Department of CSE
(School of Technology)
MATH2361: Probability and Statistics (No. of hrs/week: 3 Credits: 3)
@Semester –IV

UNIT-I
Measures of Dispersion

Dr Malikarjuna Reddy Doodipala
MSc, M.Phil, PGDCA, Ph D,
Associate Professor
Department of Mathematics

Learning Objectives

By the end of this topic, students should be able to:

- Learn Measures of dispersion
- Understand how to find dispersion or variation of the data
- Know the Measures of coefficient of dispersion
- Understand how to find coefficient dispersion or variation of the data
- Get an idea on Partition values
- Analyze statistical data by measures of dispersion using-MS-Excel

Learning Outcomes

Upon successful completion of this topic, students will be able to:

- Learn Measures of dispersion
- Understand how to find dispersion or variation of the data
- Learn Measures of coefficient of dispersion
- Understand how to find coefficient dispersion or variation of the data
- Partition values
- Analyze statistical data by measures of dispersion using-MS-Excel

Prerequisite: Data and Data Sets

- Data & data set
- Central tendencies

What is Dispersion in QT?

- Dispersion is the state of getting scattered or spread.
- Statistical dispersion means the extent to which a numerical data(variable) is likely to vary about an average value.
- In other words, dispersion helps to understand the distribution of the data.
- Measures of variation give information on the **spread** or **variability** or **dispersion** of the data values.
- The measure of **dispersion** shows how the data is spread or scattered around the mean.
- In statistics, the measures of dispersion help to interpret the variability of data
- i.e. to know how much homogenous or heterogeneous the data is.
- In simple terms, it shows how hugged or scattered the variable is

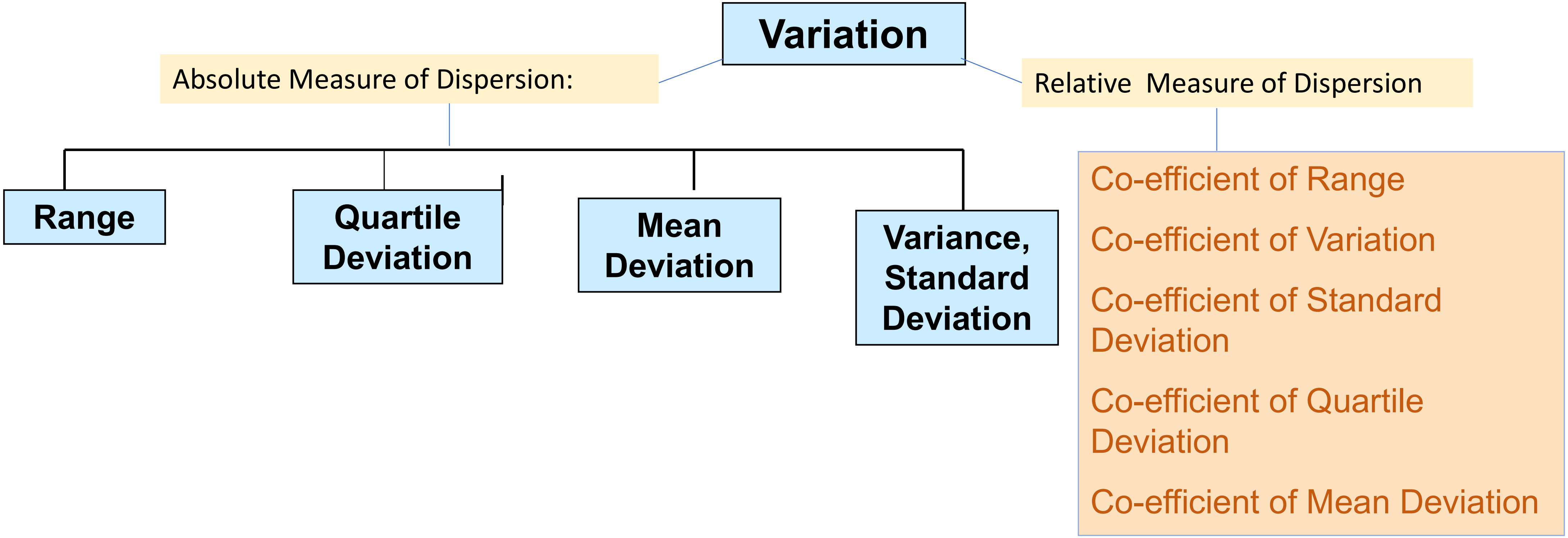
Types : Measures of Dispersion(Scattered ness)

- There are two main types of dispersion methods in statistics which are:
- Absolute Measure of Dispersion:
  - An absolute measure of dispersion contains the same unit as the original data set. Absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations.
  - It includes range, standard deviation, quartile deviation, etc.

Types : Measures of Dispersion(Scattered ness)

- The relative measures of depression are used to compare the distribution of two or more data sets.
- This measure compares values without units.
- Common relative dispersion methods includes(based on) Range, Quartiles, MD & SD.

Types :Measures of Dispersion(Scattered ness)

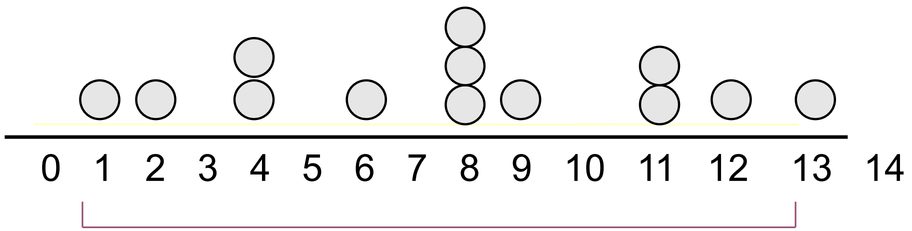


Absolute Measures of Dispersion: The Range

- Simplest measure of dispersion
- Difference between the largest and the smallest values:

$$\text{Range} = X_{\text{largest}} - X_{\text{smalles}} = A - B$$

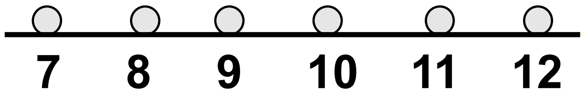
Example:

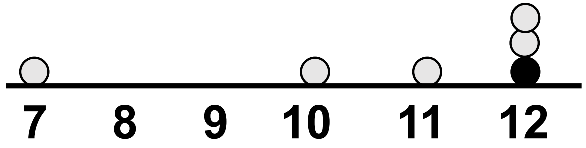


Range = 13 - 1 = 12

Measures of Dispersion: Why The Range is Crude measure

- Ignores the way in which data are distributed





Range = 12 - 7 = 5 | **Range = 12 - 7 = 5**

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**5**

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**120**

Range = 120 - 1 = 119

Range of Property worth

Frequency data

Property (Rs)	Boundaries	Frequency(000)
Lower	Upper	*f*
0	9999	3417
10000	24999	1303
25000	39999	1240
40000	49999	714
50000	59999	642
60000	79999	1361
80000	99999	1270
100000	149999	2708
150000	199999	1633
200000	299999	1242
300000	499999	870
500000	999999	367
1000000	1999999	125
2000000	4000000	41
	Total	16933

The range=difference Between Two extreme observations (A-B) is

=4 000 000 – 0

=4 000 000

Partition values :Quartile

- In statistics, Quartiles are the set of values which has three points dividing the data set into four identical parts.
- We ordinarily deal with a large amount of numerical data, in statistics.
- There are several concepts and formulas, which are extensively applicable in various researches and surveys.
- One of the best applications of quartiles is defined in box and whisker plot.

- Quartiles are the values that divide a list of numerical data into three quarters.
- The middle part of the three quarters measures the central point of distribution and shows the data which are near to the central point.
- The lower part of the quarters indicates just half information set which comes under the median and the upper part shows the remaining half, which falls over the median.
- In all, the quartiles depict the distribution or dispersion of the data set.

Partition values :Quartile

Quartiles Definition

- Quartiles divide the entire set into four equal parts.
- So, there are three quartiles, first, second and third represented by Q1, Q2 and Q3, respectively.
- Q2 is nothing but the median, since it indicates the position of the item in the list and thus, is a positional average.
- To find quartiles of a group of data, we have to arrange the data in ascending order.
- In the median, we can measure the distribution with the help of lesser and higher quartile.
- Apart from mean and median, there are other measures in statistics, which can divide the data into specific equal parts.

Partition values :Quartile

- A median divides a series into two equal parts. We can partition values of a data set mainly into three different ways:

Quartiles

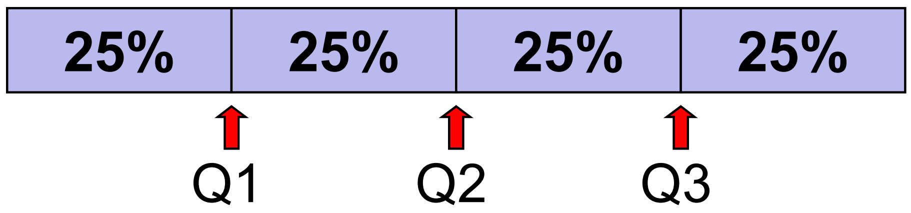
Deciles

Percentiles

- Quartiles split the ranked data into 4 segments with an equal number of values per segment
- Deciles split the ranked data into 10 segments with an equal number of values per segment
- Percentiles split the ranked data into 100 segments with an equal number of values per segment

Quartile Measures

- Quartiles split the ranked data into 4 segments with an equal number of values per segment



- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% of the observations are smaller and 50% are larger)
- Only 25% of the observations are greater than the third quartile

Quartile Measures: Locating Quartiles

Find a quartile by determining the value in the appropriate position in the ranked data, where

First quartile position: **Q_1 = (n+1)/4** ranked value

Second quartile position: **Q_2 = (n+1)/2** ranked value

Third quartile position: **Q_3 = 3(n+1)/4** ranked value

where **n** is the number of observed values

Quartile Measures:Locating Quartiles

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 is in the (9+1)/4 = 2.5 position of the ranked data

so use the value half way between the 2nd and 3rd values,

so **Q_1 = 12.5**

Q_1 and Q_3 are measures of non-central location
Q_2 = median, is a measure of central tendency

Quartile Measures Calculating The Quartiles: Example

Sample Data in Ordered Array:	**11 12 13 16 16 17 18 21 22**

(n = 9)

Q_1 is in the (9+1)/4 = 2.5 position of the ranked data,

so **Q_1 = (12+13)/2 = 12.5**

Q_2 is in the (9+1)/2 = 5^{th} position of the ranked data,

so **Q_2 = median = 16**

Q_3 is in the 3(9+1)/4 = 7.5 position of the ranked data,

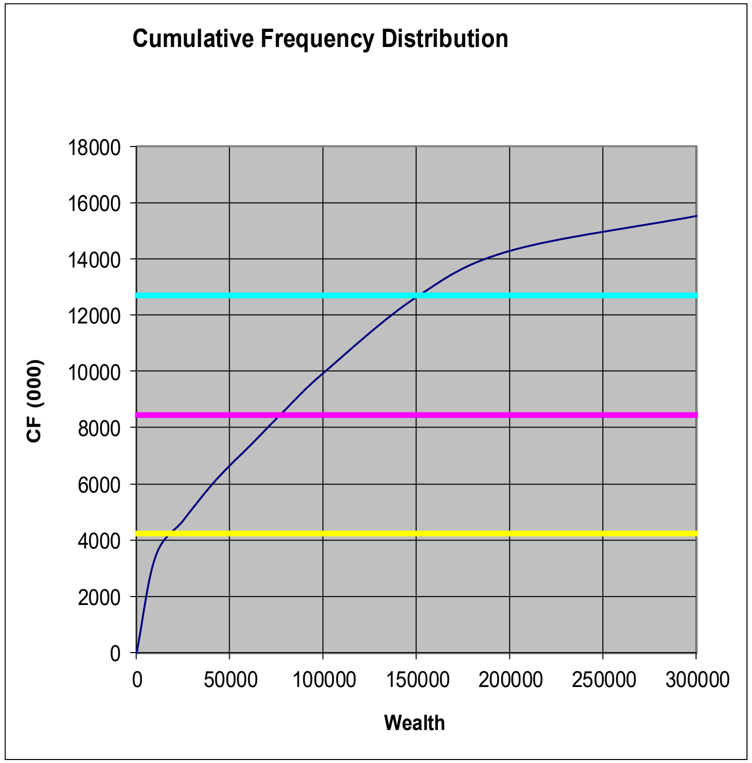
so **Q_3 = (18+21)/2 = 19.5**

Q_1 and Q_3 are measures of non-central location
Q_2 = median, is a measure of central tendency

Quartile Measures: Calculation Rules

- When calculating the ranked position use the following rules
  - If the result is a whole number, then it is the ranked position to use
  - If the result is a fractional half (e.g. 2.5, 7.5, 8.5, etc.) then average the two corresponding data values.
  - If the result is not a whole number or a fractional half then round the result to the nearest integer to find the ranked position.

Quartiles of Property (Wealth)



$$Jth\ quartile\ formula$$

$$Qj = l + \frac{\frac{jN}{4} - cf}{N} * h$$

J=1, 2, 3

The Lower Quartile Q_1 = 19 396

The Upper Quartile Q_3 = 151 370

The Inter-Quartile Range

IQR=151 370-19 396

= 131 974

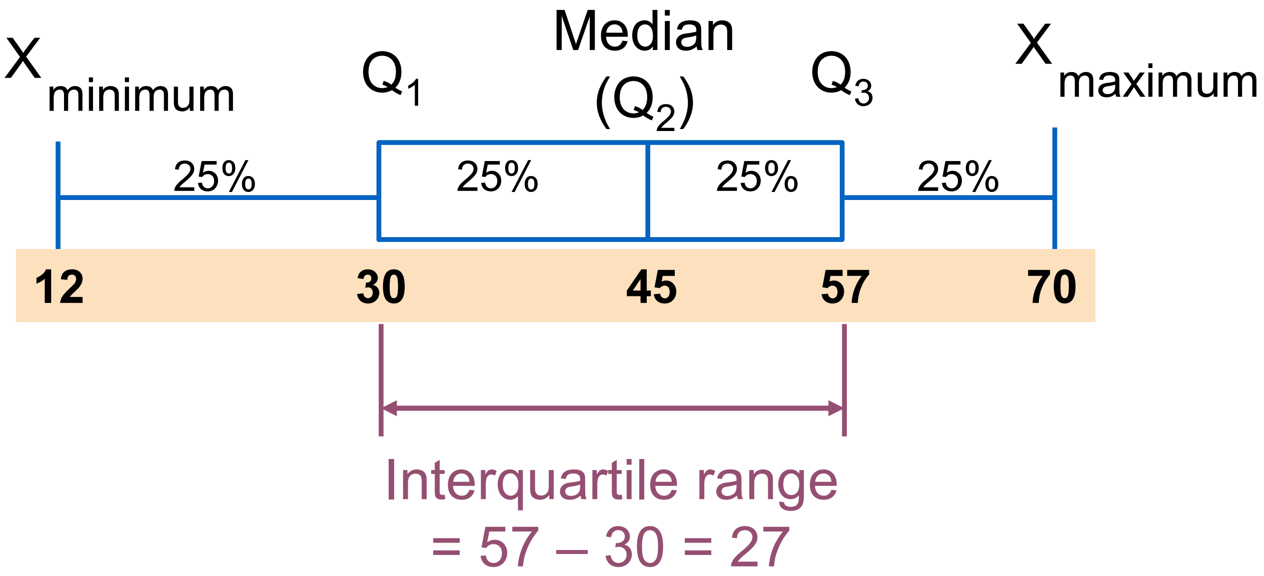
Quartile Measures:
Quartile Deviation or The semi-Interquartile Range (IQR)

- The IQR is $Q_3 – Q_1$ and measures the spread in the middle 50% of the data
- The IQR is a measure of variability that is not influenced by outliers or extreme values
- Measures like Q_1, Q_3, and IQR that are not influenced by outliers are called resistant measures
- Quartile deviation is defined as half of the distance between the third and the first quartile.
- It is also called Semi Interquartile range. If Q1 is the first quartile and Q3 is the third quartile, then the formula for deviation is given by;

Quartile deviation = $(Q_3-Q_1)/2$

Calculating The Interquartile Range

Example:

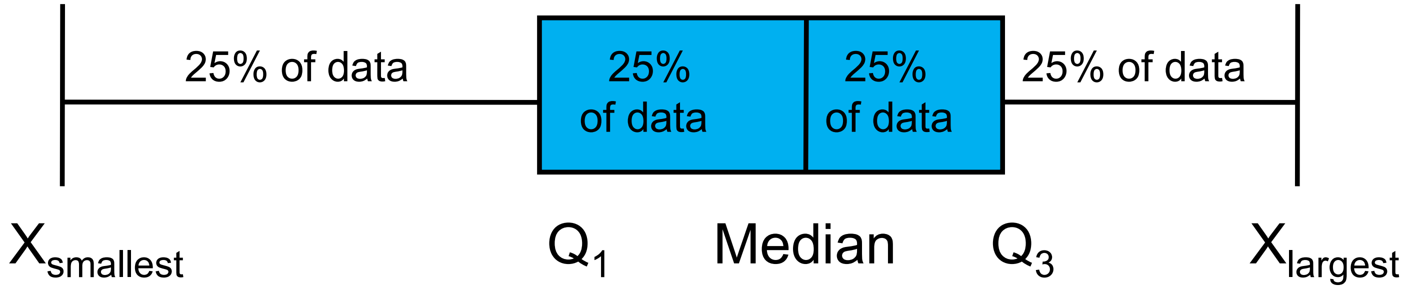


The Boxplot or Box and Whisker Diagram

- The Boxplot: A Graphical display of the data.

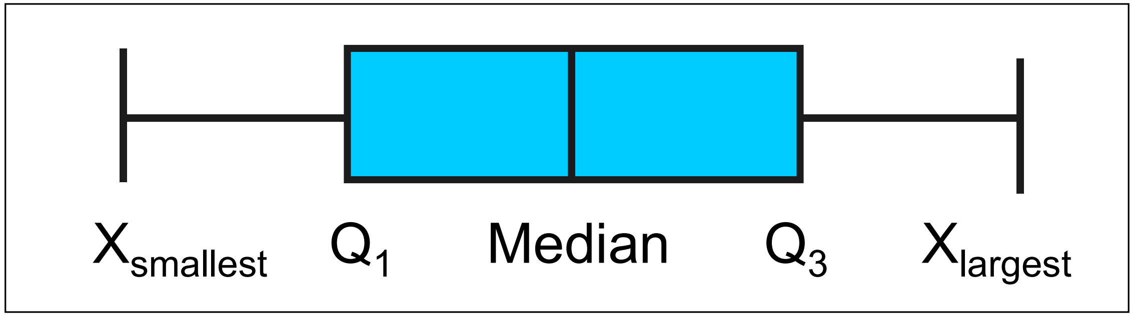
$X_{smallest}$ -- Q_1 -- Median -- Q_3 -- $X_{largest}$

Example:



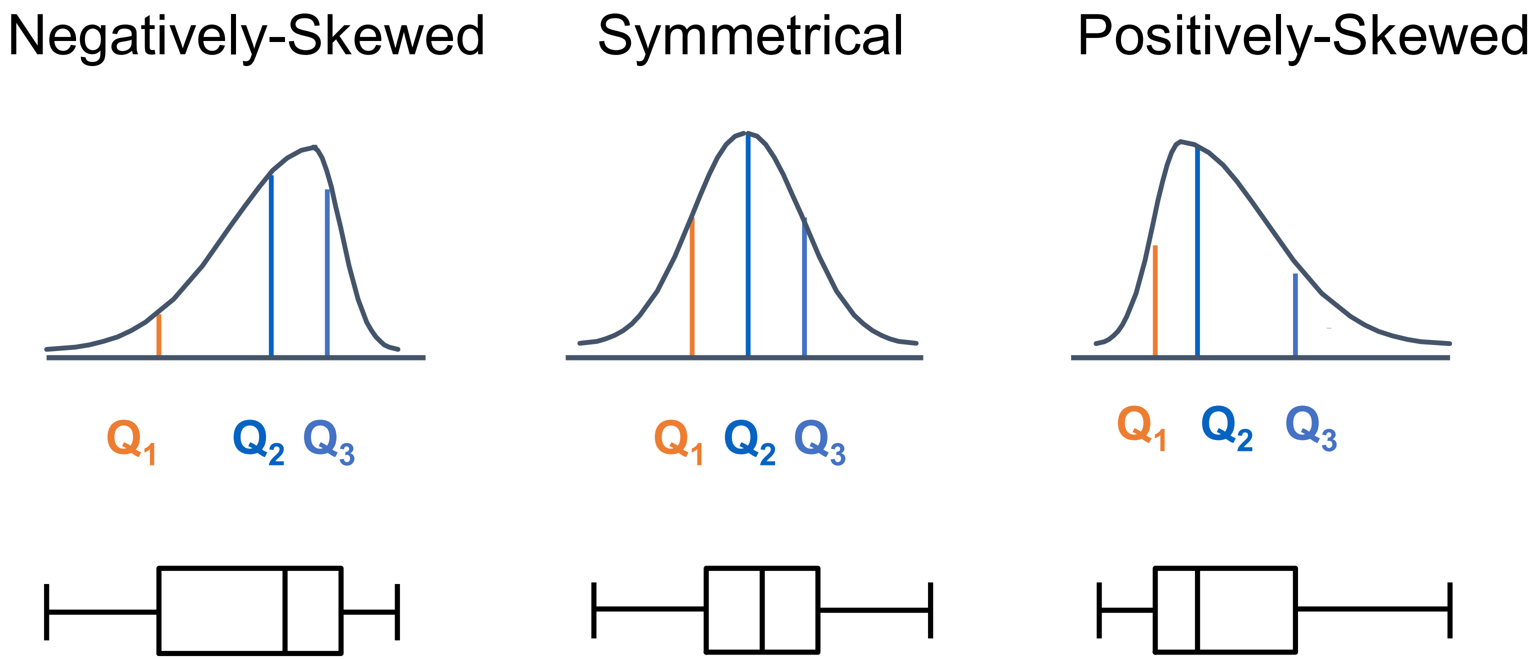
Shape of Boxplots

- If data are symmetric around the median then the box and central line are centered between the endpoints



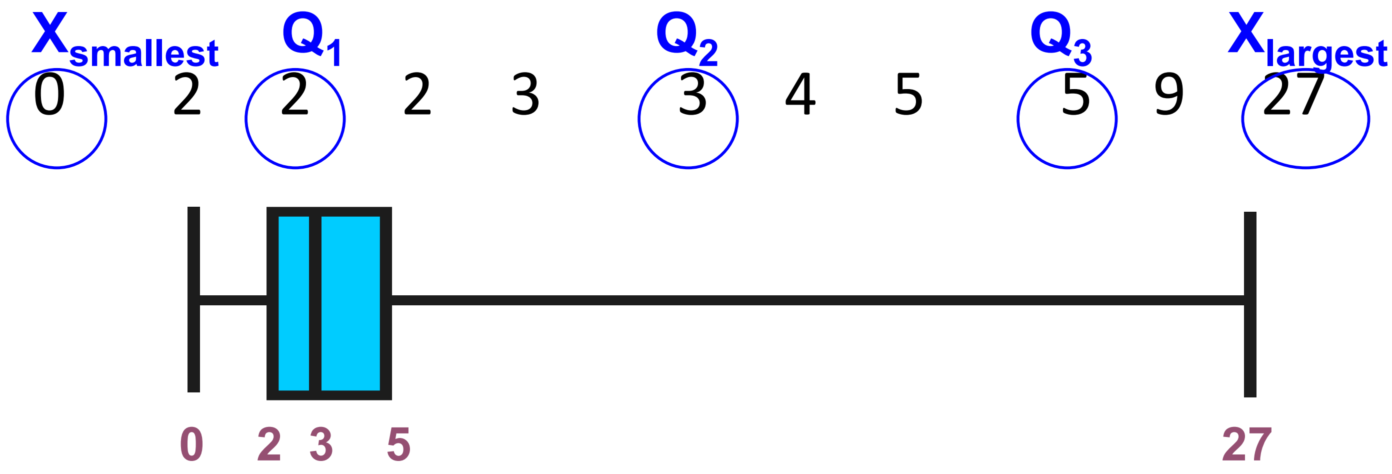
- A Boxplot can be shown in either a vertical or horizontal orientation

Distribution Shape and The Boxplot



Boxplot Example

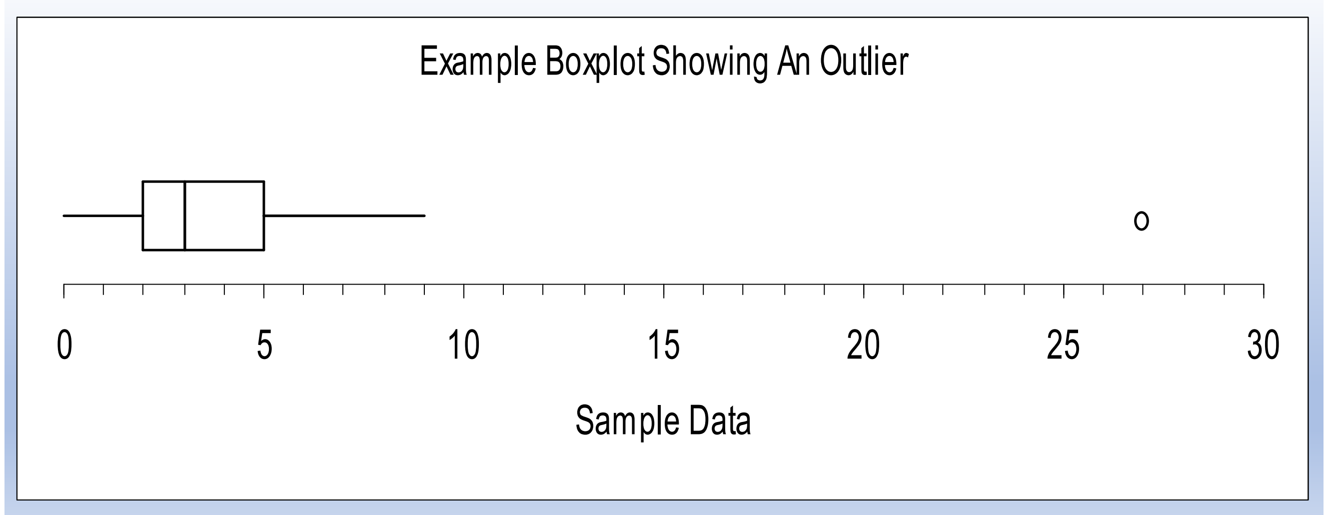
- Below is a Boxplot for the following data:



- The data are positively skewed.

Boxplot example showing an outlier

- The boxplot below of the same data shows the outlier value of 27 plotted v separately
- A value is considered an outlier if it is more than 1.5 times the interquartile range below Q_1 or above Q_3



Measures of Dispersion: The Mean Deviation (About mean)

- Average (approximately) of Absolute deviations of values from the mean

  - Mean Deviation (MD): $\boxed{MD = \frac{\sum_{i=1}^{n} |X_i - \overline{X}|}{n}}$ For Raw Data or Ungrouped

    $\boxed{MD = \frac{\sum_{i=1}^{n} f_i |X_i - \overline{X}|}{N}}$ For Frequency Data

Where $\overline{X}$ = arithmetic mean

n = size, N= Total frequency

X_i = i^{th} value of the variable X

f_i= Frequency of ith Variable

Measures of Dispersion: The Variance

- Average (approximately) of squared deviations of values from the mean

  - Sample variance: $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

formula for Variance –Frequency distribution

- Sample Variance
  with frequency table

$$s^2 = \frac{\sum x^2 f}{n-1} - \bar{x}^2$$

$\overline{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

f = frequency

For A Population: The Variance σ^2

- Average of squared deviations of values from the mean

  - Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Where μ = population mean

N = population size

X_i = i^{th} value of the variable X

Measures of Dispersion: The Standard Deviation s

- Most commonly used measure of variation
- Shows variation about the mean
- Is the **square root of the variance**
- Has the same units as the original data
  - Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

For A Population: The Standard Deviation σ

- Most commonly used measure of variation
- Shows variation about the mean
- Is the **square root of the population variance**
- Has the same units as the original data

- Population standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Approximating the Standard Deviation from a Frequency Distribution

- Assume that all values within each class interval are located at the midpoint of the class

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

Where
n = number of values or sample size
x = midpoint of the j^{th} class
f = number of values in the j^{th} class

Summary : Measures of Dispersion

Range	$X_{\text{largest}} - X_{\text{smallest}}$	Total Spread
Standard Deviation (Sample)	$\sqrt{\dfrac{\sum (X_i - \bar{X})^2}{n-1}}$	Dispersion about Sample Mean
Standard Deviation (Population)	$\sqrt{\dfrac{\sum (X_i - \mu_X)^2}{N}}$	Dispersion about Population Mean
Variance (Sample)	$\dfrac{\Sigma (X_i - \bar{X})^2}{n-1}$	Squared Dispersion about Sample Mean

Measures of Dispersion: The Standard Deviation(SD)

Steps for Calculating Standard Deviation

1. Calculate the difference between each value and the mean.
2. Square each difference.
3. Add the squared differences.
4. Divide this total by n-1 to get the sample variance.
5. Take the square root of the sample variance to get the sample standard deviation.

Measures of Dispersion: Sample Standard Deviation: Calculation Example

Sample Data (X_i) :

10	12	14	15	17	18	18	24

n = 8 Mean = $\overline{X}$ = 16

$$S = \sqrt{\frac{(10-\overline{X})^2 + (12-\overline{X})^2 + (14-\overline{X})^2 + \cdots + (24-\overline{X})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

$$= \sqrt{\frac{130}{7}} = \boxed{4.3095}$$

A measure of the "average" scatter around the mean

Standard Deviation of Wealth

The Distribution of Marketable Wealth

Wealth Lower	Boundaries Upper	Mid interval(000) x	Frequency(000) f	*fx squared*
0	9999	5.0	3417	85425
10000	24999	17.5	1303	399043.75
25000	39999	32.5	1240	1309750
40000	49999	45.0	714	1445850
50000	59999	55.0	642	1942050
60000	79999	70.0	1361	6668900
80000	99999	90.0	1270	10287000
100000	149999	125.0	2708	42312500
150000	199999	175.0	1633	50010625
200000	299999	250.0	1242	77625000
300000	499999	400.0	870	139200000
500000	999999	750.0	367	206437500
1000000	1999999	1500.0	125	281250000
2000000	4000000	3000.0	41	369000000
		Total	16933	1187973644

Mean = 131.443

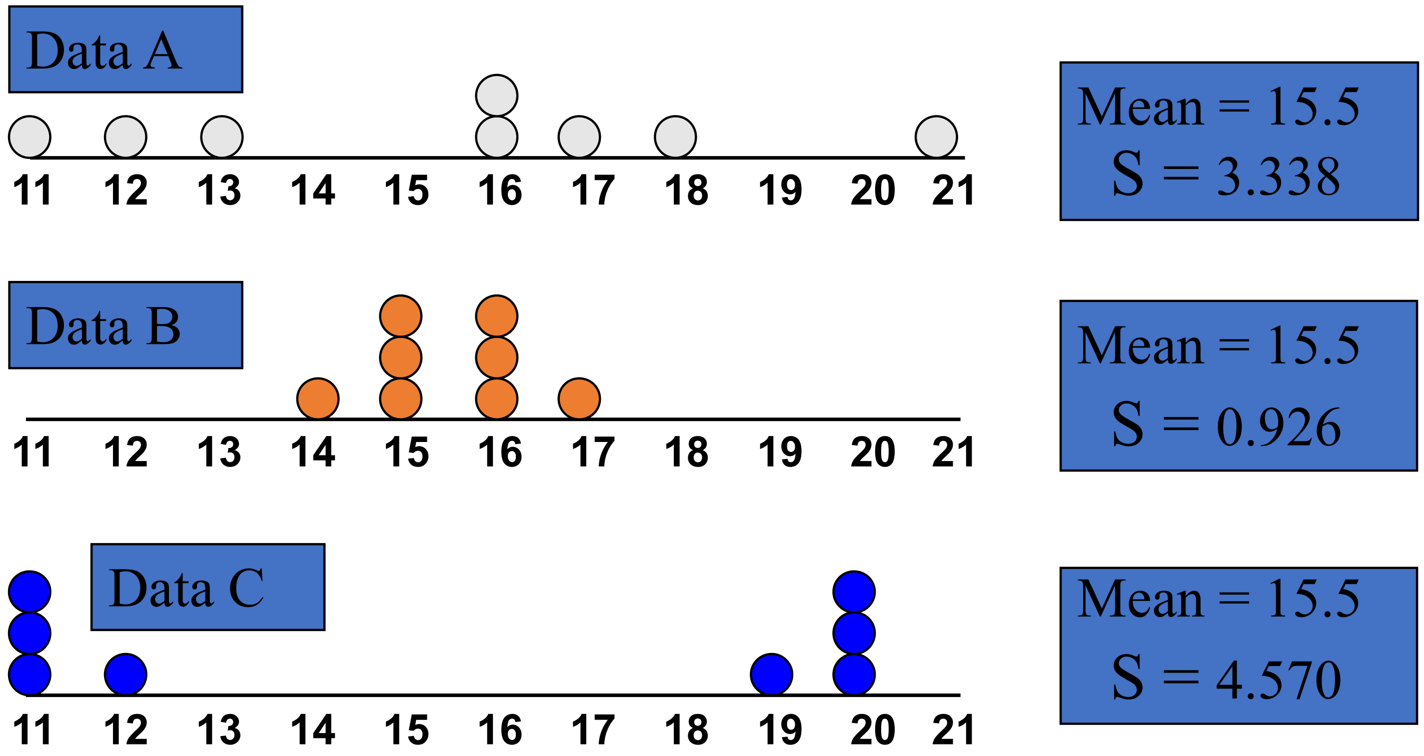
Variance = $\frac{1187973644}{16933}$ _ 131.443 squared

Variance = 52880.043

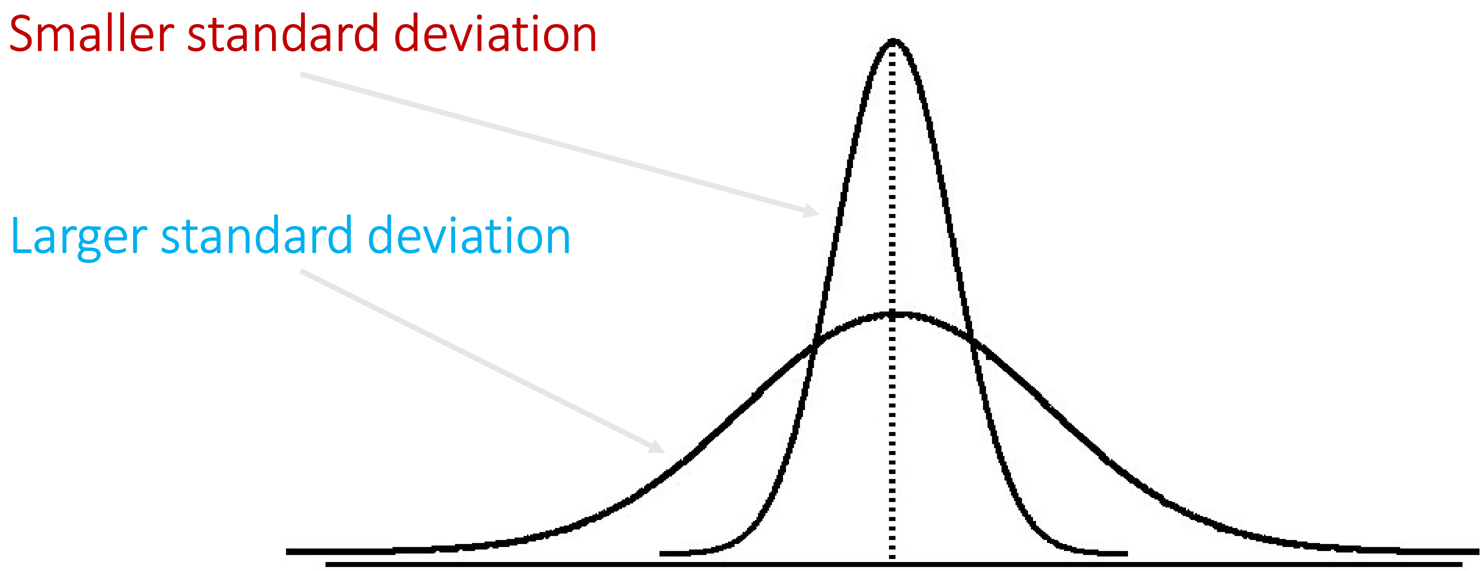
Standard deviation = 229.957

Standard deviation = 229 957

Measures of Dispersion: Comparing Standard Deviations



Measures of Dispersion: Comparing Standard Deviations



Measures of Dispersion: Summary Characteristics

- The more the data are spread out, the greater the range, variance, and standard deviation.
- The less the data are spread out, the smaller the range, variance, and standard deviation.
- If the values are all the same (no variation), all these measures will be zero.
- None of these measures are ever negative.

Relative Measures : Coefficient of Dispersion

- The coefficients of dispersion are calculated (along with the measure of dispersion) when two series are compared, that differ widely in their averages.
- The dispersion coefficient is also used when two series with different measurement unit, are compared. It is denoted as C.D.
- The common coefficients of dispersion are:

C.D. In Terms of Variation Measures	Coefficient of dispersion
Range	C.D. = $(X_{max} - X_{min}) / (X_{max} + X_{min})$
Quartile Deviation	C.D. = $(Q3 - Q1) / (Q3 + Q1)$
Standard Deviation (S.D.)	C.D. = S.D. / Mean
Mean Deviation	C.D. = Mean deviation/Average

Measures of Dispersion: The Coefficient of Variation

- Measures **relative variation**
- Always in percentage (%)
- Shows **variation relative to mean**
- Can be used to compare the variability of two or more sets of data
- measured in different units

$$CV = \left(\frac{S}{\overline{X}}\right) * 100$$

The Coefficient of Variation

- Coefficient of Variation of a population:

$$\mathrm{CV} = \left(\frac{\sigma}{\mu}\right)*100$$

- This can be used to compare two distributions directly to see which has more dispersion because it does not depend on units of the distribution.

Measures of Dispersion:
Comparing Coefficients of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{S}{\overline{\overline{X}}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_B = \left(\frac{S}{\overline{\overline{X}}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

Coefficient of Variation of Wealth

Coefficient of variation = $\frac{\sigma}{\mu}$

=229.957 / 131.443

= 1.749

The standard deviation is 1.75% of the mean.

Sample statistics versus population parameters

Measure	Population Parameter	Sample Statistic
Mean	μ	$\overline{X}$
Variance	σ^2	S^2
Standard Deviation	σ	S

Pitfalls in Numerical Descriptive Measures

- Data analysis is **objective**
  - Should report the summary measures that best describe and communicate the important aspects of the data set

- Data interpretation is **subjective**
  - Should be done in fair, neutral and clear manner