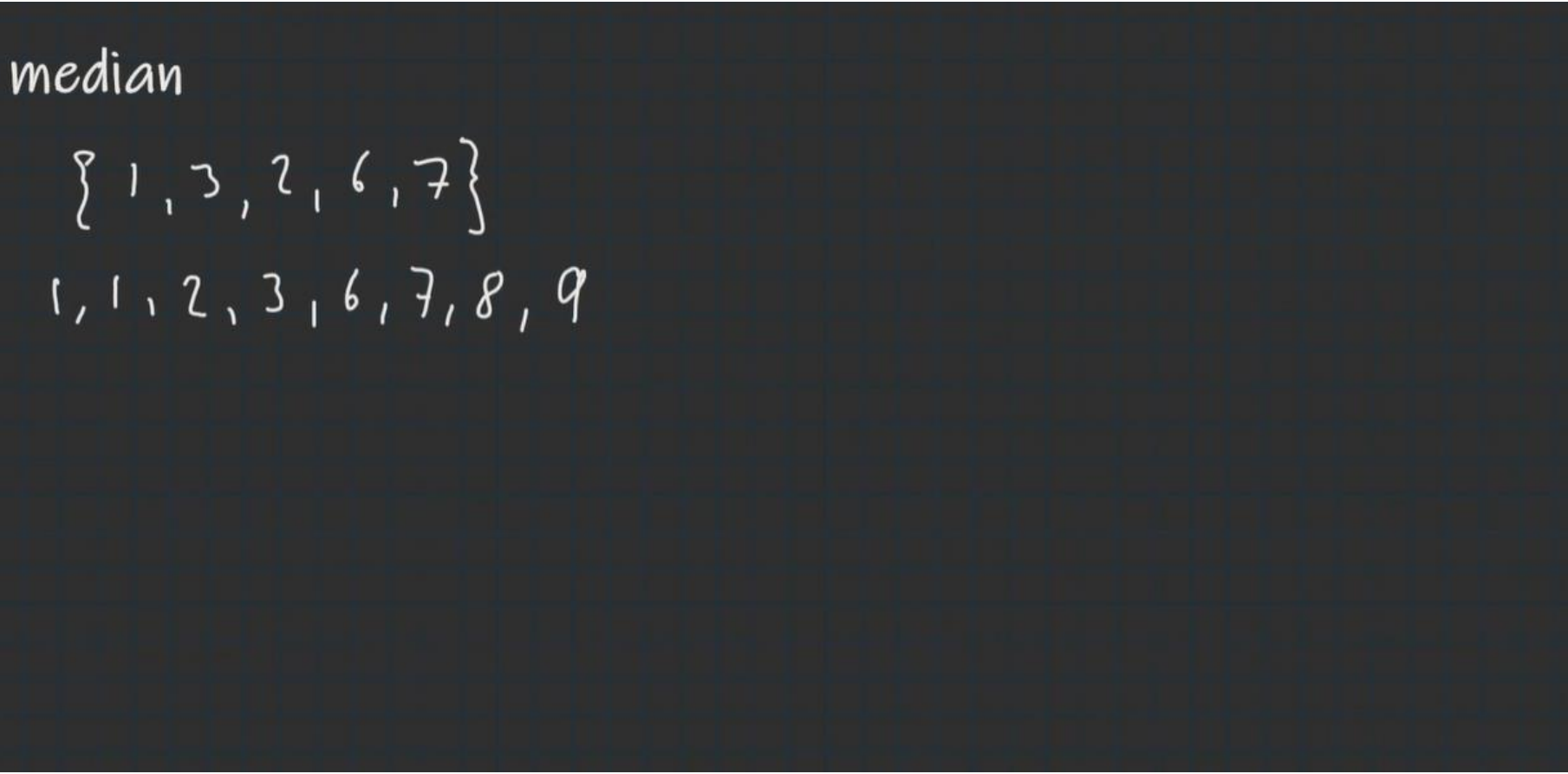
Median

- If there are an odd number of values, find one number in the middle with the same number on the left side as the numbers on the right side, and the result will be the median.

- If there are an even number of values, then there will be two numbers in the middle; add them up together and divide by two, and the result will be the median.



Median - Example



Ms. Excel

Syntax

MEDIAN(number1, [number2], ...)

The MEDIAN function syntax has the following arguments:

- **Number1, number2, ...** Number1 is required, subsequent numbers are optional. 1 to 255 numbers for which you want the median.

Mode

Mode

- To calculate the mode, count the occurrence of the numbers in the data set.

- {1,1,2,3,4,8}
  In this case, the number 1 occurs twice, whereas all other numbers occur once. 1 is the mode.

- {1,1,2,3,4,8,8}
  In this case, the number 1 and 8 occurs twice, whereas all of the numbers occur once. This data set has no mode, or it is a bi-modal.

Ms. Excel

Syntax

MODE(number1,[number2],...)

The MODE function syntax has the following arguments:

- **Number1** Required. The first number argument for which you want to calculate the mode.
- **Number2,...** Optional. Number arguments 2 to 255 for which you want to calculate the mode. You can also use a single array or a reference to an array instead of arguments separated by commas.

Ms. Excel

The MODE function measures central tendency, which is the location of the center of a group of numbers in a statistical distribution. The three most common measures of central tendency are:

- **Average** which is the arithmetic mean, and is calculated by adding a group of numbers and then dividing by the count of those numbers. For example, the average of 2, 3, 3, 5, 7, and 10 is 30 divided by 6, which is 5.
- **Median** which is the middle number of a group of numbers; that is, half the numbers have values that are greater than the median, and half the numbers have values that are less than the median. For example, the median of 2, 3, 3, 5, 7, and 10 is 4.
- **Mode** which is the most frequently occurring number in a group of numbers. For example, the mode of 2, 3, 3, 5, 7, and 10 is 3.

For a symmetrical distribution of a group of numbers, these three measures of central tendency are all the same. For a skewed distribution of a group of numbers, they can be different.

Mean or Median?

Mean or Median

- Use the median if the mean does not work that well.
- The mean does not work If the dataset is not symmetrical, but it's skewed.
- The most common case is if the values are categorical when using the mode instead of the mean or the median.

Mean or Median?

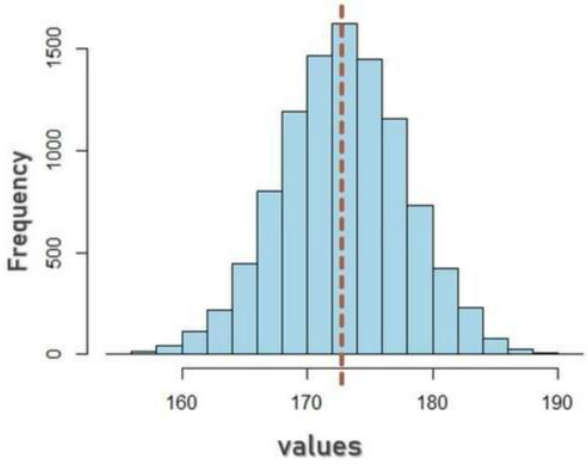
Example: Product price

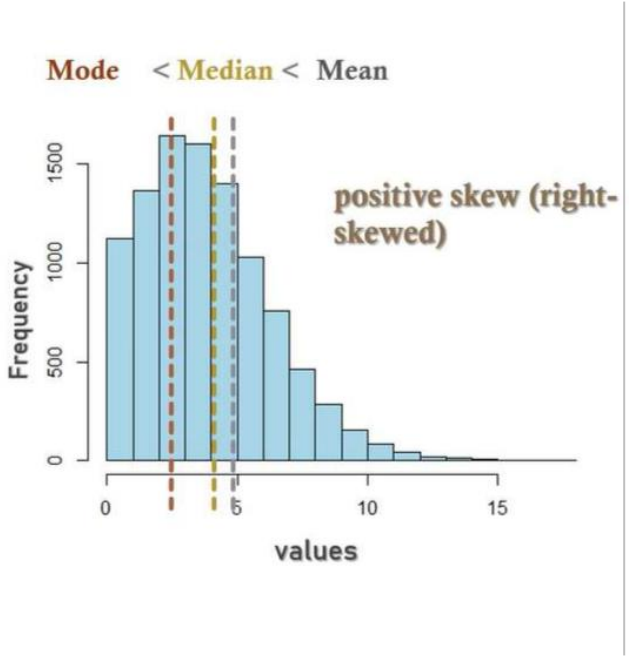
Price
21
4
9
14
29
2
194

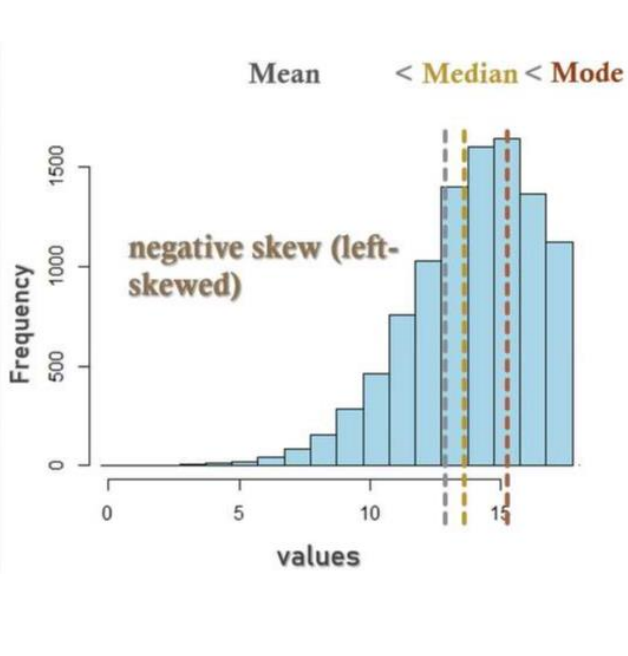
Skewness

Skewness

The skewness measures the symmetry of the dataset.







Formula for Calculation of the Skewness

$$\gamma_m = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sigma^3}$$

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}^3}$$

$\gamma_m > 0$ positive skew (right-skewed)

$\gamma_m < 0$ negative skew (left-skewed)

Range and IQR

Range and IQR

To calculate the range:

- Find the highest and lowest number in the dataset, then subtract the minimum value from the maximum value.

To calculate the interquartile range:

- First, bring the numbers into order.
- Then divide the numbers in half.
- After drawing a line exactly in the middle, find the number in the left partition's middle. Also, the same thing is on the right side.
- Now, subtract the left number from the right, which is the result for the interquartile range.

Range and IQR (Interquartile range)

13	6
14	7

4
3
4
4
2
2
2
3
2
3
1
6
1
3

$6 - 1 = 5$

1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 6

$4 - 2 = 2$

2, 2, 3, 4, 5, 5, 6, 6, 6, 6, 7, 7

$\frac{3+4}{2} = \frac{7}{2} = 3.5$

$\frac{6+6}{2} = \frac{12}{2} = 6$

$6 - 3.5 = 2.5$

Ms. Excel IQR

IQR = Quartile 3 - Quartile 1

QUARTILE(array,quart)

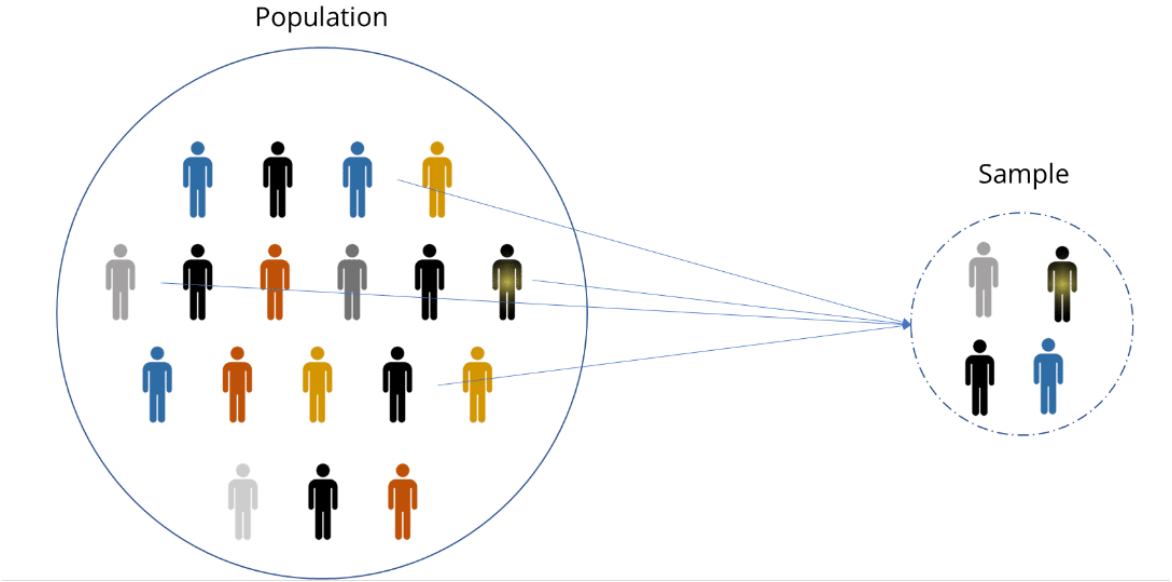
The QUARTILE function syntax has the following arguments:

- **Array** Required. The array or cell range of numeric values for which you want the quartile value.
- **Quart** Required. Indicates which value to return.

If quart equals	QUARTILE returns
0	Minimum value
1	First quartile (25th percentile)
2	Median value (50th percentile)
3	Third quartile (75th percentile)
4	Maximum value

Sample vs. Population

Sample vs. Population



Parameters of population

- Mean or average (μ)
- Population size (N)
- Variance (σ^2)

Parameters of sample

- Mean or average ($\bar{X}$)
- Population size (n)
- Variance (S^2)

Variance and Standard Deviation

Variance and Standard Deviation

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Sample Variance

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

Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$

Sample Standard Deviation

$$s = \sqrt{s^2}$$

Variance & Standard Deviation

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$\bar{x} = 3 = \frac{\sum_{i=1}^{5} x_i}{5}$$

0, 1, 3, 5, 6

$$\bar{x} = 3 \quad s^2 = \frac{(0-3)^2 + (1-3)^2 + (3-3)^2 + (5-3)^2 + (6-3)^2}{4}$$

Ms. Excel

Syntax

VAR.S(number1,[number2],...)

The VAR.S function syntax has the following arguments:

- **Number1** Required. The first number argument corresponding to a sample of a population.
- **Number2, ...** Optional. Number arguments 2 to 254 corresponding to a sample of a population.

Syntax

VAR.P(number1,[number2],...)

The VAR.P function syntax has the following arguments:

- **Number1** Required. The first number argument corresponding to a population.
- **Number2, ...** Optional. Number arguments 2 to 254 corresponding to a population.

Impact of Scaling and Shifting

Impact of Scaling and Shifting

- Scaling means multiplying all of the values in the data set with a certain value.
- Shifting means adding or subtracting a certain value.

Impact of scaling & shifting

scaling: × / ÷

shifting: + / −

1, 2, 4, 5, 7, 5 —(+3)→ 4, 5, 7, 8, 10, 8

mean	4	+3	$\frac{42}{6} = 7$
median	4.5	+3	7.5
mode	5	+3	8
range	6	no change	6
IQR	3	no change	3
variance	4		4

Impact of scaling & shifting

scaling: × / ÷

shifting: + / −

1, 2, 4, 5, 7, 5 → ×2 → 2, 4, 8, 10, 14, 10

mean	4	× 2	$\frac{48}{6} = 8$
median	4.5	× 2	9
mode	5	× 2	10
range	6	× 2	12
IQR	3	× 2	6
variance	4	$\times 2^2$	16

Thank you