INDIAN INSTITUTE OF TECHNOLOGY MADRAS

सिद्धिर्भवति कर्मजा

IIT Madras

ONLINE DEGREE

Statistics for Data Science -1

Lecture 3.4: Describing Numerical Data- Measures of dispersion

Usha Mohan

Indian Institute of Technology Madras

Introduction- why do we need a measure of dispersion

- Consider the two data sets given below
  - Dataset 1: 3, 3, 3, 3, 3
  - Dataset 2: 1, 2, 3, 4, 5

Introduction- why do we need a measure of dispersion

- Consider the two data sets given below
  - Dataset 1: 3, 3, 3, 3, 3
  - Dataset 2: 1, 2, 3, 4, 5
- The measures of central tendency for both the data sets are

	Dataset 1	Dataset 2

Introduction- why do we need a measure of dispersion

- Consider the two data sets given below
  - Dataset 1: 3, 3, 3, 3, 3
  - Dataset 2: 1, 2, 3, 4, 5
- The measures of central tendency for both the data sets are

	Dataset 1	Dataset 2
Mean	3	3

Introduction- why do we need a measure of dispersion

- Consider the two data sets given below
  - Dataset 1: 3, 3, 3, 3, 3
  - Dataset 2: 1, 2, 3, 4, 5
- The measures of central tendency for both the data sets are

	Dataset 1	Dataset 2
Mean	3	3
Median	3	3

Introduction- why do we need a measure of dispersion

- Consider the two data sets given below
  - Dataset 1: 3, 3, 3, 3, 3
  - Dataset 2: 1, 2, 3, 4, 5
- The measures of central tendency for both the data sets are

	Dataset 1	Dataset 2
Mean	3	3
Median	3	3
Mode	3	Not available

Introduction- why do we need a measure of dispersion

- Consider the two data sets given below
  - Dataset 1: 3, 3, 3, 3, 3
  - Dataset 2: 1, 2, 3, 4, 5
- The measures of central tendency for both the data sets are

	Dataset 1	Dataset 2
Mean	3	3
Median	3	3
Mode	3	Not available

- The mean, median are same for both the datasets. However, the datasets are not same. They are different.

Measures of dispersion

- To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set.

Measures of dispersion

- To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set.
- Such descriptive measures are referred to as
  - measures of dispersion, or
  - measures of variation, or
  - measures of spread.

Measures of dispersion

- To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set.
- Such descriptive measures are referred to as
  - measures of dispersion, or
  - measures of variation, or
  - measures of spread.
- In this course we will be discussing about the following measures of dispersion.

Measures of dispersion

- To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set.
- Such descriptive measures are referred to as
  - measures of dispersion, or
  - measures of variation, or
  - measures of spread.
- In this course we will be discussing about the following measures of dispersion.
  1. Range.

Measures of dispersion

- To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set.
- Such descriptive measures are referred to as
  - measures of dispersion, or
  - measures of variation, or
  - measures of spread.
- In this course we will be discussing about the following measures of dispersion.
  1. Range.
  2. Variance.

Measures of dispersion

- To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set.
- Such descriptive measures are referred to as
  - measures of dispersion, or
  - measures of variation, or
  - measures of spread.
- In this course we will be discussing about the following measures of dispersion.
  1. Range.
  2. Variance.
  3. Standard deviation.

Measures of dispersion

- To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set.
- Such descriptive measures are referred to as
  - measures of dispersion, or
  - measures of variation, or
  - measures of spread.
- In this course we will be discussing about the following measures of dispersion.
  1. Range.
  2. Variance.
  3. Standard deviation.
  4. Interquartile range.

Range

Definition

The range of a data set is the difference between its largest and smallest values.

Range

Definition

The range of a data set is the difference between its largest and smallest values.

- The range of a data set is given by the formula

$$\text{Range} = \text{Max - Min}$$

where Max and Min denote the maximum and minimum observations, respectively.

Range

Definition

The range of a data set is the difference between its largest and smallest values.

- The range of a data set is given by the formula

$$\text{Range} = \text{Max - Min}$$

where Max and Min denote the maximum and minimum observations, respectively.

	Dataset 1	Dataset 2
	3,3,3,3,3	1,2,3,4,5

Range

Definition

The range of a data set is the difference between its largest and smallest values.

- The range of a data set is given by the formula

$$\text{Range} = \text{Max} - \text{Min}$$

where Max and Min denote the maximum and minimum observations, respectively.

	Dataset 1	Dataset 2
	3,3,3,3,3	1,2,3,4,5
Max	3	5

Range

Definition

The range of a data set is the difference between its largest and smallest values.

- The range of a data set is given by the formula

$$\text{Range} = \text{Max - Min}$$

where Max and Min denote the maximum and minimum observations, respectively.

-

	Dataset 1	Dataset 2
	3,3,3,3,3	1,2,3,4,5
Max	3	5
Min	3	1

Range

Definition

The range of a data set is the difference between its largest and smallest values.

- The range of a data set is given by the formula

$$\text{Range} = \text{Max} - \text{Min}$$

where Max and Min denote the maximum and minimum observations, respectively.

	Dataset 1	Dataset 2
	3,3,3,3,3	1,2,3,4,5
Max	3	5
Min	3	1
Range	0	4

Range sensitive to outliers

- Range is sensitive to outliers. For example consider two datasets as given below

Range sensitive to outliers

- Range is sensitive to outliers. For example consider two datasets as given below

	Dataset 1	Dataset 2
	1,2,3,4,5	1,2,3,4,15

Range sensitive to outliers

- Range is sensitive to outliers. For example consider two datasets as given below

	Dataset 1	Dataset 2
	1,2,3,4,5	1,2,3,4,15
Max	5	15

Range sensitive to outliers

- Range is sensitive to outliers. For example consider two datasets as given below

	Dataset 1	Dataset 2
	1,2,3,4,5	1,2,3,4,15
Max	5	15
Min	1	1

Range sensitive to outliers

- Range is sensitive to outliers. For example consider two datasets as given below

	Dataset 1	Dataset 2
	1,2,3,4,5	1,2,3,4,15
Max	5	15
Min	1	1
Range	4	14

Range sensitive to outliers

- Range is sensitive to outliers. For example consider two datasets as given below

	Dataset 1	Dataset 2
	1,2,3,4,5	1,2,3,4,15
Max	5	15
Min	1	1
Range	4	14

- Though the two datasets differ only in one datapoint, we can see that this contributes to the value of Range significantly. This happens because the range takes into consideration only the Min and Max of the dataset.

Variance

Variance

- In contrast to the Range, the variance takes into account all the observations.

Variance

- In contrast to the Range, the variance takes into account all the observations.
- One way of measuring the variability of a data set is to consider the deviations of the data values from a central value

Population variance and sample variance

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the dataset has n observations.

- The variance is computed using the following formulae

Population variance and sample variance

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the dataset has n observations.

- The variance is computed using the following formulae
  - Population variance: $\sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N}$

Population variance and sample variance

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the dataset has n observations.

- The variance is computed using the following formulae
  - Population variance: $\sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N}$
  - Sample variance: $s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}$

Population variance and sample variance

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the dataset has n observations.

- The variance is computed using the following formulae
  - Population variance: $\sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N}$
  - Sample variance: $s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}$
- The numerator is the sum of squared deviations of every observation from its mean.

Population variance and sample variance

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the dataset has n observations.

- The variance is computed using the following formulae
  - Population variance: $\sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N}$
  - Sample variance: $s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}$
- The numerator is the sum of squared deviations of every observation from its mean.
- The denominator for computing population variance is N, the total number of observations.

Population variance and sample variance

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the dataset has n observations.

- The variance is computed using the following formulae
  - Population variance: $\sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N}$
  - Sample variance: $s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}$
- The numerator is the sum of squared deviations of every observation from its mean.
- The denominator for computing population variance is N, the total number of observations.
- The denominator for computing sample variance is $(n-1)$. The reason for this will be clear in forthcoming courses on statistics.

Example

- Recall marks of students obtained by ten students in an exam is
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66

Example

- Recall marks of students obtained by ten students in an exam is
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66
- The mean was computed to be 59.

Example

- Recall marks of students obtained by ten students in an exam is
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66
- The mean was computed to be 59.
- The deviations of each data point from its mean is given in the table below:

	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	9	
2	79	20	
3	38	-21	
4	68	9	
5	35	-24	
6	70	11	
7	61	2	
8	47	-12	
9	58	-1	
10	66	-7	
Total	**590**	**0**	

	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	441
4	68	9	81
5	35	-24	576
6	70	11	121
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	-7	49
Total	**590**	**0**	**1898**

	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	441
4	68	9	81
5	35	-24	576
6	70	11	121
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	-7	49
Total	**590**	**0**	**1898**

1. Population variance $= \frac{1898}{10} = 189.8$

	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	441
4	68	9	81
5	35	-24	576
6	70	11	121
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	-7	49
Total	**590**	**0**	**1898**

1. Population variance $= \frac{1898}{10} = 189.8$
2. Sample variance $= \frac{1898}{9} = 210.88$

Adding a constant

Adding a constant

- Let $y_i = x_i + c$ where c is a constant then $new\ variance = old\ variance$

Adding a constant

- Let $y_i = x_i + c$ where c is a constant then
  $new\ \ variance = old\ \ variance$
- Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66. has sample variance 210.88

Adding a constant

- Let $y_i = x_i + c$ where c is a constant then
  $new\ variance = old\ variance$
- Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66. has sample variance 210.88
- Suppose the teacher has decided to add 5 marks to each student.
- Then the data is
  73, 84, 43, 73, 40, 75, 66, 52, 63, 71

Adding a constant

- Let $y_i = x_i + c$ where c is a constant then $new\ \ variance = old\ \ variance$
- Example: Recall the marks of students 68,79,38,68,35,70,61,47,58,66. has sample variance 210.88
- Suppose the teacher has decided to add 5 marks to each student.
- Then the data is 73, 84, 43, 73, 40, 75, 66, 52, 63, 71
- The variance of the new dataset is $\frac{1898}{9} = 210.88$
- In general, adding a constant does not change variability of a dataset, and hence it is the same.

Multiplying a constant

Multiplying a constant

- Let $y_i = x_i c$ where c is a constant then

$$new\ variance = c^2 \times old\ variance$$

Multiplying a constant

- Let $y_i = x_i c$ where c is a constant then

$$new\ variance = c^2 \times old\ variance$$

- Example: Recall the marks of students 68,79,38,68,35,70,61,47,58,66.
  We already know variance for this data is 210.88

Multiplying a constant

- Let $y_i = x_i c$ where c is a constant then

$$new\ \ variance = c^2 \times old\ \ variance$$

- Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66.
  We already know variance for this data is 210.88
- Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.

Multiplying a constant

- Let $y_i = x_i c$ where c is a constant then

$$new\ \ variance = c^2 \times old\ \ variance$$

- Example: Recall the marks of students
68,79,38,68,35,70,61,47,58,66.
We already know variance for this data is 210.88
- Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.
- Then the data becomes
27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
The mean of new dataset is 23.6

Multiplying a constant

- Let $y_i = x_i c$ where c is a constant then

$$new\ \ variance = c^2 \times old\ \ variance$$

- Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66.
  We already know variance for this data is 210.88
- Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.
- Then the data becomes
  27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
  The mean of new dataset is 23.6
- The sum of squared deviations from mean = 303.68 and the variance $= \frac{303.68}{9} = 33.74$. We can verify that $33.74 = 0.4^2 \times 210.88$.

Standard deviation

- Another very useful measure of dispersion is the standard deviation.

Definition

The quantity

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

which is the square root of sample variance is the sample standard deviation.

Units of standard deviation

Units of standard deviation

- The sample variance is expressed in units of square units if original variable. For example, instead of marks if the data were weights of 10 students measured in kilograms. Then the unit of variance would be $(kilogram)^2$

Units of standard deviation

- The sample variance is expressed in units of square units if original variable. For example, instead of marks if the data were weights of 10 students measured in kilograms. Then the unit of variance would be $(kilogram)^2$
- The sample standard deviation is measured in the same units as the original data. That is, for instance, if the data are in kilograms, then the units of standard deviation are also in kilograms.

Adding a constant

Adding a constant

- Let $y_i = x_i + c$ where c is a constant then $new\ variance = old\ variance$

Adding a constant

- Let $y_i = x_i + c$ where c is a constant then
  $new\ variance = old\ variance$
- Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66. has sample variance 210.88

Adding a constant

- Let $y_i = x_i + c$ where c is a constant then $new\ variance = old\ variance$
- Example: Recall the marks of students 68,79,38,68,35,70,61,47,58,66. has sample variance 210.88
- Suppose the teacher has decided to add 5 marks to each student.
- Then the data is 73, 84, 43, 73, 40, 75, 66, 52, 63, 71

Adding a constant

- Let $y_i = x_i + c$ where c is a constant then $new\ variance = old\ variance$
- Example: Recall the marks of students 68,79,38,68,35,70,61,47,58,66. has sample variance 210.88
- Suppose the teacher has decided to add 5 marks to each student.
- Then the data is 73, 84, 43, 73, 40, 75, 66, 52, 63, 71
- The variance of the new dataset is $\frac{1898}{9} = 210.88$
- the standard deviation of the new dataset is $\sqrt{210.88} = 14.522$
- In general, adding a constant does not change variability of a dataset, and hence it is the same.

Multiplying a constant

Multiplying a constant

- Let $y_i = x_i c$ where c is a constant then
  $new\ variance = c^2 \times old\ variance$

Multiplying a constant

- Let $y_i = x_i c$ where c is a constant then $new\ \ variance = c^2 \times old\ \ variance$
- Example: Recall the marks of students 68,79,38,68,35,70,61,47,58,66.
  We already know variance for this data is 210.88

Multiplying a constant

- Let $y_i = x_i c$ where c is a constant then
  $new\ variance = c^2 \times old\ variance$
- Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66.
  We already know variance for this data is 210.88
- Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.

Multiplying a constant

- Let $y_i = x_i c$ where c is a constant then
  $new\ variance = c^2 \times old\ variance$
- Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66.
  We already know variance for this data is 210.88
- Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.
- Then the data becomes
  27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
  The mean of new dataset is 23.6

Multiplying a constant

- Let $y_i = x_i c$ where c is a constant then
  $new\ \ variance = c^2 \times old\ \ variance$
- Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66.
  We already know variance for this data is 210.88
- Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.
- Then the data becomes
  27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
  The mean of new dataset is 23.6
- The sum of squared deviations from mean = 303.68 and the variance $= \frac{303.68}{9} = 33.74$.
- The standard deviation of the newdata set is $\sqrt{33.74} = 5.808$.
  We can verify $5.808 = 0.4 \times 14.522$

Section summary

- Measures of dispersion
  1. Range
  2. Variance: population variance and sample variance.
  3. Standard deviation.
- Impact of adding a constant or multiplying with a constant on the measures.