VARIABILITY

Santosh Chhatkuli

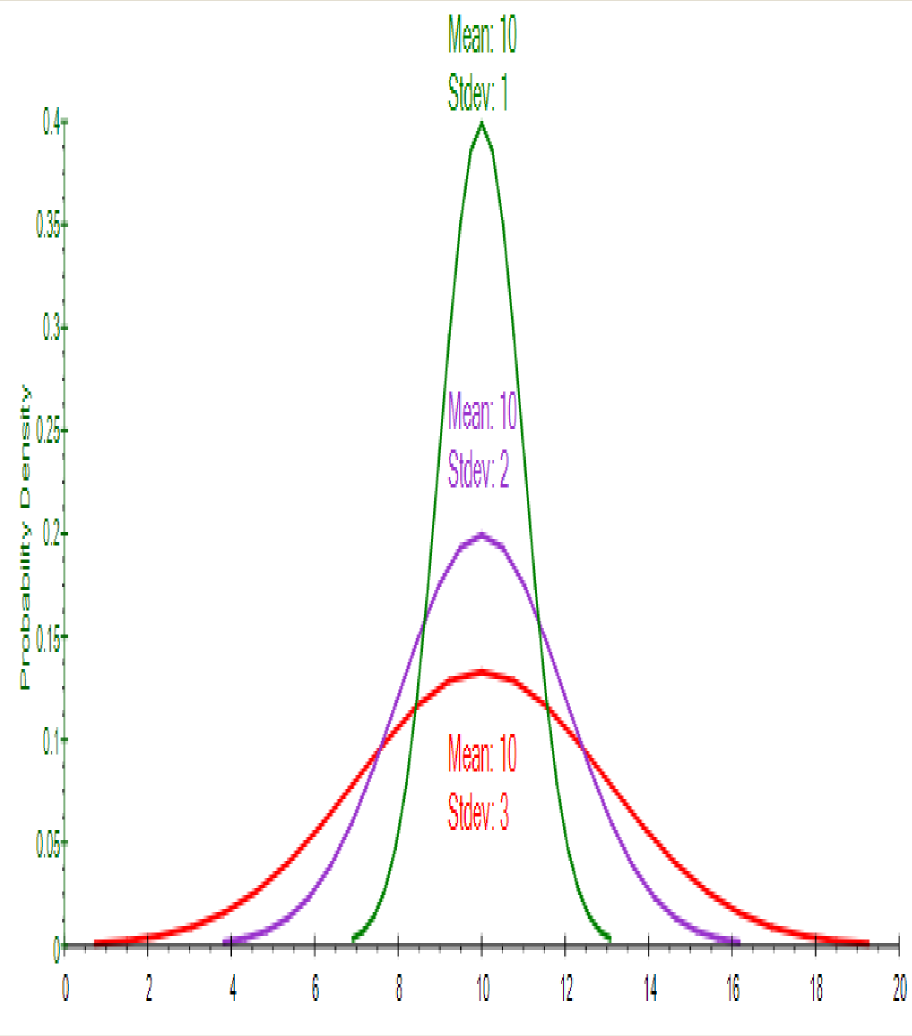
What is variation?

The degree to which numerical data tend to spread about an average value is called variation. It is the second important property that describes a set of numerical data.

The study of average alone is inadequate to give us a complete idea about distribution.

Two or more distributions can be following:

- Same in terms of central value and spread
- Same in terms of central value but different in terms of variation
- Different in terms of central vale but same in terms of variation
- Different in terms of central vale and variation



Consider three hypothetical data sets.

Data set I:	**7**	**8**	**9**	**10**	**11**
Data set II:	**3**	**6**	**9**	**12**	**15**
Data set III:	**1**	**5**	**9**	**13**	**17**

The three distribution are all same in terms of average since they have identical mean i.e. 9 but are different in terms of variation. Data set 3 is most variable since it has high range (16) of values and data set 1 is most consistent because it has small range (4) of values.

When assessing the variability of a data set, there are two key components:

1. How spread out are the data values near the center?
2. How spread out are the tails?

Measures of variation or spread may be either absolute of relative.

Absolute measures of variation

Absolute measures of variation refer to an absolute magnitude of the average of deviations in individual values and are expressed in the same units in which the variable are measured.

1. Range
2. Inter-fractile range/Inter-percentile range
3. Mean Deviation
4. Standard Deviation
5. Variance

Relative measures of variation

A **measure of relative** variation is the ratio of a measure of absolute variation to an appropriate average. In comparing the variability of two or more distributions, relative measures are useful.

1. Coefficient of range
2. Coefficient of inter-fractile range
3. Coefficient of Mean Deviation
4. Coefficient of variation

Range

Range is the difference between the highest and lowest observed value.

Range (R) = $x_{max} - x_{min}$

Coefficient of range = $\frac{x_{max} - x_{min}}{x_{max} + x_{min}}$

Example:

The scores of individual students in the examination and coursework component of a module.

Student	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Coursework	27	44	39	23	41	48	37	34	40	43	30	43	29	27
Examination	12	47	26	25	38	45	35	35	41	39	32	25	18	30

Range of score for coursework is 48 – 23 = 25 and that for examination is 47 – 12 = 35. This indicates that there was greater variation in the students' performance in the examination than in the coursework for this module.

Coefficient of range for coursework = 48 – 23 / 48 + 23 = 0.3521 = 35.21 %

Coefficient of range for Examination = 47 – 12 / 47 + 12 = 0.5932 = 59.32 %

Properties

- It is simple but crude measure of variation
- It is used to get idea of spread of whole distribution
- It gives a quick sense of its spread
- It is based upon two observations only (highest and lowest). So it is not reliable measure.
- It is heavily influence by extreme values.
- It ignores the nature of variation among all the other observations. The spread near the center of the data is not captured at all.

Uses:

1. Analysis of stock data; the difference between highest and lowest value of stock.
2. Analysis of temperature difference per day.

Fractiles or Partitioning values

A **fractile** is the cut off point for a certain fraction of a sample. A fractile is also known as Quantile

Fractiles (Quantiles) are:

1. Quartiles
2. Deciles
3. Percentiles

Quartiles:

Quartiles are partitioning values which divide the data in four equal parts and in three points. The three partitioning points are:

$\mathbf{Q_1}$ = First Quartile or Lower Quartile.
$\mathbf{Q_2}$ = Second Quartile and equals to median.
$\mathbf{Q_3}$ = Third Quartile or Upper Quartile.

Deciles:

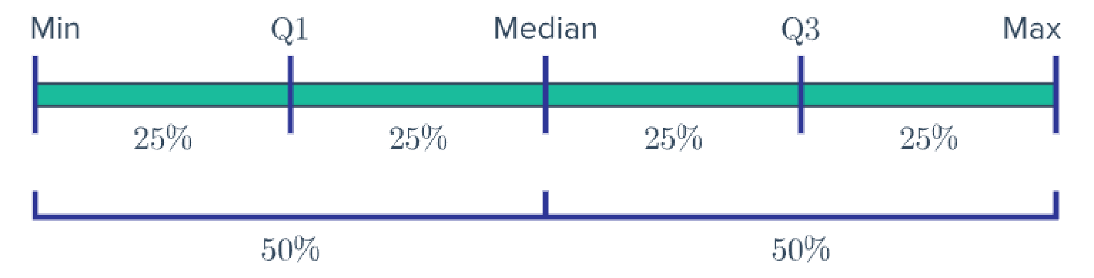
Deciles are partitioning values which divide the whole distribution in ten equal parts and in nine points. There are nine partitioning points and are symbolized by $\mathbf{D_1, D_2 ... D_9}$

Percentiles are those partitioning values which divide the data set into hundred equal parts and in Obviously, points. Obviously, there are ninety-nine partitioning points are symbolized by $\mathbf{P_1, P_2, ..., P_{99}}$.

Inter-fractile Range

Inter-Quartile range or Middle 50 % rang or Quartile Deviation = $Q_3 - Q_1$

It considers the spread in the middle 50 % of the data and therefore it is not influenced by extreme values. A large IQR indicates a large amount of variability among the middle 50 % of the observations and small IQR indicates that the main bulk of the data is fairly concentrated



Semi-interquartile range = $\frac{Q_3 - Q_1}{2}$

Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

Higher the coefficient of quartile deviation, more variable the data set is.

The inter-fractile range between 5th and 95th percentile = Middle 90 % range = $P_{95} - P_5$

It considers middle 90 % variability.

3. Middle 80 % range = $P_{90} - P_{10} = D_9 - D_1$

It give middle 80 % variability in the data set

4. Middle 60 % range = $P_{80} - P_{20} = D_8 - D_2$

5. Middle 50 % range = $P_{75} - P_{25}$

5. Middle 40 % range = $P_{70} - P_{30} = D_7 - D_2$

Computing fractiles

Individual data/Raw data/Ungrouped data

Example: The following data represent the length of life in years, measured to the nearest tenth, of 30 similar fuel pumps:

2.0	3.0	0.3	3.3	1.3	0.4
0.2	6.0	5.5	6.5	0.2	2.3
1.5	4.0	5.9	1.8	4.7	0.7
4.5	0.3	1.5	0.5	2.5	5.0
1.0	6.0	5.6	6.0	1.2	0.2

(a) Find first, second and third quartile

(b) Find interquartile range (middle 50 % range)

(c) Find coefficient of quartile deviation

(d) Find D_4 and P_{80}

(e) Construct Box-and-whisker plot and comment on the shape of distribution

Solution:

Here the variable (X) = Life of fuel pumps (in years)

Sample size (n) = 30

Data Array

0.2	0.2	0.2	0.3	0.3	0.4
0.5	**0.7**	1.0	1.2	1.3	**1.5**
1.5	1.8	**2.0**	**2.3**	2.5	3.0
3.3	4.0	4.5	4.7	**5.0**	5.5
5.6	5.9	6.0	6.0	6.0	6.5

Q_1 = The value of $\frac{1(n+1)}{4}$ th ordered data

= The value of 7.75th ordered data

= The value of 8th ordered data

= 0.7 years

Q_2 = The value of $\frac{2(n+1)}{4}$ th ordered data

= The value of 15.5th ordered data

= Average value of 15th and 16th ranked data

= $\frac{2+2.3}{2}$ = 2.15 years

Q_3 = The value of $\frac{3(n+1)}{4}$ th ordered data

= The value of 23.25th ordered data

= The value of 23rd ordered data

= 5.0 years

IQR = Middle 50 % range

= $Q_3 - Q_1$ = 5.0 – 0.7 = 4.3 years

Semi inter-quartile range = $\frac{IQR}{2}$ = 2.15 years

Coefficient of Quartile Deviation

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{5.0 - 0.7}{5.0 + 0.7} = \frac{4.3}{5.7} = 75.44\ \%$$

D_4 = The value of $\frac{4(n+1)}{10}$ th ordered data

= The value of 12.4th ordered data

= The value of 12th ordered data

= 1.5 yers

P_{80} = The value of $\frac{80\,(n+1)}{100}$ th ordered data

= The value of 24.8th ordered data

= The value of 25th ordered data

= 5.6 years

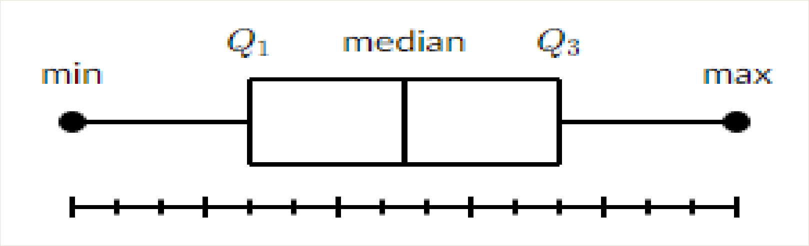
Box-and-whisker plot

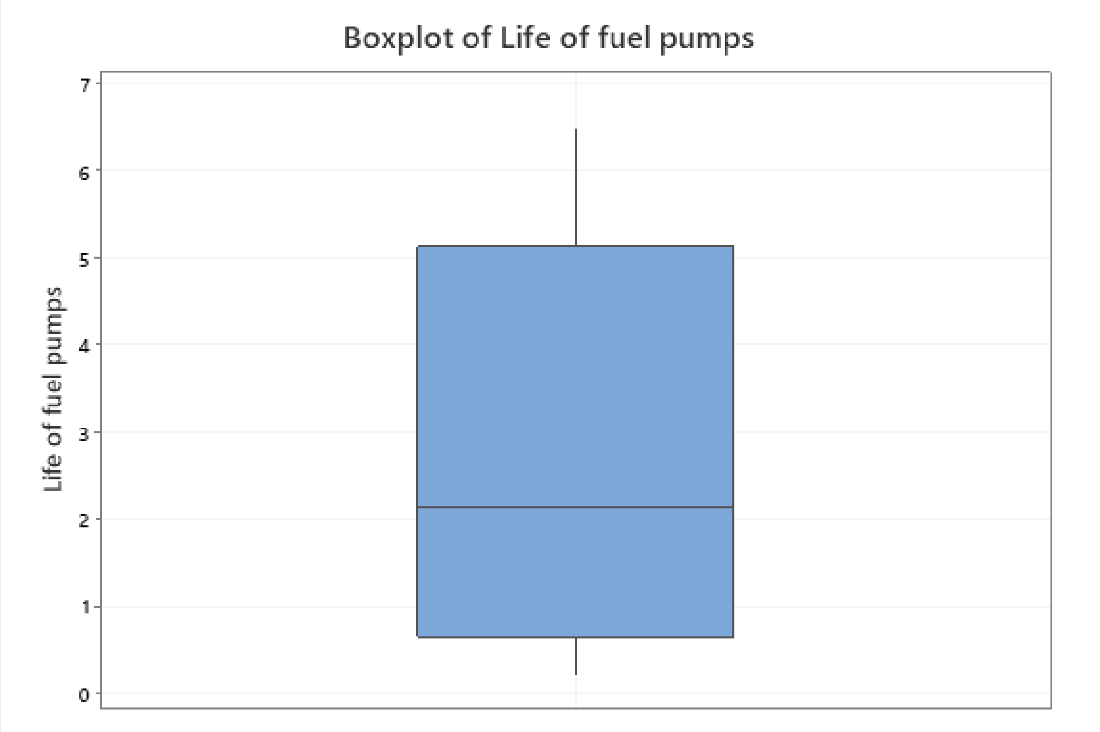
A box and whisker plot also called a box plot, displays the five-number summary of a set of data. This tool is used to detect symmetry or non-symmetry of a data set. It also tells variability in the middle 50 % of the data and in the tails (right or left)

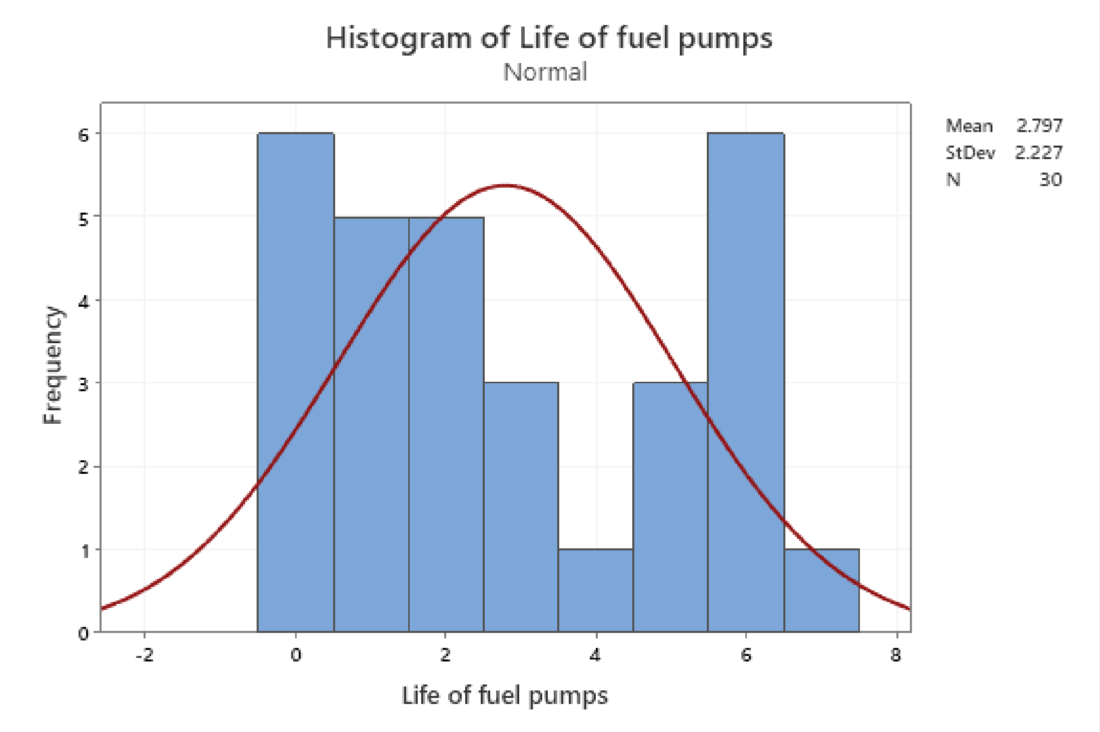
The five number summary is:

1. X_{min} = Lowest observation in the data set
2. Q_1 = First Quartile
3. Q_2 = Second Quartile
4. Q_3 = Third Quartile
5. X_{max} = Highest observation in the data set

In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.




Boxplot of Life of fuel pumps
Life of fuel pumps
0
1
2
3
4
5
6
7


Histogram of Life of fuel pumps
Normal
Frequency
Life of fuel pumps
Mean 2.797
StDev 2.227
N 30
-2
0
2
4
6
8
0
1
2
3
4
5
6

Computing partitioning values in frequency distribution

Example:

The grades of students in Statistics module is given in following frequency distribution

Grades	No. of students	Cumm. Frequency	
52 - 58	2	2	
58 - 64	12	14	
64 - 70	10	24	
70 - 76	19	43	
76 - 82	16	59	
82 - 88	9	68	
88 - 94	7	75	
94 - 100	5	80	
Total	80		

(a) Compute Q1, Q2 and Q3

(b) Compute middle 50 % range

(c) Compute middle 80 % range

(d) Draw a box-and-whisker plot and comment on the nature of the distribution

Solution:

Here the variable of interest (X) = Grades of students in Statistics

Size of the sample (n) = 80

Computing Quartiles

First Quartile (Q1)

Q_1 = Size of $\frac{1x\,n}{4}$ th ordered data

= 20^{th} ordered data

Hence, Q_1 falls in the class 64 – 70.

The actual Q_1 is given by,

$$Q_1 = L_{Q_1} + \frac{\frac{n}{4} - c.f.}{f} * h = 64 + \frac{20 - 14}{10} * 6$$

= 67.6

Second Quartile (Q2)

Q_2 = Value of $\frac{2xn}{4}$ th ordered data

= 40th ordered data

Hence, Q_2 falls in the class 70 -76

Then actual Q_2 is given by,

Q_2 = $L_{Q_2} + \frac{\frac{2n}{4} - c.f.}{f} * h = 70 + \frac{40 - 24}{19} * 6$

= 75.05

Third Quartile Q_3

Q_3 = Value of $\frac{3\,x\,n}{4}$ th ordered data

= 60th ordered data

Hence, the third quartile lies in the class 82 – 88

The actual Q_3 is given by the formula

Q_3 = $L_{Q_3} + \frac{\frac{3n}{4} - c.f.}{f} * h = 82 + \frac{60 - 59}{9} * 6$

= 82.67

The inter-quartile range (middle 50 % range)

= Q3 – Q1 = 82.67 – 67.6 = 15.07

The range of variability of marks of middle 50 % of students is mere 15.07

Tenth Percentile P_1

P_{10} = Value of $\frac{10\ x\ n}{100}$ th ordered data

= 8^{th} ordered data

Hence, tenth percentile P10 lies in the class 58 – 64

The actual P_{10} class is given by

$$P_{10} = L_{P_{10}} + \frac{\frac{10n}{100} - c.f.}{f} * h = 58 + \frac{8-2}{12} * 6 = 61$$

Ninetieth Percentile P_{90}

P90 = Value of $\frac{90\ x\ n}{100}$ th ordered data

= 72^{nd} ordered data

Hence, the actual P90 class falls in the class 88 – 94

Now, the actual P90 is calculated as follows

$$P_{90} = L_{P_{90}} + \frac{\frac{90n}{100} - c.f.}{f} * h = 88 + \frac{72-68}{7} * 6 = 91.43$$

The middle 80 % range = $P_{90} – P_{10}$ = 91.43 – 61 = 30.43

Hence, the variability of marks for the middle 80 % (64) of students is 30.43.

Box-and-Whisker plot

The five number summary is given by

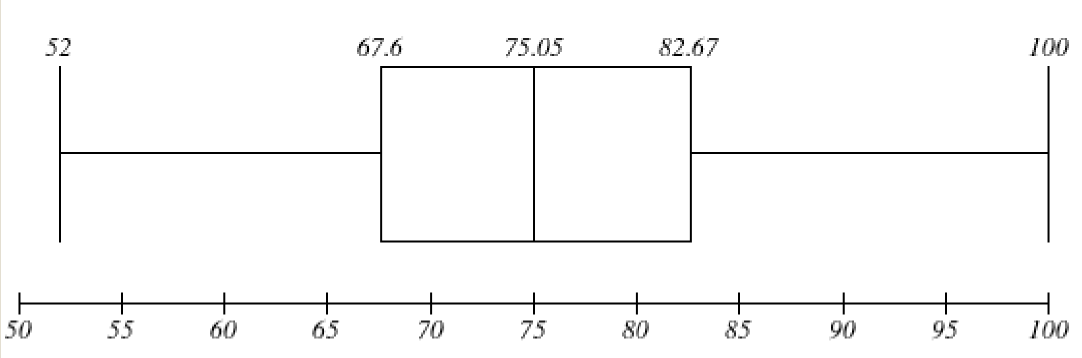
X_{min} = 52

Q_1 = 67.6

$Q_2 = M_d$ = 75.05

Q_3 = 82.67

X_{max} = 100



The length of stay of hospital patients is given in the following table.

LOS (days) (X)	No. of patients (f)
1	2
2	6
3	6
4	5
5	11
6	6
7	8
8	5
9	3
10	1
11	2
12	3
Total	n $=\sum f = 58$

(a) Compute Q1, Q2 and Q3

(b) Compute inter-quartile range

(c) Compute semi inter quartile range

(d) Compute coefficient of quartile deviation

(e) Construct a box-and-whisker plot

(f) Comment on the shape of the distribution of length of stay of hospital patients

Standard Deviation

It is a statistic that measures the dispersion or variation of a dataset relative to its mean. It is an absolute measure of dispersion that expresses variation in the same units as the original data.

It is defined as the positive square root of the average of the square of the deviations of the measurements from their arithmetic mean.

Notation:

Sample standard deviation = s

Population standard deviation = σ

Sample S.D.

Raw Data

For a sample of n observations x1, x2, …, xn , the standard deviation is given by,

$$s = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}} = \sqrt{\frac{(X_1-\bar{X})^2+\ldots+(X_2-\bar{X})^2}{n-1}}$$

A computationally easier formula is given by,

$$s = \sqrt{\frac{1}{n-1}\left\{\sum X^2 - n.\bar{X}^2\right\}}$$

Note: The divisor n -1 is used instead of n because the sample standard deviation 'S' will be much closer to population standard deviation 'σ' when we use divisor n – 1.

Ungrouped frequency distribution

For ungrouped frequency distribution we use following formula

Definitional Formula

$$s = \sqrt{\frac{\sum f(X-\bar{X})^2}{n-1}}$$

Computational Formula

$$s = \sqrt{\frac{1}{n-1}\left\{\sum f X^2 - n\bar{X}^2\right\}}$$

Grouped frequency distribution

For grouped frequency distribution the formula are:

Definitional or Direct Formula

$$s = \sqrt{\frac{\sum f(m-\bar{X})^2}{n-1}}$$

Computational Formula

$$s = \sqrt{\frac{1}{n-1}\left\{\sum f m^2 - n\bar{X}^2\right\}}$$

Example 1: The random sample of 29 calls made by office staff of a company during a week is given below:

6.8	2.3	4.8	8.3	15.9	18.7
11.8	5.6	15.9	10.4	15.3	12.3
9.1	10.4	7.2	14.5	11.2	15.3
19.8	7.6	17.7	11.1	9.0	13.2
12.0	3.7	8.0	13.4	12.5	

Compute standard deviation of duration of call.

Solution:

Variable (X) = Duration of calls by staff

Sample size (n) = 29

The sample standard deviation is given by,

$$s = \sqrt{\frac{1}{n-1}\left\{\sum X^2 - n\,.\bar{X}^2\right\}}$$

Here,

$\Sigma X = 6.8 + 2.3 + \ldots + 12.5 = 323.8$

$\Sigma X^2 = 6.8^2 + 2.3^2 + \ldots + 12.5^2 = 4174.54$

Now,

Mean $\bar{X} = \frac{\sum X}{n} = \frac{323.8}{29} = 11.1655$ mins

$$\text{S.D.(s)} = \sqrt{\frac{1}{n-1}\left\{\sum X^2 - n\,\bar{X}^2\right\}} = \sqrt{\frac{1}{29-1}\{4174.54 - 29 \text{ x } 124.6684\}} = 4.3910 \quad \text{mins}$$

Example 2: A professor in Statistics class was rated by his 30 students on a scale of 1 to 5 where rating of 5 was considered as excellent and the rating of 1 was considered as poor with increasing scale from poor to excellence. These ratings are recorded individually in the following table.

4	3	1	1	4
5	2	2	4	4
2	3	3	4	3
5	4	4	5	3
3	2	2	3	4
4	5	4	5	1

(a) Construct an ungrouped frequency distribution

(b) Compute standard deviation of rating obtained by Professor

Solution:

Here variable (X) = Rating score given by students

Sample size (n) = 30

The frequency distribution for this data is

Rate	**1**	**2**	**3**	**4**	**5**
Frequency	3	5	7	10	5

The sample S.D. is given by,

$$s = \sqrt{\frac{1}{n-1}\left\{\sum f\,X^2 - n\,\bar{X}^2\right\}}$$

Table

Rate (X)	1	2	3	4	5	Total
Frequency (f)	3	5	7	10	5	30
f . X	3	10	21	40	25	99
f . X2	3	20	63	160	125	371

Sample mean is given by

$$\bar{X} = \frac{\sum f X}{n} = \frac{99}{30} = 3.3$$

Sample SD is given by

$$s = \sqrt{\frac{1}{n-1}\left\{\sum fX^2 - n.\bar{X}^2\right\}} = \sqrt{\frac{1}{30-1}\{371 - 30 * 3.3^2\}} = 1.24$$

Computing S.D. in grouped frequency distribution

Example 3: The CEO of an IT company wants to study the pattern of absenteeism of employees over a given year. The data from the files of a sample of 50 employees shows the following distribution of number days of these employees were absent.

No. of days absent	0 to 2	3 to 5	6 to 8	9 to 11	12 to 14	Total
No. of employees	15	20	8	5	2	50

Compute standard deviation of the distribution

Solution:

The variable (X) = No. of days employee of company remain absent

Frequency (f) = No. of students

First we calculate mean of the distribution and then S.D.

Computation table

X	0 to 2	3 to 5	6 to 8	9 to 11	12 to 14	Total
f	15	20	8	5	2	**50**
m	1	4	7	10	13	**35**
fm	15	80	56	50	26	**227**
fm2	15	320	392	500	338	**1565**

The mean of the distribution is

$$\bar{X} = \frac{\sum f\,m}{n} = \frac{227}{50} = 4.54$$

Now the SD of this distribution is

$$s = \sqrt{\frac{1}{n-1}\left\{\sum fm^2 - n\bar{X}^2\right\}} = \sqrt{\frac{1}{50-1}\{1565 - 50 * 4.54^2\}} = 3.3$$

Hence, the distribution of number of days of absent has mean of 4.54 days and standard deviation of 3.3 days.

Properties of S.D.

1. The standard deviation (or variance) is better measure of variation in the data because the measures take into account of every observation in the data set.
2. It is best and powerful measure of variation.
3. It is least affected by sampling fluctuations.
4. It gives greater weight to the extreme values and less to those which are near the mean. It is unduly affected by extreme observations.

Uses of standard deviation

1. **Spread of the data**: The standard deviation is useful in describing departure of individual items in a distribution from the mean i.e. the spread of data. If standard deviation or variance of the data set is large, then data are more dispersed from the mean and in this case the mean becomes less representative.
2. **Comparing consistency of data sets**: To check the variability or consistency of two or more samples, we compare their sample standard deviation values. Comparison is possible or sensible when the samples are related or measurements are done on the same variable on different occasions or different group.
3. **Comparing standard scores or Z scores**: The sample standard scores tell us how many standard deviations a particular sample observation lies below or above the sample mean. The sample standard score is computed by formula:

$$Z = \frac{Data - Mean}{SD} = \frac{X - \bar{X}}{S}$$

It measures the deviation of X from the mean in terms of standard deviations. The standardized variable Z is often used in educational testing, where it is known as a standard score.

4. To know the percentage value in the range: The standard deviation helps tell us how a set of data clusters or distributes around its mean. In other words, S.D. used to determine the number of data value that fall within a specified interval in a distribution.

Empirical rule for normally distributed data:

For normal distribution, the following rule of clustering of data obeys:

(a) The interval includes **68.26 %** of observations i.e. about 68 % of the values fall within ± one standard deviation from the mean.

(b) The interval includes **95.44** % of observations.

(c) The interval includes **99.74** % of observations.

Pooled SD or Combined SD

The combined SD of two related samples is given by,

$$S_c = \sqrt{\frac{Combined\ sum\ of\ squares\ of\ deviations}{total\ sample\ size - 1}}$$

$$= \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2+n_1(\bar{X}_1-\bar{X}_c)^2+n_2(\bar{X}_2-\bar{X}_c)^2}{n_1+n_2-1}}$$

If we have k samples, the generalized formula is,

$$S_c = \sqrt{\frac{\sum_{i=1}^{k}(n_i-1)s_i^2 + \sum_{i=1}^{n} n_i(\bar{X}_i-\bar{X}_c)^2}{n_1+n_2+\dots+n_k-1}}$$

If samples are of equal sizes then,

$$S_c = \sqrt{\frac{(n-1)\sum_{i=1}^{k} s_i^2 + n\sum_{i=1}^{n}(\bar{X}_i-\bar{X}_c)^2}{nk-1}}$$

Symbol:

k = No. of independent samples

n_1, n_2, … etc. are no. of measurements made for different samples (sizes of different samples)

$S_1, S_2, \dots$. etc. are within sample standard deviations or S.D. of respective samples.

Example: Consider a large retail chain with stores in two different regions: Region A and Region B. The company wants to understand the overall customer satisfaction across both regions. Full score for the satisfaction is 100

Region	Sample size	Sample mean	Sample sd
A	100	85	10
B	150	82	12

(a) Calculate overall mean satisfaction score

(b) Calculate the overall standard deviation of score or combined standard deviation

Coefficient of Variation

Coefficient of Variation is the relative measure of variation and it measures the scatterness in the data relative to the mean. It is the percentage variation in the mean, standard deviation being considered as the total variation in the mean. It relates $\bar{X}$ and s by expressing standard deviation as percentage of the mean.

Sample coefficient of variation is given by,

$$C.V. = \frac{s}{\bar{X}} \text{ x } 100\ \%$$

Notation:

$\bar{X}$ = Sample mean

S = Sample standard deviation

The population coefficient of variation is given by,

$$C.V. = \frac{\sigma}{\mu} \text{ x } 100\ \%$$

Notation:

μ = Population mean

σ = Population standard deviation

Higher coefficient of variation means high variability or more spread of the data from their mean.

High C.V. ---- Less consistency of data.

Low C.V. ---- More consistency of data.

Note:

- C.V. is particularly useful when we have to compare the variability of two or more data sets that are expressed in different units of measurement.
- C.V. can also be useful when comparing two or more sets of data that are measured in the same units but differ to such an extent that a direct comparison of the respective standard deviations is not very helpful.
- C.V. is used to compare variability of same character/variable in two different groups like comparing variability of weight among male and female students
- C.V. is can also be used to compare variability of two different character/variable in same group like comparing variability of weight and height among a group of students.

Example

The following are the weights (Kg) and heights (cm) of the 14 patients.

Weight

83.9	99.0	63.8	71.3	65.3	79.6
70.3	69.2	56.4	66.2	88.7	59.7
64.6	78.8				

Height

185	180	173	168	175	183
184	174	164	169	205	161
177	174				

(a)For each variable, compute the mean, standard deviation, and coefficient of variation.

(b)Which set of measurements is more variable, weight or height? On what do you base your answer?

Solution

Let variable (X1) = Weight of patients (kg)

and variable (X2) = Height of patients (cm)

Weight

$n_1 = 14$

$\Sigma X_1 = 83.9 + \ldots + 78.8 = 1016.8$

$\sum X_1^2 = 83.9\text{^}2 + \ldots + 78.8\text{^}2 = 75703.1$

Now,

$$\bar{X}_1 = \frac{\sum X_1}{n_1} = \frac{1016.8}{14} = 72.63$$

$$s_1 = \sqrt{\frac{1}{n_1 - 1}\{\sum X_1^2 - \ n_2\,\bar{X}_2^2\ \}}$$

$$= \sqrt{\frac{1}{14.-1}\{75703.1 - 14\ *\ 72.63^2\}}$$

$= 11.94$

$$CV_1 = \frac{S_1}{\bar{X}_1}\ * 100\ \%$$

$$= \frac{11.94}{72.63}\ * 100\ \%$$

$= 16.44\ \%$

Height

$$n_1 = 14$$

$$\sum X_2 = 185 + 180 + \ldots + 174 = 2472$$

$$\sum X_2^2 = 185^2 + 180^2 + \ldots + 174^2 = 438032$$

Now,

$$\bar{X}_1 = \frac{\sum X_2}{n_2} = \frac{2472}{14} = 176.57$$

$$s_2 = \sqrt{\frac{1}{n_2 - 1}\{\sum X_2^2 - n_2\,\bar{X}_2^2\}}$$

$$= \sqrt{\frac{1}{14.-1}\{438032 - 14 * 176.57^2\}}$$

$$= 10.94$$

$$CV_2 = \frac{S_2}{\bar{X}_2} * 100\ \%$$

$$= \frac{10.944}{176.57} * 100\ \%$$

$$= 6.196\ \%$$

Conclusion: Since coefficient of variation of Weight (16.44 %) is greater than coefficient of variation of height (6.196 %), we can conclude that the weight is more variable characteristic than height among the patients.

Example 4:

Two automatic filling machines A and B are used to fill tea in 500 grams bag. A random sample of 100 bag on each machine showed the following results:

Tea Contents (in gm)	Machine A	Machine B
485 – 490	12	10
490 – 495	18	15
495 – 500	20	24
500 – 505	22	20
505 – 510	24	18
510 – 515	4	13
Total	100	100

Comment on the performance of two machines on the basis of following measures:

(a) Average filling

(b) Variability in filling

(c) Consistency in filling

Solution

Let X_1 = Weight of tea in bags filled by machine A

X_2 = Weight of tea in bags filled by machine B

Weight	f1	f2	m	f1m	f1m^2	f2m	f2m^2	
485 - 490	12	10	487.5	5850	2851875	4875	2376562.5	
490 - 495	18	15	492.5	8865	4366013	7387.5	3638343.75	
495 - 500	20	24	497.5	9950	4950125	11940	5940150	
500 - 505	22	20	502.5	11055	5555138	10050	5050125	
505 - 510	24	18	507.5	12180	6181350	9135	4636012.5	
510 - 515	4	13	512.5	2050	1050625	6662.5	3414531.25	
Total	100	100		49950	24955125	50050	25055725	
		Machine A	Machine B					
	Mean	499.5	500.5					
	SD	7.177405626	7.5878691					
	CV	1.436918043	1.5160578					

Machine A

$$\bar{X}_1 = \frac{\sum f_1 m}{n_1} = \frac{499.50}{100} = 499.50$$

$$s_1 = \sqrt{\frac{1}{n_1 - 1}\left\{\sum f_1 m^2 - n_1\,\bar{X}_1^2\right\}}$$
$$= \sqrt{\frac{1}{100 - 1}\{24955125 - 100 * 499.5^2\}}$$
$$= 7.18$$

$$CV_1 = \frac{s_1}{\bar{X}_1} * 100\%$$
$$= \frac{7.18}{499.50} * 100\%$$
$$= 1.44\ \%$$

Machine B

$$\bar{X}_2 = \frac{\sum f_2 m}{n_2} = \frac{500.50}{100} = 500.5$$

$$s_2 = \sqrt{\frac{1}{n_2 - 1}\left\{\sum f_2 m^2 - n_2\,\bar{X}_2^2\right\}}$$
$$= \sqrt{\frac{1}{100 - 1}\{25055725 - 100 * 500.5^2\}}$$
$$= 7.59$$

$$CV_2 = \frac{s_2}{\bar{X}_2} * 100\%$$
$$= \frac{7.59}{500.5} * 100\%$$
$$= 1.52\ \%$$

Conclusion:

(a) Machine B has slightly higher mean (500.5) than machine A (499.5)

(b) Machine A has less variability (7.18) in filling the teabag than machine B (7.59)

(c) Machine A is more consistent (1.44 %) in filling the teabag than machine B (1.52%)

(d) Though there is no significant difference between two machine in terms of average, standard deviation and consistency in filling the teabags with tea, we can say machine A is more closer to the confirmation to the quality than machine B.

Example 3: In a certain test for which the pass marks is 30, the distribution of marks of passing candidates classified by sex (boys and girls) were as given below.

Marks	No. of boys	No. of girls
30 – 34	5	15
35 – 39	10	25
40 – 44	15	30
45 – 49	30	14
50 – 54	5	5
55 – 59	5	1
Total	70	90

(a) Find the mean and standard deviation of marks obtained by the boys and girls in the test.

(b) Compute coefficient of variation of marks obtained by boys and girls in the test and comment on the consistency of performance for two groups.

Example 1:

The scores of individual students in the examination and coursework component of a module are given below:

Student	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Coursework	27	44	39	23	41	48	37	34	40	43	30	43	29	27
Examination	12	47	26	25	38	45	35	35	41	39	32	25	18	30

Compute the coefficient of variation for both component and conclude which component is more variable for a module?

Example 2: 60 students were asked how many books they had read over the past 12 months. The results are listed in the frequency distribution table below.

No. of books	0	1	2	3	4	5	6	7	8
No. of students	1	6	8	10	13	8	5	6	3

(a) Compute mean and standard deviation

(b) Compute coefficient of variation.

Mean Deviation

The mean absolute deviation (MAD) is calculated by summing the absolute deviations around the arithmetic mean (mostly of the time) and dividing by the number of observations.

The mean deviation or mean absolute deviation from an average A is given by,

$$\text{M.D.} = \frac{\sum|X - A|}{n}$$

Where,

X = Data

A = Average (arithmetic mean or median or mode)

The coefficient of mean deviation from an average A is given by,

$$\text{Coefficient of M.D.} = \frac{Mean\ Deviaiton\ from\ an\ average\ A}{A}$$

It doesn't give satisfactory result when deviations are taken from mode as mode is ill-defined.