Statistics II

Class 7th April 2020

Measures of dispersion

Meaning of dispersion: Though average gives us one single value representing the entire data, but it can't be adequate in case all the observations are not same. It is thus necessary to present the variability of the series since even though the average of the two given series may be same but they may have wide variation. e.g.

Series A	**Series B**	**Series c**
100	100	1
100	105	2
100	102	5
100	101	9
100	92	483
Total :500	500	500
Average : 100	100	100

Since the average of three given series is 100 but the observations in three series is widely different. Thus it is essential to understand the variations of items which is called dispersion.

Definition: "dispersion is the measure of variation of the items" (Bowley)

"Dispersion or spread is the degree of the scatter or variation of the variable about a central value" (Brooks & Dick)

What are the significance of measuring dispersion?

1. To determine the reliability of an average: if dispersion is small, average is highly representative of the series.
2. It serves as the base to control the variability: in case of quality control in production units

3. To compare two or more series with regard to their variability: higher variation little uniformity.

Methods of measuring dispersion

1. Range
2. Interquartile Range and Quartile Deviation
3. Mean deviation/average deviation
4. Standard deviation

Absolute and relative measures of dispersion: absolute measures of dispersions are expressed in the same unit in which original data is given. But if the two sets of data are represented in two different units or their average size is very different or unmatchable e.g. salary of professionals and labourers, relative measure of dispersion is used in such cases.

Measure of relative dispersion is the ratio of a measure of absolute dispersion to an appropriate average. It is also called **coefficient of dispersion.**

I. Range

It is simplest method of studying dispersion. It is defined as difference between value of largest and smallest item.

Range = L-S

L= Largest item, S= smallest item

Coefficient of range: L-S/L+S

Exercise I. Calculate range of the price fluctuation in a given week. Also calculate coefficient of range.

200, 150, 210, 190, 169, 250, 170

Solution :

Range = L- S, where L= 250, S= 150

Thus range= 250-150 = 100

Coefficient of range: L – S / L+S

= 250 – 150 / 250+ 150

= 100/400 = 0.25

Exercise II. Calculate coefficient of range from the following data

Marks	**No. of students**
10-20	8
20-30	10
30-40	12
40-50	6
50-60	3
60-70	2
70-80	9

(Hint: for continuous series, range is calculated as difference between the upper limit of the upper class and lower limit of the lower class.

Coefficient of range: L –S / L+ S

Complete the exercise.....

Merits of Range:

1. Simplest of all the methods of dispersion and easy to compute
2. Take less time in computation so quick picture about dispersion can be obtained.

Uses of range:

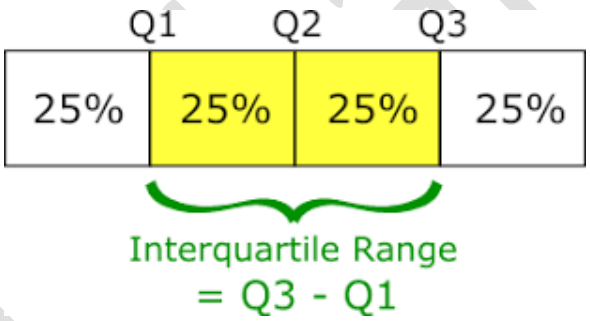
1. Weather forecast: minimum and maximum temperature
2. Share market fluctuations
3. Quality control
4. Everyday life, prices of items

Limitations of range:

1. It does not consider all the observation in the given series, only the two extreme items
2. It cannot tell us anything about the character of distribution
3. Cannot be calculated in case of open ended series.

II. Interquartile range and quartile deviation

Range which includes middle 50% of items in a distribution, one quarter of observations at both the end are excluded in calculating inter-quartile range. Thus interquartile range represents the difference between first quartile and third quartile.



Interquartile range : $\mathbf{Q_3 - Q_1}$

Quartile deviation: If interquartile range is reduced to half i.e. semi-interquartile range is called as quartile deviation.

Quartile deviation (Q.D.) = $\mathbf{Q_3 - Q_1 / 2}$

Q.D. = $\frac{Q3-Q1}{}$

2

If quartile deviation is very small it means high uniformity of central 50% of items.

Coefficient of quartile deviation: $Q_3 - Q_1 / Q_3 + Q_1$

Calculation of quartile deviation:

Individual series

Exercise III: 20, 28, 40,12,30,15,50

Solution :

Arrange the series in ascending or descending order

12, 15, 20, 28, 30, 40, 50

Q1 = size of $N+1/4^{th}$ item

= $7+1/4^{th}$ item i.e. 2^{nd} item = 15

Q3 = size of $3(N+1)/4^{th}$ item

= size of $3 (7+1)/4^{th}$ item

= size of 6^{th} item i.e. 40

$$Q.D. = \frac{Q3-Q1}{2}$$

= 40-15/2

= 12.5

Coefficient of quartile deviation = Q3 - Q1/Q3 + Q1

= 40- 15/ 40+15

= 25 / 55 = 0.455

Discrete series

Exercise IV : Calculate coefficient of quartile deviation from the following data:

Marks	Frequency
10	5
20	7
30	16
40	6
50	10
60	1

Hint :

First Calculate C.f .

Calculate Q3, calculate Q1

Coefficient of quartile deviation : Q3-Q1/ Q3+ Q1

Solution

Marks	Frequency	C.f.
10	5	5
20	7	12
30	16	28
40	6	34
50	10	44
60	1	45
	∑f=45	

Q1= Size of $\frac{N+1_{th}}{4}$ item

$= \frac{45+1}{4}$th item

11.5^{th} item

Since 11.5^{th} item lies in cf12, so Q1= 20

Q3 = **Size of $\frac{3(N+1)_{th}}{4}$ item**

= size of $\frac{3(45+1)}{4}$th item

= 34.5^{th} item

Since 34.5^{th} item lies in cf 44, so Q3= 50

Quartile deviation = $\frac{Q3-Q1}{2}$

Q.D.= $\frac{50-20}{2}$

=15

Coefficient of Q.D.= $\frac{Q3-Q1}{Q3+Q1}$

$= \frac{50-20}{50+20}$

=30/70

= 0.428

Continuous series

Quartile class:

For first quartile: size of $N/4^{th}$ item

For third quartile: size of $3N/4^{th}$ item

$\mathbf{Q_1} = L + \frac{\frac{N}{4} - c.f}{f} \times i$

Q3 = $L + \frac{\frac{3N}{4} - c.f}{f} \times i$

Quartile deviation = Q3 - Q1

Exercise V : Calculate quartile deviation and its coefficient from the given data

Marks	**Frequency**
Less than 35	12
35 - 40	60
40 – 45	90
45 – 50	25
50 - 55	7
55 and above	6

Solution:

Marks	**Frequency**	**c.f.**
30- 35	12	12
35 - 40	60	72
40 – 45	90	162
45 – 50	25	187
50 - 55	7	194
55 -60	6	200
	∑f=200	

Hint :

First calculate cumulative frequency

Determine Q1 class by = size of N/4th item and determine quartile class through C.f.

Calculate Q1 using the formula as given above : $L + \frac{\frac{N}{4} - c.f0}{f} \times i$

Determine Q 3 class by = size of 3N/4th item and determine quartile class through c.f.

Calculate Q3 using the formula as given above : $L + \frac{\frac{3N}{4} - c.f0}{f} \times i$

Calculate quartile deviation = Q3 – Q1

Calculate coefficient of quartile deviation = $\frac{Q3-Q1}{Q3+Q1}$

Merit of quartile deviation

1. It is useful in case of open ended distribution
2. It is also useful in skewed distribution as 1/4th of items at both the ends are excluded.

Limitation

1. Since it ignores 50% of the items so can't be regarded as adequate method of studying dispersion.
2. Not capable of mathematical operations

References for practice questions:

https://www.mathsisfun.com/data/quartiles.html

https://www.wallstreetmojo.com/quartile-deviation/

Class on 13th April 2020

Standard deviation

Standard deviation is most important and widely used measure of studying dispersion. It was introduced by Karl Pearson in 1823. It is also known as root mean square deviation. It is denoted by the Greek letter σ (sigma). Greater the standard deviation, greater would be the variability from the mean in the given series.

Advantages of standard deviation over mean deviation

1. In case of mean deviation algebraic signs are ignored while standard deviation considers algebraic signs.
2. Mean deviation can be computed either from mean or median, but standard deviation is strictly computed from mean.
3. It is best method of dispersion as it is governed by strict mathematical formula.

Individual series

Rule for the formula : root mean square deviation

1. Actual mean method $\sigma = \sqrt{\frac{x^2}{N}}$ where x = $X - \bar{X}$
2. Assumed mean method

$$\sigma = \sqrt{\frac{\sum d2}{N} - (\frac{\Sigma d}{N})^2}$$

where d= X- A , A= assumed mean

Discrete series

3. Actual mean method $\sigma = \sqrt{\frac{\Sigma f x^2}{\Sigma f}}$ where x = $X - \bar{X}$
4. Assumed mean method

$$\sigma = \sqrt{\frac{\sum fd2}{\Sigma f} - (\frac{\Sigma fd}{\Sigma f})^2}$$

Continuous series

Step deviation method

$$\sigma = \sqrt{\frac{\sum fd'2}{\Sigma f} - (\frac{\Sigma fd'}{\Sigma f})^2} \text{ X i}$$

Where d′= d/i, d= (m-A) , i = class interval

Exercise X: Calculate standard deviation from the following data

240,260,290,245,255,288,272,263,277,251

X	**d(X-A) A= 264**	**d^2**
240	-24	576
260	-4	16
290	26	676
245	-19	361
255	-9	81
288	24	576
272	8	64
263	-1	1
277	13	169
251	-13	169
∑X= 2641	∑d= 1	**∑ d^2 =2689**

Solution:

Step 1: calculation of mean

$$\bar{X} = \frac{\sum X}{N}$$

$$= \frac{2641}{10}$$

= 264.1

Since actual mean is in fractions, so we can take assumed mean for calculation of standard deviation

$$\sigma = \sqrt{\frac{\sum d2}{N} - (\frac{\Sigma d}{N})^2}$$

d= (X-A), A is assumed mean
let A= 264

$$\sigma = \sqrt{\frac{2689}{10} - (\frac{1}{10})^2}$$

$$\sigma = \sqrt{268.9 - 0.01}$$

$$\sigma = \sqrt{268.89}$$

σ = 16.398 Ans.

Calculation of standard deviation in discrete series

Exercise XI: Annual salary of a factory workers is given below. Calculate the standard deviation of the salaries.

Salary (X)	No. of persons(f)
45	3
50	5
55	8
60	7
65	9
70	7
75	4
80	7

Solution:

(X)	(f)	d=(X-A)	d^2	fd	f d^2
45	3	-15	225	-45	675
50	5	-10	100	-50	500
55	8	-5	25	-40	200
60	7	0	0	0	0

65	9	5	25	45	225
70	7	10	100	70	700
75	4	15	225	60	900
80	7	20	400	140	2800
	∑f= 50			∑fd=80	$\sum f d^2$= 6000

Let us use assumed mean method

Let A = 60

$$\sigma = \sqrt{\frac{\sum fd2}{\Sigma f} - (\frac{\Sigma fd}{\Sigma f})^2}$$

$$\sigma = \sqrt{\frac{6000}{50} - \left(\frac{80}{50}\right)^2}$$

$$\sigma = \sqrt{120 - (1.6)^2}$$

$\sigma = \sqrt{120-2.56}$

$\sigma = \sqrt{120-2.56}$

$\sigma = \sqrt{117.44}$

10.83 Answer

Exercise XII: Calculate standard deviation from the following:

Age	f
0-10	15
10-20	15
20-30	23
30-40	22
40-50	25
50-60	10
60-70	5
70-80	10

Solution:

X	f	m	d(m-A)	d'=d/i	d'^2	fd'	f d'^2
0-10	15	5	-40	-4	16	-60	240
10-20	15	15	-30	-3	9	-45	135
20-30	23	25	-20	-2	4	-46	92
30-40	22	35	-10	-1	1	-22	22
40-50	25	45	0	0	0	0	0
50-60	10	55	10	1	1	10	10
60-70	5	65	20	2	4	10	20
70-80	10	75	30	3	9	30	90
	$\sum$ f = 125					$\sum$ fd'= -123	$\sum$ fd'^2 =609

$$\sigma = \sqrt{\frac{\sum fd'^2}{\Sigma f} - (\frac{\Sigma fd'}{\Sigma f})^2} \text{ X i}$$

Let A = **45**

$$\sigma = \sqrt{\frac{609}{125} - (\frac{-123}{125})^2} \text{ X } 10$$

$$\sigma = \sqrt{4.87 - (-0.98)^2} \text{ X } 10$$

$$\sigma = \sqrt{4.87 - (0.96)} \text{ X } 10$$

$$\sigma = \sqrt{3.91} \text{ X } 10$$

= 1.97 X10

= 19.76 Answer

Merits of standard deviation

1. It included each and every item and governed by strict formula.
2. It is capable of further algebraic operation and further statistical measures like correlation, less affected by fluctuations of sampling

Limitations

1. As compared to other measures, it involves difficult calculations

Practice question 1:

Calculate standard deviation of age from the given data

Age (X)	No. of persons (f)
20-25	170
25-30	110
30-35	80
35-40	45
40-45	40
45-50	35

Practice question 2

X	**f**
3.5	3
4.5	7
5.5	22

6.5	60
7.5	85
8.5	32
9.5	8

Relationship among various measures of dispersion

$Q.D = \frac{2}{3}\sigma$ $\quad$ $M.D = \frac{4}{5}\sigma$

Class on 16th April 2020

Coefficient of variation

Standard deviation is absolute measure of dispersion, its relative measure is known as coefficient of variation. It is calculated as :

$C.V. = \frac{\sigma}{\bar{X}} \times 100$

Note: always use actual mean as average even if you are using assumed mean to calculate standard deviation.

Coefficient of variation is **used to compare the variation or consistency of two series. Lower the value of coefficient of variation, higher the consistency of the series.**

Step 1 : Find out the mean

Step 2: Find out the standard deviation

Step 3: Use formula $\frac{\sigma}{\bar{X}} \times 100$

Step 4: Compare C.V. of the given series, the series with less CV will be more consistence.

Exercise XIII: Given below are the prices of two commodities over a period of ten days. Find out which commodity is more stable.

Commodity X	**Commodity Y**
30	**110**
59	**105**
50	**107**
54	**103**
57	**108**
59	**105**
51	**107**
52	**101**
48	**103**
50	**101**

Solution

Commodity X			Commodity Y		
Series 1	$x = (X - \bar{x})$	x^2	Series 2	$x = (X - \bar{x})$	x^2
30	-21	441	110	5	25
59	9	81	105	0	0
50	-1	1	107	2	4
54	3	9	103	-2	4
57	6	36	108	3	9
59	8	64	105	0	0
51	0	0	107	2	4
52	1	1	101	-4	16
48	-3	9	103	-2	4
50	-1	1	101	-4	16
$\sum X = 510$		$\sum x^2 = 643$	$\sum X = 1050$		$\sum x^2 = 82$

Calculation for Series 1

Mean = $\frac{\sum X}{N}$

=$\frac{510}{10}$

= 51

Standard deviation : σ = $\sqrt{\frac{\sum x^2}{N}}$ where x = X -$\bar{X}$

Standard deviation : σ = $\sqrt{\frac{643}{10}}$

σ = √64.3

8.02

C.V. for series 1: = $\frac{\sigma}{\bar{X}}$ x 100

= $\frac{8.02}{51}$ x 100

15.72 for series 1

Calculation for Series 2

Mean = $\frac{\sum X}{N}$

Mean = $\frac{1050}{10}$

= 105

Standard deviation : σ = $\sqrt{\frac{\sum x^2}{N}}$ where x = X -$\bar{X}$

Standard deviation : σ = $\sqrt{\frac{82}{10}}$

=2.86

C.V. for series 2: = $\frac{\sigma}{\bar{X}}$ x 100

== $\frac{2.86}{105}$ x 100

=2.72 for series 2

Since coefficient of variation for series 2 is less than series 1. Therefore series 2 is more consistent. Therefore price of commodity Y is more stable.

Exercise XIV: Following data shows marks of a class. Calculate average marks, standard deviation and coefficient of variation of the marks.

Marks	No. of students
20-30	8
30-40	14
40-50	12
50-60	18
60-70	13
70-80	9
80-90	6

Solution:

X	f	m	d=(m-A)	d^2	fd	fd^2
20-30	8					
30-40	14					
40-50	12					
50-60	18					
60-70	13					
70-80	9					
80-90	6					
	$\sum f$				$\sum fd$	$\sum f d^2$

Practice question 1:

Below are the marks scored by two students in a class Calculate coefficient of variation and state which student is more consistent.

X	Y
32	20
28	31
47	48
63	52
71	62
39	95
10	10
60	65
96	37
14	80

Source for more practice question

https://www.mathsisfun.com/data/standard-deviation.html

https://www.superprof.co.uk/resources/academic/maths/probability/normal-distribution/standard-deviation-problems.html

https://www.examsolutions.net/tutorials/exam-questions-continuous-data-standard-deviation/

VARIANCE AND STANDARD DEVIATION

Recall that the range is the difference between the upper and lower limits of the data. While this is important, it does have one major disadvantage. It does not describe the variation among the variables. For instance, both of these sets of data have the same range, yet their values are definitely different.

90, 90, 90, 98, 90 Range = 8

1, 6, 8, 1, 9, 5 Range = 8

To better describe the variation, we will introduce two other measures of variation—*variance* and *standard deviation* (the variance is the square of the standard deviation). These measures tell us how much the actual values differ from the mean. The larger the standard deviation, the more spread out the values. The smaller the standard deviation, the less spread out the values. This measure is particularly helpful to teachers as they try to find whether their students' scores on a certain test are closely related to the class average.

To find the standard deviation of a set of values:

a. Find the mean of the data
b. Find the difference (deviation) between each of the scores and the mean
c. Square each deviation
d. Sum the squares
e. Dividing by one less than the number of values, find the "mean" of this sum (the **variance***)
f. Find the square root of the variance (the **standard deviation**)

*Note: In some books, the variance is found by dividing by n. In statistics it is more useful to divide by n -1.

EXAMPLE

Find the variance and standard deviation of the following scores on an exam:

92, 95, 85, 80, 75, 50

SOLUTION

First we find the mean of the data:

$$\text{Mean} = \frac{92+95+85+80+75+50}{6} = \frac{477}{6} = 79.5$$

Then we find the difference between each score and the mean (deviation).

Score	Score - Mean	Difference from mean
92	92 – 79.5	+12.5
95	95 – 79.5	+15.5
85	85 – 79.5	+5.5
80	80 – 79.5	+0.5
75	75 – 79.5	-4.5
50	50 – 79.5	-29.5

Next we square each of these differences and then sum them.

Difference	Difference Squared
+12.5	156.25

+15.5	240.25
+5.5	30.25
+0.5	0.25
-4.5	20.25
-29.5	870.25
Sum of the squares →	1317.50

The sum of the squares is 1317.50.

Next, we find the "mean" of this sum (the variance). $\frac{1317.50}{5}$ = 263.5

Finally, we find the square root of this variance. $\sqrt{263.5} \approx 16.2$

So, the standard deviation of the scores is 16.2; the variance is 263.5.

EXAMPLE

Find the standard deviation of the average temperatures recorded over a five-day period last winter:

18, 22, 19, 25, 12

SOLUTION

This time we will use a table for our calculations.

Temp		Temp – mean = deviation	Deviation squared
18		18 – 19.2 = -1.2	1.44
22		22 – 19.2 = 2.8	7.84
19		19 – 19.2 = -0.2	0.04
25	mean	25 – 19.2 = 5.8	33.64
12	↓	12- 19.2 = -7.2	51.84
96 ÷ 5 = 19.2			94.80 ← sum of squares

To find the variance, we divide 5 – 1 = 4. $\frac{94.8}{4}$ = 23.7

Finally, we find the square root of this variance. $\sqrt{23.7} \approx 4.9$

So the standard deviation for the temperatures recorded is 4.9; the variance is 23.7. Note that the values in the second example were much closer to the mean than those in the first example. This resulted in a smaller standard deviation.

We can write the formula for the standard deviation as $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$

where

$\sum$	means "the sum of"
x_i	represents each value *x* in the date
$\bar{x}$	is the mean of the x_i values
n	is the total of x_i values

PRACTICE PROBLEMS

1. Find the variance and standard deviation for the five states with the most covered bridges:
   Oregon: 106 Vermont: 121 Indiana: 152 Ohio: 234 Pennsylvania: 347

2. Find the variance and standard deviation of the heights of five tallest skyscrapers in the United States:
   Sears Tower (Willis Building): 1450 feet Empire State Building: 1250 feet
   One World Trade Center: 1776 feet Trump Tower: 1388 feet 2 World Trade Center: 1340 feet

3. Find the variance and standard deviation of the scores on the most recent reading test:
   7.7, 7.4, 7.3, and 7.9

4. Find the variance and standard deviation of the highest temperatures recorded in eight specific states:
   112, 100, 127, 120, 134, 118, 105, and 110.

SOLUTIONS

1. 9956.5; 99.8 2. 40,449.2; 201.12 3. 0.076; 0.275 4. 127.6; 11.3