<u>Variance:</u>

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

x̄ = 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,

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

Now putting the values in the formula we get,

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

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

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	$(\text{xi-mean})^2$	f.(xi-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
- x̄ = 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-mean)^2	fi.(xi-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***

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

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

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;

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

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

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