lecture 6. Measures of Dispersion and Variability

Variance:

Variance in statistics is a measurement of the spread between numbers in a data set. That is, it measures how far each number in the set is from the mean and therefore from every other number in the set, so Variance defined as the average of the

squared differences from the mean. Variance measures how far a data set is spread out:

$$V = \frac{\sum_{i=1}^{n}(Xi - \bar{X})^2}{N}$$

Where:

V = Variance

xi = Value of each data point

$\bar{x}$ = Mean

N = Number of data points

Variance can be negative. A zero value means that all of the values within a data set are identical.

If the variance is low that's mean the data collect near average, while If the variance is high the data will spread from the average.

Problem 1:

The heights (in cm) of students of a class is given to be 163, 158, 167, 174, 148. Find the variance.

Solution:

To find the variance, we need to find the mean of the given data and total members in the data set.

Total number of elements, N = 5

$$\bar{X} = \frac{163 + 158 + 167 + 174 + 148}{5} = 162$$

The formula for variance is,

$$\mathrm{V} = \frac{\sum_{i=1}^{n}(Xi - \bar{X})^2}{N}$$

Now putting the values in the formula we get,

$$\mathrm{V} = \frac{(162-163)^2 + (158-163)^2 + (167-163)^2 + (174-163)^2 + (148-163)^2}{5}$$

$$\mathrm{V} = \frac{(-1)^2 + (-5)^2 + (4)^2 + (11)^2 + (-15)^2}{5} = 77.6$$

Hence, the variance is found to be 77.6

variance for grouped data

Find the variance of the following data:

$$\text{variance} = \frac{\sum_{i=1}^{n} fi(Xi - \bar{X})^2}{\sum_{i=1}^{n} fi}$$

$$\bar{x} = \frac{\sum fi.xi}{\sum fi}$$

Find the variance of the following data:

Classes	Frequency(f)
30-34	4
35-39	5
40-44	2
45-49	9
TOTAL	***20***

Solution:

Classes	Frequency(fi)	Mid classes (Xi)	fi. Xi	xi - mean	$(xi\text{-}mean)^2$	$f.(xi\text{-}mean)^2$
30-34	4	32	128	32-41= -9	81	4*81= 324
35-39	5	37	185	37-41= -4	16	5*16=80
40-44	2	42	84	42-41= 1	1	2*1=2
45-49	9	47	423	47-41= 6	36	9*36= 324
TOTAL	***20***					***730***

$$\text{Mean } \bar{x} = \frac{\sum \mathbf{fi}*\mathbf{xi}}{\sum \mathbf{fi}} = \frac{820}{20} = 41$$

$$\text{variance} = \frac{\sum_{i=1}^{n} fi(Xi - \bar{X})^2}{\sum_{i=1}^{n} fi} = \frac{730}{20} = 36.5$$

Standard deviation

Standard deviation is a measure of dispersement in statistics. “Dispersement” tells you how much your data is spread out. Specifically, it shows you how much your data is spread out around the mean or average.

It is the most robust and widely used measure of dispersion since, unlike the range and inter-quartile range, it takes into account every variable in the dataset.

For example, are all your scores close to the average? Or are lots of scores way above (or way below) the average score?

When the values in a dataset are pretty tightly bunched together the standard deviation is small. When the values are spread apart the standard deviation will be relatively large. The standard deviation is usually presented in conjunction with the mean and is measured in the same units.

$$S.D = \sqrt{\frac{\sum_{i=1}^{n}(Xi - \bar{X})^2}{n}}$$

Where:

- xi = Value of each data point
- $\bar{x}$ = Mean
- N = Number of data points

For example, suppose we have five climatic stations and have recorded rainfall in mm as follows (60,47,17,43,30). Calculate the standard deviation for them.

Solution:

$$\bar{X} = \frac{60 + 47 + 17 + 43 + 30}{5} = 39.4$$

so the mean (average) height is 39.4 mm.

Now we calculate each station's difference from the Mean:

$$= \frac{(60 - 39.4)^2 + (47 - 39.4)^2 + (17 - 39.4)^2 + (43 - 39.4)^2 + (30 - 39.4)^2}{5}$$

$$= \frac{424.36 + 57.76 + 501.76 + 12.96 + 88.36}{5}$$

= 217.04

Now the Standard Deviation is

$$S.D = \sqrt{\frac{\sum_{i=1}^{n}(Xi - \bar{X})^2}{n}}$$

$$S.D = \sqrt{217.04}$$

S.D = 14.73

And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (14.73 mm) of the Mean:

So, using the Standard Deviation we have a "standard" way of knowing what is normal rainfall, and what is extra-large rainfall or extra small rainfall.

Standard deviation for grouped data

$$S.D = \sqrt{\frac{\sum_{i=1}^{n} fi(Xi - \bar{X})^2}{\sum_{i=1}^{n} fi}}$$

$$\bar{x} = \frac{\sum fi.xi}{\sum fi}$$

Example: Find the standard deviation of the following data:

Classes	Frequency(f)
30-34	4
35-39	5
40-44	2
45-49	9
TOTAL	***20***

Solution:

Classes	Frequency(fi)	Mid classes (xi)	fi.xi	xi- mean	$(x\text{-mean})^2$	$fi.(xi\text{-mean})^2$
30-34	4	32	128	32-41= -9	81	4*81= 324
35-39	5	37	185	37-41= -4	16	5*16=80
40-44	2	42	84	42-41= 1	1	2*1=2
45-49	9	47	423	47-41= 6	36	9*36= 324
TOTAL	***20***					***730***

$$\text{Mean } \bar{x} = \frac{\sum \mathbf{fi}*\mathbf{xi}}{\sum \mathbf{fi}} = \frac{820}{20} = 41$$

$$S.D = \sqrt{\frac{\sum_{i=1}^{n} fi(Xi - \bar{X})^2}{\sum_{i=1}^{n} fi}} = \sqrt{\frac{730}{20}} = 6.04$$

Coefficient of variation (CV):

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.

The coefficient of variation is helpful when using the risk/reward ratio to select investments. For example, in finance, the coefficient of variation allows investors to determine how much volatility, or risk, is assumed in comparison to the amount of return expected from investments.

Ideally, the coefficient of variation formula should result in a lower ratio of the standard deviation to mean return, meaning the better risk-return trade-off. Note that if the expected return in the denominator is negative or zero, the coefficient of variation could be misleading.

$$\textbf{Coefficient of variation (CV)} = \frac{\textbf{standard deviation}}{\boldsymbol{Mean}}$$

Example: Find CV of {13,35,56,35,77}

Solution:

Number of terms (N) = 5

Mean:

$$\text{Mean} = \frac{13 + 35 + 56 + 35 + 77}{5} = 43.2$$

$$S.D = \sqrt{\frac{\sum_{i=1}^{n}(Xi - \bar{X})^2}{n}}$$

$$S.D = \sqrt{\frac{(13 - 43.2)^2 + (35 - 43.2)^2 + (56 - 43.2)^2 + (35 - 43.2)^2 + (77 - 43.2)^2}{5}}$$

S.D=24.25

$$\text{Coefficient of variation (CV)} = \frac{\text{standard deviation}}{Mean}$$

$$\text{Coefficient of variation (CV)} = \frac{24.25}{43.2}$$

$$\text{Coefficient of variation (CV)} = 0.5614$$

Standard Error:

The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation. In statistics, a sample mean deviates from the actual mean of a population—this deviation is the standard error of the mean.

It is used to measure the amount of accuracy by which the given sample represents its population.

When you take measurements of some quantity in a population, it is good to know how well your measurements will approximate the entire population.

A large standard error would mean that there is a lot of variability in the population, so different samples would give you different mean values.

A small standard error would mean that the population is more uniform, so your sample mean is likely to be close to the population mean.

$$\text{Standard Error (SE)} = \frac{Standard\ Deviation}{\sqrt{N}}$$

Where: N is the number of observation.

Example

Calculate the standard error of the given data:

(5, 10, 12, 15, 20)

Solution: First we have to find the mean of the given data;

Mean = (5+10+12+15+20)/5 = 62/5 = 10.5

Now, the standard deviation can be calculated as;

$$S = \sqrt{\frac{(5-10.5)^2 + (10-10.5)^2 + (12-10.5)^2 + (15-10.5)^2 + (20-10.5)^2}{5}}$$

After solving the above equation, we get;

S = 5.35

Therefore, SE can be estimated with the formula;

$$\text{Standard Error (SE)} = \frac{Standard\ Deviation}{\sqrt{N}}$$

$$\text{SE} = \frac{5.35}{\sqrt{5}} = 2.39$$

In statistics and mathematics, the deviation is a measure that is used to find the difference between the observed value and the expected value of a variable. In simple words, the deviation is the distance from the center point. Similarly, the mean deviation is used to calculate how far the values fall from the middle of the data set. In this article, let us discuss the definition, formula, and examples in detail.

Mean Deviation Definition

The mean deviation or average deviation is defined as a statistical measure that is used to calculate the average deviation from the mean value of the given data set. The mean deviation of the data values can be easily calculated using the below procedure.

$$M.D = \frac{\sum|Xi - \bar{X}|}{N}$$

Σ: represents the addition of values

X: represents each value in the data set

$\bar{X}$: represents the mean value of the data set

N: represents the number of data values

| | : represents the absolute value, which ignores the "-" symbol

$$M.D = \frac{\sum|Xi - \bar{X}| * fi}{\sum fi}$$

Determine the mean deviation for the data values

13 ،7 ،15 ،6 ،9 ،12 ،8

$$\bar{x} = \frac{\sum x_i}{n} = \frac{70}{7} = 10$$

x_i	$x_i - \bar{x}$	$\left\|x_i - \bar{x}\right\|$
8	**-2**	**2**
12	**2**	**2**
9	**-1**	**1**
6	**-4**	**4**
15	**5**	**5**
7	**-3**	**3**
13	**3**	**3**
$\sum 70$	**0**	**20**

$$\therefore M.D = \frac{\sum\left|x_i - \bar{x}\right|}{n} = \frac{20}{7} = 2.85$$

Calculate the mean deviation about the mean for the given data.

16−14	14−12	12−10	10−8	8−6	6−4	**Classes**
2	7	6	8	5	3	**Frequency**

Solution

C	F_i	x_i	$F_i x_i$	$\bar{X} - X_i$	$\left\|\bar{x} - x_i\right\|$	$F_i\left\|\bar{x} - x_i\right\|$
6−4	**3**	**5**	**15**	**5−**	**5**	**15**
8−6	**5**	**7**	**35**	**3−**	**3**	**15**
10−8	**8**	**9**	**72**	**1−**	**1**	**8**
12−10	**6**	**11**	**66**	**1**	**1**	**6**
14−12	**7**	**13**	**91**	**3**	**3**	**21**
16−14	**2**	**15**	**30**	**5**	**5**	**10**
المجموع	**31**		**309**	**0**		**75**

$$\overline{X} = \frac{\sum f_i x_i}{\sum f_i} = \frac{309}{31} = 10$$

$$M.D = \frac{\sum f_i \left| x_i - \overline{x} \right|}{\sum f_i} = \frac{75}{31} = 2.4$$

Advantages and disadvantages of measures of dispersion

Measures of Variability	Advantages	Disadvantages
Range	It is easier to compute	The value of range is affected by only two extreme scores
	It can be used as a measure of variability where precision is not required	It is not very stable from sample to sample
		It is not sensitive to total condition of the distribution
		It is dependent on sample size, being greater when sample size is greater
Inter quartile Range	It is less sensitive to the presence of a few very extreme scores than is standard deviation	The sampling stability of IQR is good but it is not up to that of standard deviation
	If the distribution is skewed, IQR is a good measure of variation.	
Standard Deviation	It is resistant to sampling variation It is of high use both in descriptive and inferential statistics	It is responsive to exact position of each score in the distribution
		It is more sensitive than IQR to the presence of few extreme scores in the distribution
Variance	provide a summary of individual observations around the mean	sensitive to outliers
Coefficient of variation	used to compare two or more distribution that have different means	Does not vary with the magnitude of the mean

Measures of Shape: Skewness and Kurtosis

The measure of central tendency and measure of dispersion can describe the distribution but they are not sufficient to describe the nature of the distribution.

For this purpose, we use other two statistical measures that compare the shape to the normal curve called Skewness and Kurtosis.

Skewness and Kurtosis are the two important characteristics of distribution that are studied in descriptive statistics

1-Skewness

Skewness is a statistical number that tells us if a distribution is symmetric or not. A distribution is symmetric if the right side of the distribution is similar to the left side of the distribution.

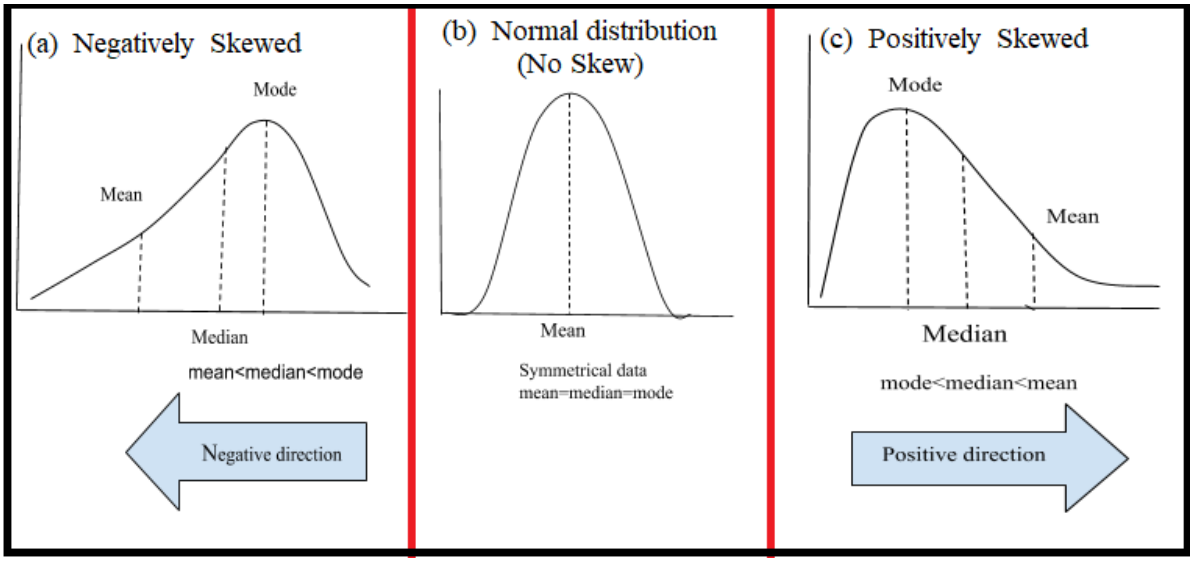
If a distribution is symmetric, then the Skewness value is 0.

i.e. If a distribution is Symmetric (normal distribution):

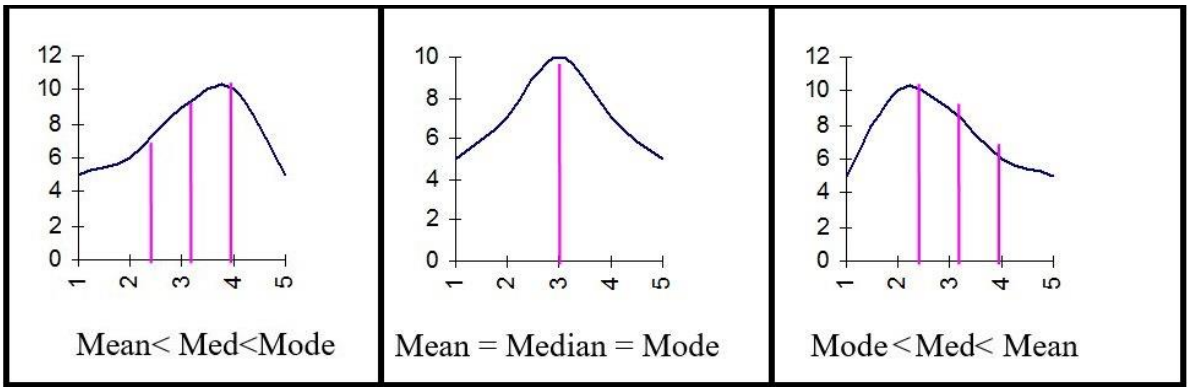
median= mean= mode, (Skewness value is 0)

If Skewness is greater than 0, then it is called right-skewed or that the right tail is longer than the left tail. If Skewness is less than 0, then it is called left-skewed or that the left tail is longer than the right tail.

For example, the symmetrical and skewed distributions are shown by curves as:



And other example



The Formula of Skewness is:

$$\text{Skewness} = \frac{\sum \left(x - \bar{x}\right)^3}{(n-1) \cdot S^3}$$

Where:
S: standard deviation
$\bar{X}$: Mean

Difference between Variance and Skewness

The following two points of difference between variance and skewness should be carefully noted.

1. Variance tells us about the amount of variability while skewness gives the direction of variability.

2. In business and economic series, measures of variation have greater practical application than measures of skewness. However, in the medical and life science field measures of skewness have greater practical applications than the variance.

Karl Pearson's Coefficient of Skewness

Karl Pearson developed two methods to find skewness in a sample.

This method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is given by

1. Pearson's Coefficient of Skewness #1 uses the mode. The formula is:

$$SK = \frac{(\bar{X} - Mo)}{SD}$$

Where $\bar{X}$:the mean,
Mo : the mode
SD: the standard deviation for the sample.

If mode is not well defined, we use the formula

Pearson's Coefficient of Skewness #2 uses the median. The formula is

$$SK = \frac{3(\bar{X} - Md)}{SD}$$

Where $\bar{X}$:the mean,

Mo = the mode,

and SD = the standard deviation for the sample.

It is generally used when you don't know the mode.

The value of this coefficient would be zero in a symmetrical distribution.

If the mean is greater than the mode, the coefficient of skewness would be positive otherwise negative.

The value of Karl Pearson's coefficient of skewness usually lies between ±1 for moderately skewed destitution.

Example:

Use Pearson's Coefficient 1 and 2 to find the skewness for data with the following characteristics:

- Mean = 70.5
- Median = 80
- Mode = 85

- Standard deviation = 19.33

Pearson's Coefficient of Skewness 1 (Mode):

Step 1: Subtract the mode from the mean: 70.5 – 85 = -14.5

Step 2: Divide by the standard deviation: -14.5 / 19.33 = -0.75

Pearson's Coefficient of Skewness 2 (Median):

Step 1: Subtract the median from the mean: 70.5 – 80 = -9.5

Step 2: Multiply Step 1 by 3: -9.5(3) = -28.5

Step 2: Divide by the standard deviation: -28.5 / 19.33 = -1.47

Remarks about Skewness

1. If the value of mean, median, and mode are the same in any distribution, then the skewness does not exist in that distribution. Larger the difference in these values, the larger the skewness;
2. If sum of the frequencies are equal on both sides of the mode then skewness does not exist;
3. If the distance of the first quartile and third quartile are the same from the median then a skewness does not exist. Similarly, if deciles (first and ninth) and percentiles (first and ninety-nine) are at equal distance from the median. Then there is no asymmetry;
4. If the sums of positive and negative deviations obtained from mean, median, or mode are equal then there is no asymmetry; and
5. If a graph of data become a normal curve and when it is folded at the middle and one-part overlap fully on the other one then there is no asymmetry

2-Kurtosis

Kurtosis is a statistical number that tells us if a distribution is taller or shorter than a normal distribution. If a distribution is similar to the normal distribution, the Kurtosis value is 0. If Kurtosis is greater than 0, then it has a higher peak compared

to the normal distribution. If Kurtosis is less than 0, then it is flatter than a normal distribution.

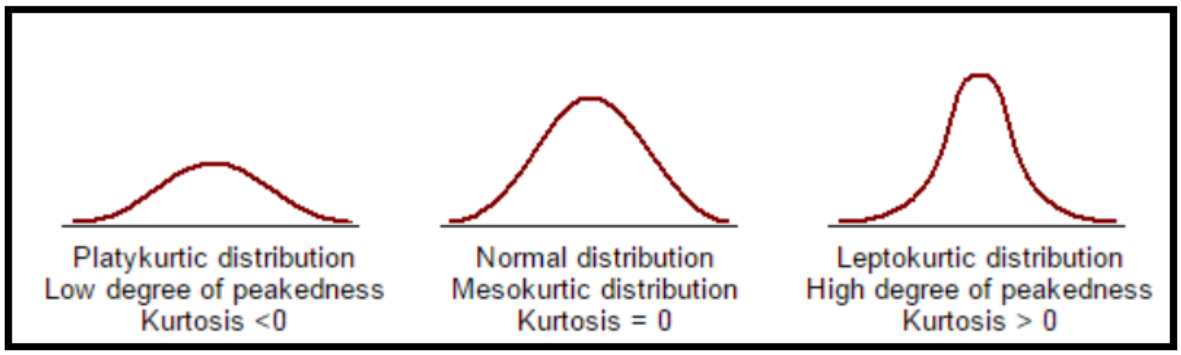
There are three types of distributions:

Leptokurtic: Sharply peaked with fat tails, and less variable.

Mesokurtic: Medium peaked

Platykurtic: Flattest peak and highly dispersed.

For example, The different types of Kurtosis:



The Formula of kurtosis is:

$$\text{Kurtosis} = \frac{\sum \left(x - \bar{x}\right)^4}{(n - 1) \cdot S^4}$$

Where:

S: standard deviation $\bar{X}$: Mean

Examples: Calculate Sample Skewness and Sample Kurtosis from the following grouped data

Class	Frequency
2 - 4	3
4 - 6	4
6 - 8	2
8 - 10	1

Solution:

Classes	Mid value (x)	f	$f \cdot x$	$(x-\bar{x})$	$f \cdot (x-\bar{x})^2$	$f \cdot (x-\bar{x})^3$	$f \cdot (x-\bar{x})^4$
2 - 4	3	3	3×3= 9	3-5.2=-2.2	3×-2.2×-2.2=14.52	14.52×-2.2= -31.944	70.27
4 - 6	5	4	4×5= 20	5-5.2=-0.2	4×-0.2×-0.2=0.16	0.16×-0.2= -0.032	0.0064
6 - 8	7	2	2×7= 14	7-5.2=1.8	2×1.8×1.8=6.48	6.48×1.8=11.664	20.98
8 - 10	9	1	1×9= 9	9-5.2=3.8	1×3.8×3.8=14.44	14.44×3.8= 54.872	208.5
---	---	---	---	---	---	---	
-TOTAL-	--	n=10	$\sum f \cdot x$=52	`--	=35.6	=34.56	=299. 79

$$\text{Mean} = \sum f \cdot x = \frac{\sum f \cdot x}{\sum f} = \frac{52}{10} = 5.2$$

Calculate Standard deviation (S.D)

$$S.D = \sqrt{\frac{\sum_{i=1}^{n} fi(Xi - \bar{X})^2}{\sum_{i=1}^{n} fi}}$$

$$S.D = \sqrt{\frac{35.6}{10}} = 1.88$$

Calculate the Skewness

$$\text{Skewness} = \frac{\sum (x - \bar{x})^3}{(n - 1) \cdot S^3}$$

$$\text{Skewness} = \frac{34.56}{9*(1.88)^3} = 0.48$$

Calculate the Kurtosis:

$$\text{Kurtosis} = \frac{\sum (x - \bar{x})^4}{(n - 1) \cdot S^4}$$

$$\text{Kurtosis} = \frac{299.79}{9*(1.88)^4} = 2.12$$

Key Differences Between Skewness and Kurtosis

This is the fundamental differences between skewness and kurtosis:

1- The characteristic of a frequency distribution that ascertains its symmetry about the mean is called skewness. On the other hand, Kurtosis means the relative pointedness of the standard bell curve, defined by the frequency distribution.

2- Skewness is a measure of the degree of lopsidedness in the frequency distribution. Conversely, kurtosis is a measure of degree of tailedness in the frequency distribution.

3- Skewness is an indicator of lack of symmetry, i.e. both left and right sides of the curve are unequal, with respect to the central point. As against this, kurtosis is a measure of data, that is either peaked or flat, with respect to the probability distribution.

4- Skewness shows how much and in which direction, the values deviate from the mean? In contrast, kurtosis explain how tall and sharp the central peak is.