3.2 Measures of Dispersion

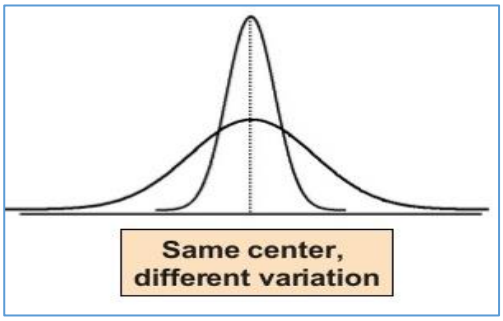
Objectives

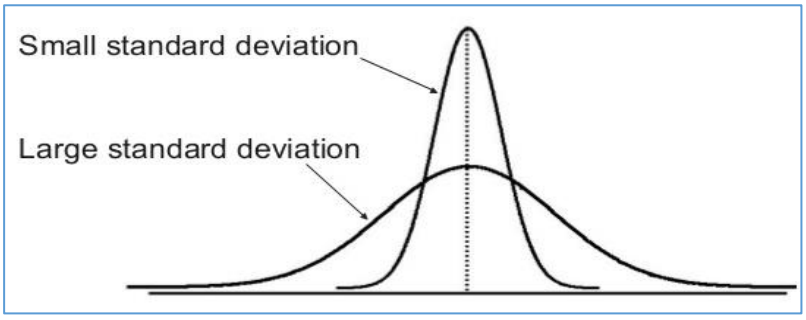
In this section, we discuss the characteristic of variation. In particular, we present measures of variation, such as the standard deviation, as tools for analyzing data. Our focus here is not only to find values of the measures of variation, but also to interpret those values. In addition, we discuss concepts that help us to better understand the standard deviation.

Measure of Variation (Dispersion):

The variation or dispersion in a set of values refers to how spread out the values are from each other. Some measures of dispersion are: Range, Variance, Standard deviation and Coefficient of variation...etc.

- The variation is small when the values are close together.
- There is no variation if the values are the same.





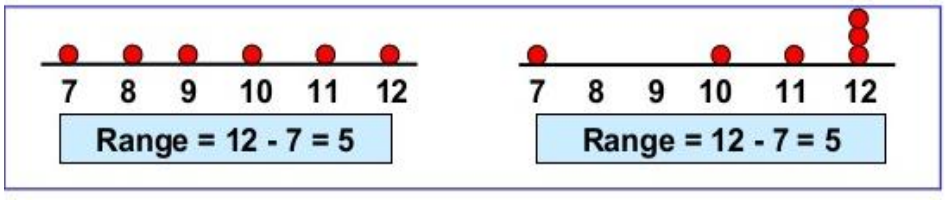
Range and Coefficient of Range:

The range is the simplest measure of variability, calculated by taking the difference between largest observation in the data set (L) and smallest observation in the data set (s) as follow:

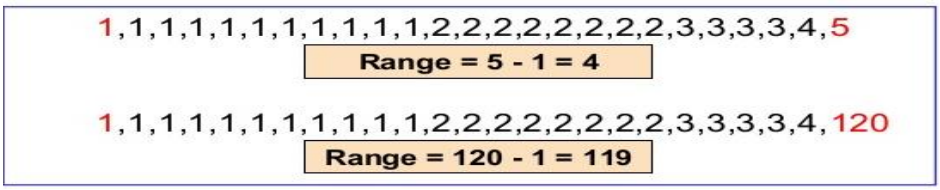
Range = L – s ; $R_{coefficient} = \frac{L-s}{L+s}$

Example (3.11) Disadvantages of Range:

1. Ignores the way in which data are distributed, for example:



2. Sensitive to outliers, for example:



Range For grouped data, is the difference between the highest class boundary and the lowest boundary.

Example (3.12):

a) Find the range of the data: 2, 2, 2, 1, 3, 3, and 9.
b) Find the range of the grouped data

Class Limit	10 – 14	15 – 19	20 - 24	25 - 29
Frequency	2	8	7	3

Solution:

a) Range = 9-1 = 8
b) Range = 29-10=19

Variance, Coefficient of Variance and Standard Deviation:

The variance is a measure that uses the mean as a point of reference. Variance and the standard deviation are useful as measures of variation of the values of a single variable for a single population (or sample).

- Variance is small when all values are close to the mean.
- Variance is large when all values are spread out from the mean.

If we want to compare the variation of two variables we cannot use the variance or the standard deviation because:

- The variables might have different units.
- The variables might have different means.

Without an understanding of the relative size of the standard deviation compared to the original data, the standard deviation is somewhat meaningless for use with the comparison of data sets. To address this problem the coefficient of variation (CV) is used.

- The coefficient of variation often used to compare the variability of two data sets. It allows comparison regardless of the units of measurement used for each set of data.
- The **larger** the coefficient of variation, the **more** the data varies.

Properties of Standard Deviation

- Measures the variation among data values.
- Values close together have a small standard deviation, but values with much more variation have a larger standard deviation.
- Has the same units of measurement as the original data.
- For many data sets, a value is unusual if it differs from the mean by more than two standard deviations.
- Compare standard deviations of two different data sets only if they use the same scale and units, and they have means that are approximately the same.
- The value of the standard deviation is usually positive. It is zero only when all of the data values are the same number. (It is never negative because it measures a distance).
- The value of the standard deviation can increase dramatically with the inclusion of one or more outliers (data values that are very far away from all of the others).
- The units of the standard deviation (such as minutes, feet, pounds, and so on) are the same as the units of the original data values.

Relations:

For ungrouped data:

	Population	Sample
Mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Variance $(unit)^2$	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ $\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \ldots + (x_N-\mu)^2}{N}$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ $S^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2}{n-1}$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$S = \sqrt{S^2}$
Coefficient of Variation	$CV = \frac{\sigma}{\mu} * 100\ \%$	$CV = \frac{S}{\bar{x}} * 100\ \%$
Where:	σ^2 (sigma-squared) is Population variance x_i is the item or observation. N is the total number of observations in population. μ is Population mean CV is coefficient of variance.	S^2 is Sample variance x_i is the item or observation. n is the total number of observations in sample. $\bar{x}$ is sample mean. CV is coefficient of variance

For grouped data:

	Population	Sample
Variance $(unit)^2$	$\sigma^2 = \frac{\sum(f\,x^2)}{\sum f} - \left(\frac{\sum f\,x}{\sum f}\right)^2$	$S^2 = \frac{1}{\sum f - 1}\left[\sum(f\,x^2) - \frac{(\sum f\,x)^2}{\sum f}\right]$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$S = \sqrt{S^2}$
Where:	f : is the frequencies and x : is the class mark (mid-point).	

*For more than one data set use:

	Mean	St.dev.	C.V.
1st data set	$\bar{x}_1$	S_1	$CV_1 = \frac{S_1}{\bar{x}_1}100\ \%$
2nd data set	$\bar{x}_2$	S_2	$CV_2 = \frac{S_2}{\bar{x}_2}100\%$

Example (3.13): Find the standard deviation for

a) Sample data: 6, 7, 8, 9, 10.

b) Grouped data:

x_i	0	1	2	3	4
f_i	2	1	2	0	2

Solution:

a) $S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$

n=5

No. of data	x_i	$\bar{x}$	$(x-\bar{x})$	$(x-\bar{x})^2$
1	6	8	-2	4
2	7	8	-1	1
3	8	8	0	0
4	9	8	1	1
5	10	8	2	4
∑				10

Variance; $S^2 = \frac{10}{5-1} = 2.5$

Standard deviation; $S = \sqrt{S^2} = 1.58$

b) $S^2 = \frac{1}{\sum f - 1}\left[\sum(f\,x^2) - \frac{(\sum f\,x)^2}{\sum f}\right]$

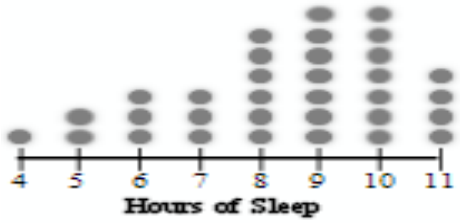
No. of data	xi	fi	fx	x^2	fx^2
1	0	2	0	0	0
2	1	1	1	1	1
3	2	2	4	4	8
4	3	0	0	9	0
5	4	2	8	16	32
∑		7	13		41

Variance; $S^2 = \frac{1}{7-1}\left[41 - \frac{(13)^2}{7}\right] = 2.81$

Standard deviation; $S = \sqrt{S^2} = 1.676$

Example (3.14):

The dot plot below shows the number of hours of sleep per night for 33 students. Find mean, median and standard deviation for the plotted data.



Example (3.15): homework

A practical example of the mean is the determination of the **mean velocity of a stream** based on measurements of travel times over a given reach of the stream using a floating device. For instance, if 10 velocities are recorded as follow, Calculate Mean, Median, Mode, Standard Deviation and Coefficient of Variation?

Velocity (m/s)	0.20	0.20	0.21	0.42	0.24	0.16	0.55	0.70	0.43	0.34

Example (3.16): homework

From 28-day **concrete cube tests** made in England in 1990, the following results of maximum load at failure in kN and compressive strength in N/mm^2 were obtained:

Max. load	950	972	981	895	908	995	646	987	940	937	846	947	827	961	935	956
Comp. strength	42.25	43.25	43.5	39.25	40.25	44.25	28.75	44.25	41.75	41.75	38.00	42.5	36.75	42.75	42.0	33.5

Calculate Mean, Median, Mode, Standard Deviation and Coefficient of Variation?

Estimation of Standard Deviation (Approximation):

Range Rule of Thumb:

- "Usual values" are values that are typical and not too extreme which are:

  Minimum usual value = mean – 2* (Standard Deviations). >>> Min. = $\bar{x} - 2S$

  Maximum usual value = mean + 2 * (Standard Deviations). >>> Max. = $\bar{x} + 2S$

- If we don't know the standard deviation we can approximate it using the range rule of thumb rule:

  $S = \frac{Range}{4}$ »»»» $S = \frac{L-s}{4}$, Where: L is the largest value and s is the smallest value.

Example (3.17):

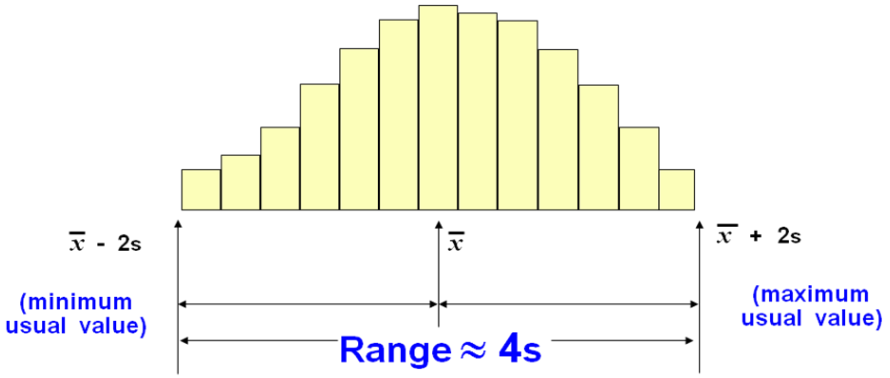
Use the **range rule of thumb** to approximate the standard deviation of the following:

a. 2, 5, 8 approximate standard deviation is (8-2)/4 = 1.5

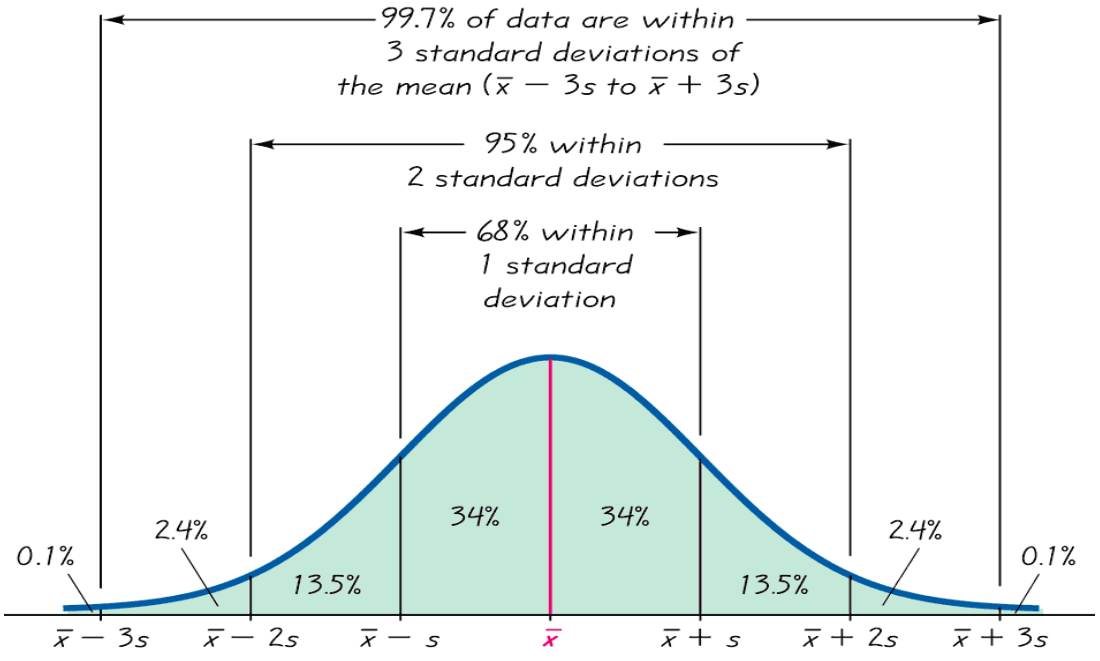
b. 36, 45, 52, 41 approximate standard deviation is (52-36)/4 = 4

Empirical Rule:

If the histogram is bell shaped distribution as shown below:



This means that:



1. Approximately **68%** of all observations fall within **one** standard deviation of the mean.
2. Approximately **95%** of all observations fall within **two** standard deviations of the mean.
3. Approximately **99.7%** of all observations fall within **three** standard deviations of the mean.

Example (3.18):

The scores for "math test" are normally distributed (approximately) with a mean **100** and standard deviation **15**. Use 68-95-99.7% rule to answer the following:

a. About what % of students have math scores **above 100**?
b. About what % of students have math scores **above 145**?
c. About what % of students have math scores **below 85**?

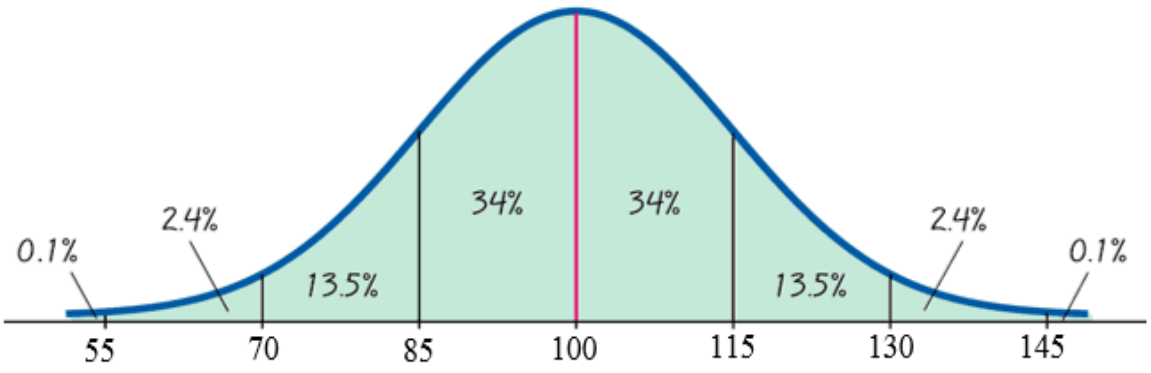
Solution:

a. $\overline{X}$ = 100,

1st Minimum usual value = mean - 1 (standard deviation) = $\overline{X}$ - 1(S) = 100- 1(15) = 85
2nd Minimum usual value = mean - 2 (standard deviation) = $\overline{X}$ - 2(S) = 100- 2(15) = 70
3rd Minimum usual value = mean - 3 (standard deviation) = $\overline{X}$ - 3(S) = 100- 3(15) = 55

And

1st Maximum usual value = mean + 1 (standard deviation) = $\overline{X}$ + 1(S) = 100+1(15) = 115
2nd Maximum usual value = mean + 2 (standard deviation) = $\overline{X}$ + 2(S) = 100+2(15) = 130
3rd Maximum usual value = mean + 3 (standard deviation) = $\overline{X}$ + 3(S) = 100+3(15) = 145



Therefore, as explained in the above figure about (34+13.5+2.4+0.1%) **50%** of students have math scores above 100.

b. About what **0.1%** of students have math scores above 145.
c. About (13.5+2.4+0.1%) **16%** of students have math scores below 85.

Example (3.19): homework

The results of a **water content of a soil** sample yields a **bell shaped data** set with a mean of **26%** and a standard deviation of **5%**, Answer for the followings:

A) What is the range for usual data?
B) What is the range for unusual data?
C) Would a water content of 10 and 60 are considered as a usual value?

Chebyshev's Theorem:

The proportion of any distribution that lies within (K) standard deviations of the mean is at least:

$$1 - \frac{1}{K^2}$$

Where; K is any positive number greater than one.

This theorem applies to all **distributions of data (any shape).** For example; this theorem says that within two standard deviations of the mean, you will find at least

$1 - \frac{1}{2^2} = 1 - \frac{1}{2^2} = \frac{3}{4} = 0.75$ or at least 75%

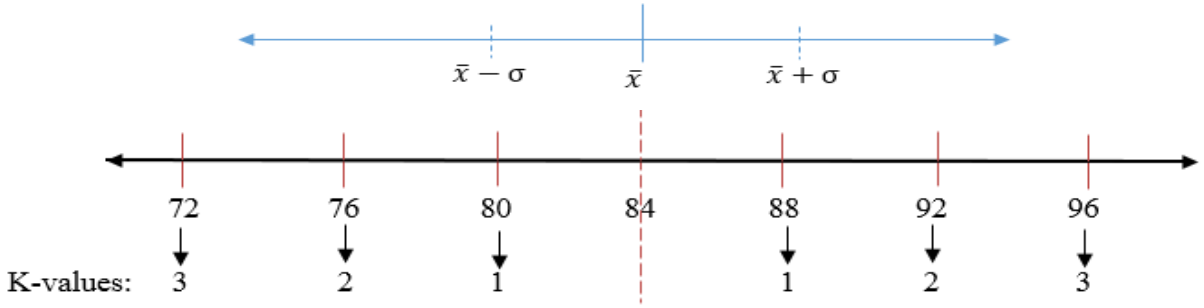
We can say that:

- K=2, at least 3/4 (75%) of all values lie within 2 standard deviations of the mean.
- K=3, at least 8/9 (89%) of all values lie within 3 standard deviations of the mean.

Example 3.20:

Suppose that the average score on a math test is an **84** with a standard deviation of **4** points. According to Chebyshev's theorem, at least what percent (%) of the tests have a grade of at least 72 and at most 96?

Solution:



K=3,

$1 - \frac{1}{K^2} = 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9}$

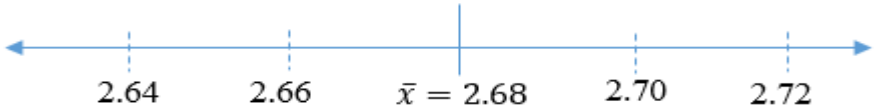
Its 0.89 or 89%

Example 3.21:

The mean value of **specific gravity results** of 20 soil samples was **2.68** with a standard deviation of **0.02**. Find the range at which at least 75% of the data will fall using Chebyshev's Theorem.

Solution:

$1 - \frac{1}{K^2} = 75\%$ >>>> $1 - \frac{1}{K^2} = 0.75$ >>>> K=2



Hence, Range of 75% specific gravity will fall between (2.64-2.72)

Example 3.22: homework

If the **annual rainfalls** in a city are 22, 37, 25, 62, 33, 51, 56, 42, 53, and 49 cm over a 10-year period, find the minimum percentage of the data values that will fall between 36 and 50cm.