Probability and Statistics (PHM111s)-Lecture 4

Part I: Introduction to Statistical Methods.

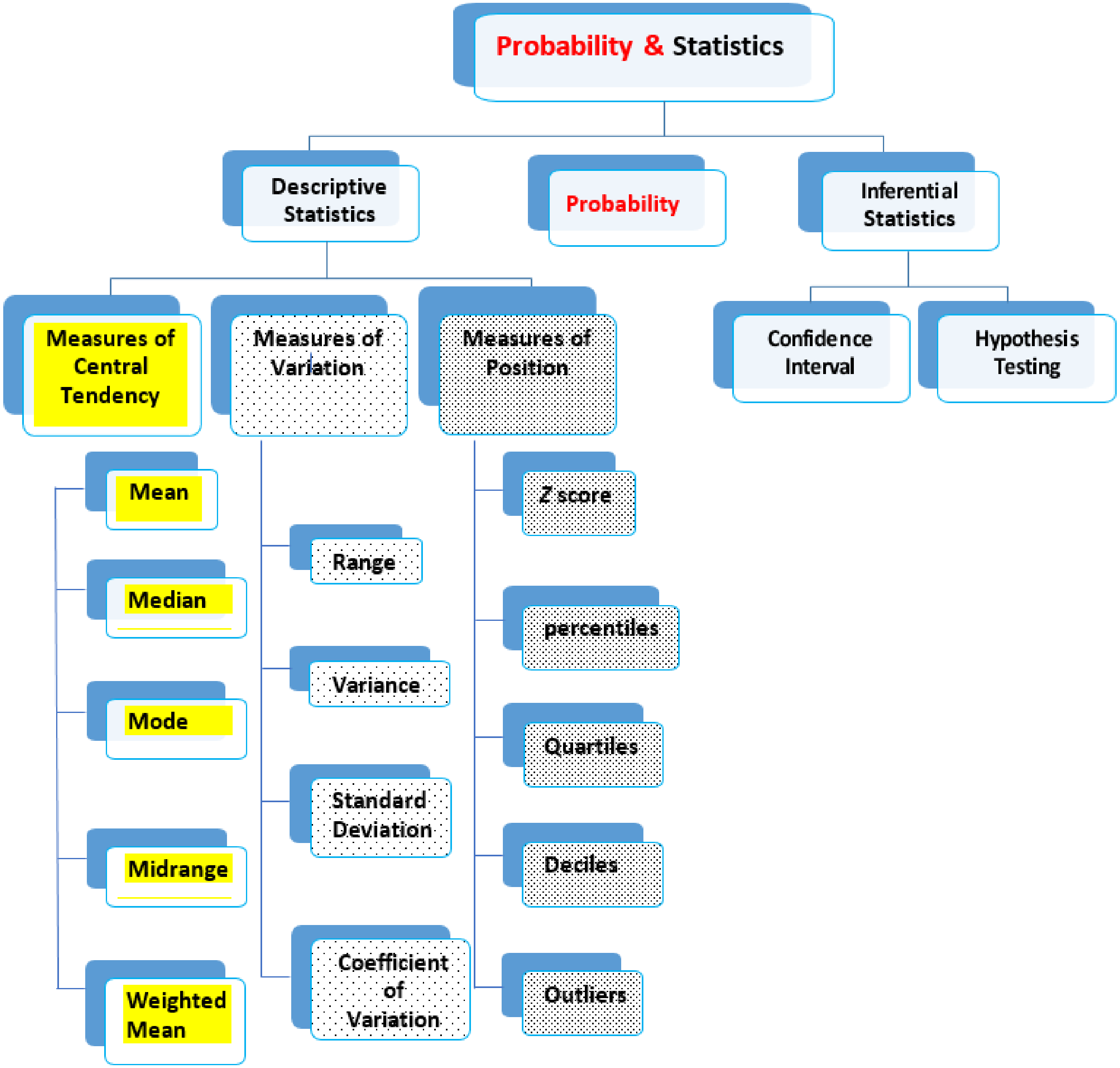
Part II: Methods of Descriptive Statistics.

1-Collecting Data.
2-Organizing Data.
3-Presenting Data.
4-Summarizing Data.

Part III: Introduction to Probability.

Part IV: Methods of Inferential Statistics.

4. Summarizing Data (cont.)



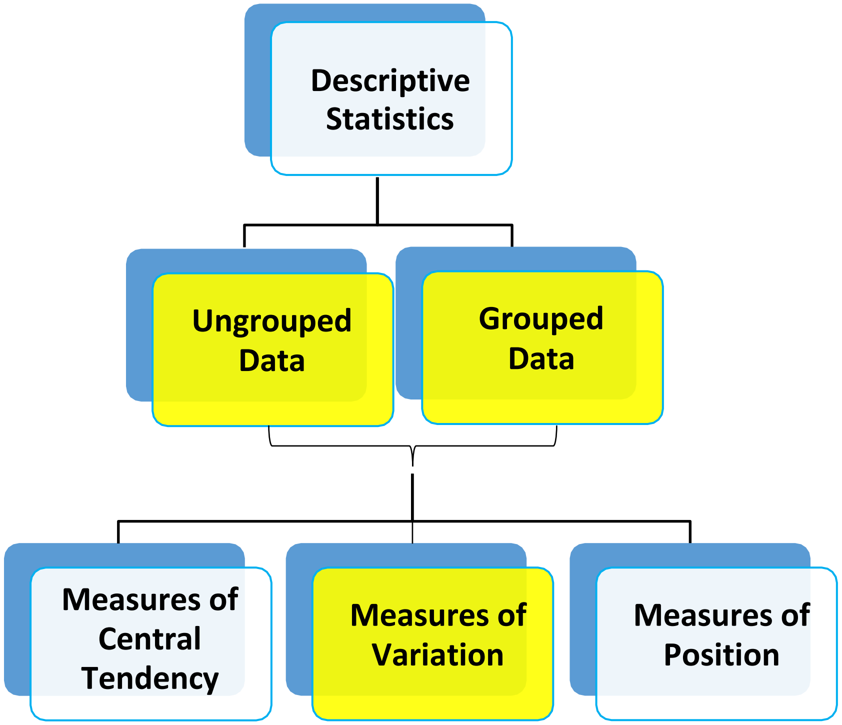
- Probability & Statistics
  - Descriptive Statistics
    - Measures of Central Tendency
      - Mean
      - Median
      - Mode
      - Midrange
      - Weighted Mean
    - Measures of Variation
      - Range
      - Variance
      - Standard Deviation
      - Coefficient of Variation
    - Measures of Position
      - Z score
      - percentiles
      - Quartiles
      - Deciles
      - Outliers
  - Probability
  - Inferential Statistics
    - Confidence Interval
    - Hypothesis Testing

Example: Imagine two sets of 5 students each and an exam of maximum mark 50 marks is given for each set, the marks of the students were as follow:

The set A: 29, 26, 35, 35, 35

The set B: 8, 35, 49, 35, 33

	Mean	Median	Mode
Set A	32	35	35
Set B	32	35	35



1- Range:

Example 1: The salaries for the staff of the XYZ Manufacturing Co. are shown here. Find the range.

Staff	**Salary**
Owner	$100,000
Manager	40,000
Sales representative	30,000
Workers	25,000
	15,000
	18,000

Solution

The range is R = $100,000 - $15,000 = $85,000.

Example: The range of set X: 55, 53, 57, 56 and 54 = 57-53 = 4

The range of set Y: 67, 73, 41, 60 and 34 = 73-34 = 39

2- A- Population Variance and Standard Deviation

The **variance** is the average of the squares of the distance each value is from the mean.

The symbol for the population variance is σ^2.

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

The **standard deviation** is the square root of the variance.

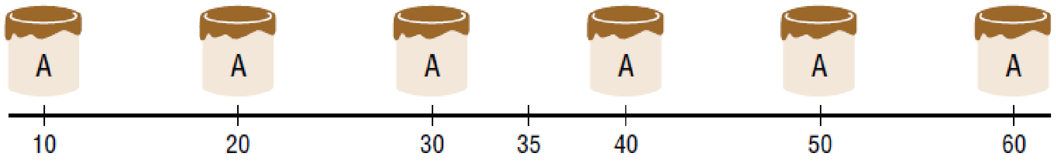
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X-\mu)^2}{N}}$$

Example 2:

Find the variance and standard deviation for brand A paint.

10, 60, 50, 30, 40, 20

Variation of paint (in months)



(a) Brand A

Solution: $\mu = \dfrac{\sum X}{N} = \dfrac{10+60+50+30+40+20}{6} = \dfrac{210}{6} = 35$

A **Values X**	**B** $X-\mu$	**C** $(X-\mu)^2$
10	-25	625
60	+25	625
50	+15	225
30	-5	25
40	+5	25
20	-15	225
		1750

Variance = 1750 ÷ 6 = 291.7

Standard deviation equals $\sqrt{291.7}$, or 17.1.

B- Sample Variance and Standard Deviation

Case 1: Ungrouped Data

The formula for the sample variance, denoted by s^2, is

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

The symbol for the sample standard deviation is s .

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

The shortcut formulas for computing the variance and standard deviation for data obtained from samples are as follows:

Variance	**Standard deviation**
$s^2 = \dfrac{n(\sum X^2) - (\sum X)^2}{n(n-1)}$	$s = \sqrt{\dfrac{n(\sum X^2) - (\sum X)^2}{n(n-1)}}$

Example 3: Find the sample variance and standard deviation for the amount of European auto sales for a sample of 6 years shown. The data are in millions of dollars.

11.2, 11.9, 12.0, 12.8, 13.4, 14.3

Solution

$$\sum X = 11.2 + 11.9 + 12.0 + 12.8 + 13.4 + 14.3 = 75.6$$

$$\sum X^2 = 11.2^2 + 11.9^2 + 12.0^2 + 12.8^2 + 13.4^2 + 14.3^2 = 958.94$$

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

$$= \frac{6(958.94) - 75.6^2}{6(6-1)}$$

$$= \frac{38.28}{30} = 1.276$$

$$s = \sqrt{1.28} = 1.13$$

Case 2: Grouped Data

$$s^2 = \frac{n(\sum f.X_m{}^2) - (\sum f.X_m)^2}{n(n-1)}$$

Example 4 Find the sample variance and the standard deviation for the frequency distribution of the following data:

Class	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2

Solution

A Class	B Frequency	C Midpoint	D $f.X_m$	E $f.X_m^2$
5.5–10.5	1	8	8	64
10.5–15.5	2	13	26	338
15.5–20.5	3	18	54	972
20.5–25.5	5	23	115	2,645
25.5–30.5	4	28	112	3,136
30.5–35.5	3	33	99	3,267
35.5–40.5	2	38	76	2,888
	$n = 20$		$\sum f.X_m = 490$	$\sum f.X_m^2 = 13{,}310$

$$s^2 = \frac{n(\sum f.X_m{}^2) - (\sum f.X_m)^2}{n(n-1)}$$

$$= \frac{20(13{,}310) - 490^2}{20(20-1)}$$

$$= \frac{266{,}200 - 240{,}100}{20(19)}$$

$$= \frac{26{,}100}{380} = 68.7$$

$$s = \sqrt{68.7} = 8.3$$

3- Coefficient of Variation

For samples,

$$\text{CVar} = \frac{s}{\bar{X}}.100\%$$

For populations,

$$\text{CVar} = \frac{\sigma}{\mu}.100\%$$

Example 5 The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

Solution

The coefficients of variation are

$$\text{CVar} = \frac{s}{\bar{X}}.100\% = \frac{5}{87}.100\% = 5.7\% \qquad \text{sales}$$

$$\text{CVar} = \frac{s}{\bar{X}}.100\% = \frac{773}{5225}.100\% = 14.8\% \qquad \text{commissions}$$

Since the coefficient of variation is larger for commissions, the **commissions are more variable than the sales**.

Example

	$\bar{X}$	s
Height	65″	3″
Weight	175 lbs	4 lbs

Solution

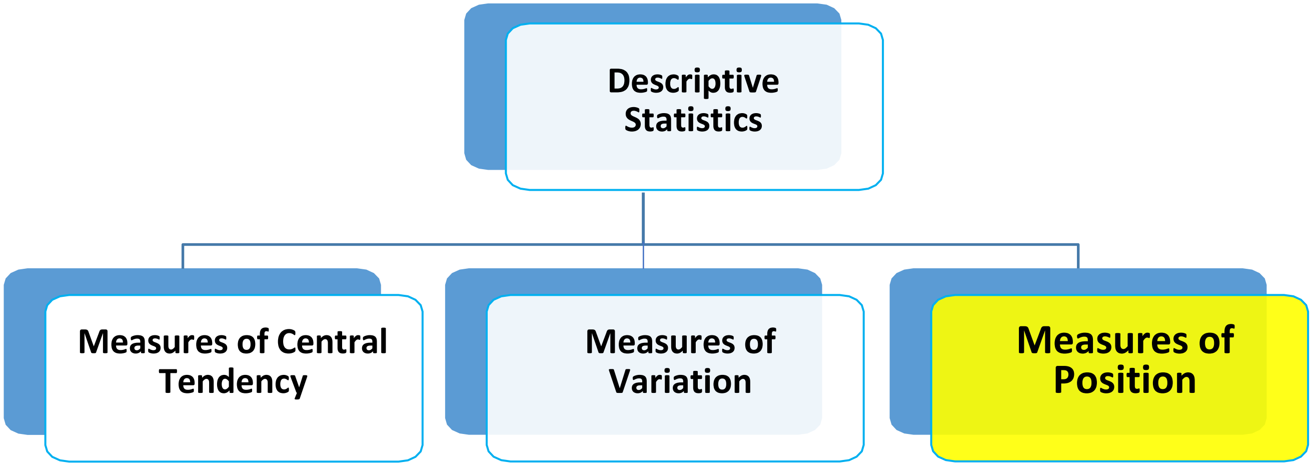
The coefficients of variation are

$$(\text{CVar})_H = \frac{s}{\bar{X}}.100\% = \frac{3}{65}.100\% = 4.6\%$$

$$(\text{CVar})_W = \frac{s}{\bar{X}}.100\% = \frac{4}{175}.100\% = 2.3\%$$

Since the coefficient of variation is larger for Height, the **Height is more variable than the Weight** (i.e. more spread).

3-3 Measures of Position



Standard Scores

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

For samples, the formula is

$$z = \frac{X - \overline{X}}{s}$$

For populations, the formula is

$$z = \frac{X - \mu}{\sigma}$$

The *z* score represents the number of standard deviations that a data value falls above or below the mean.

Example 4–6: A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

Solution

First, find the *z* scores. For calculus the z score is

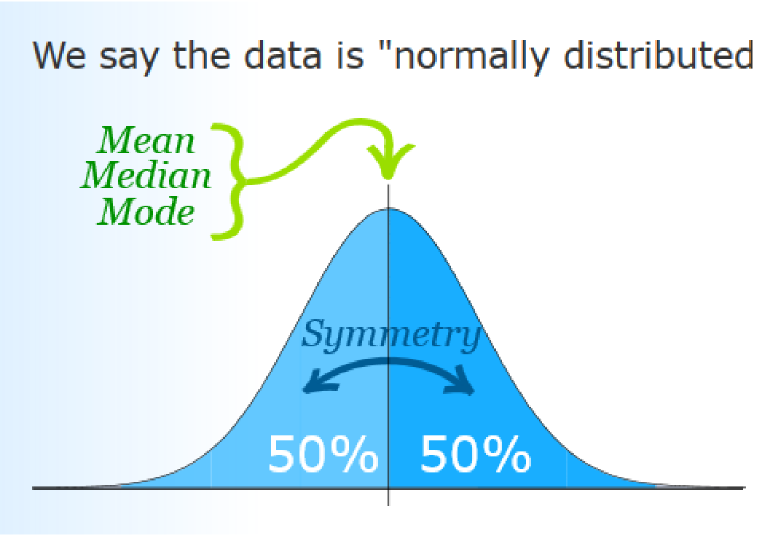
$$z = \frac{X - \overline{X}}{s} = \frac{65 - 50}{10} = 1.5$$

For history the z score is

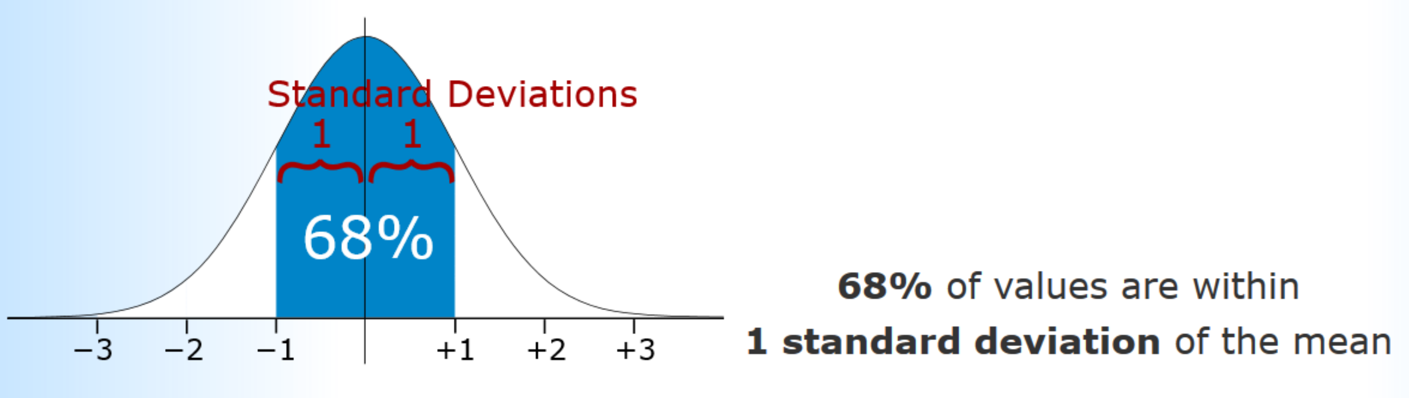
$$z = \frac{30 - 25}{5} = 1.0$$

Since the *z* score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

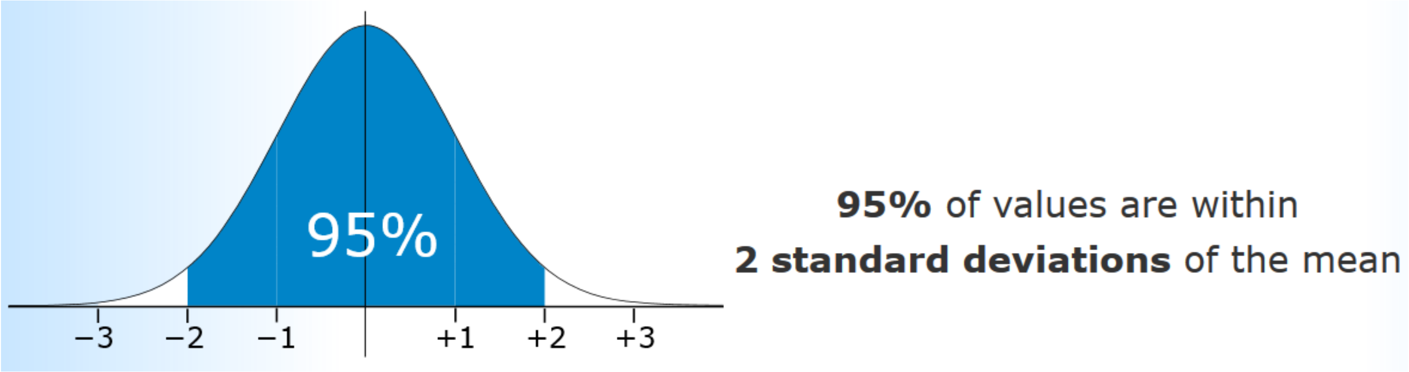
When all data for a variable are transformed into *z* scores, the resulting distribution will have a mean of 0 and a standard deviation of 1. A *z* score, then, is actually the number of standard deviations each value is from the mean for a specific distribution.



Empirical Rule:

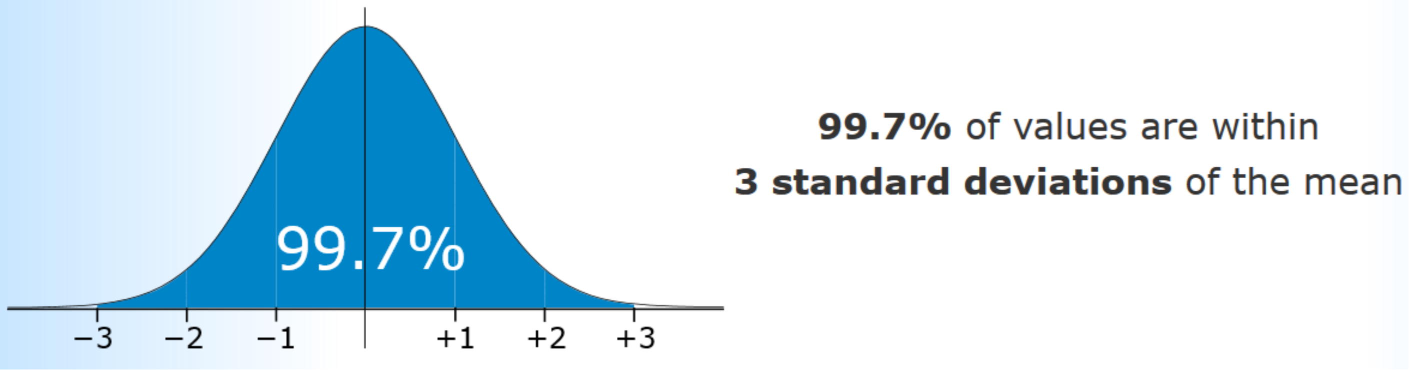


68% of values are within
1 standard deviation of the mean



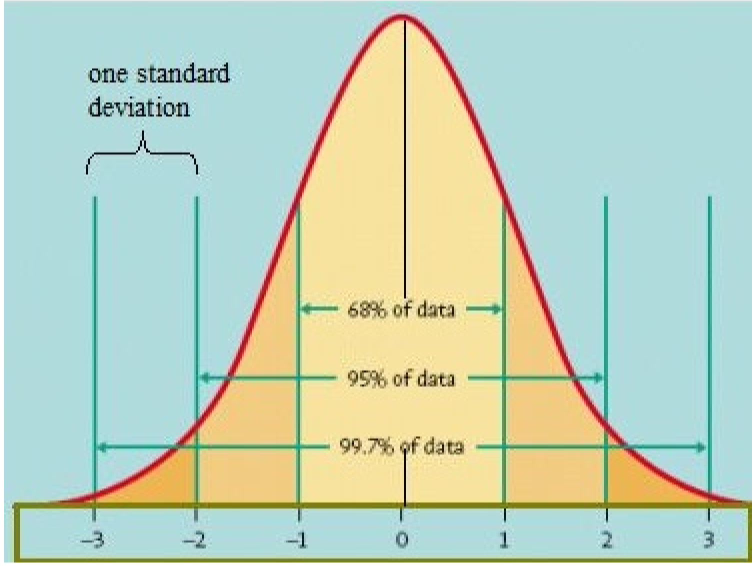
95% of values are within
2 standard deviations of the mean

"Usual"



99.7% of values are within
3 standard deviations of the mean

"Unusual" then "very rare"



Example

	Population 1	Population 2
X (individual value)	76″	86″
μ (population mean)	71.5″	80″
σ (population S.D.)	2.1″	3.3″
$z = \dfrac{X-\mu}{\sigma}$	**2.14 (<u>unusual</u>)**	**1.82 (<u>usual</u>)**

Percentiles

Percentiles divide the data set into 100 equal groups.

Percentile Formula

The percentile corresponding to a given value X is computed by using the following formula:

$$\text{Percentile rank} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} .100\%$$

and the order of the value corresponding to certain percentile is $\mathrm{c} = \dfrac{n.p}{100}$

where

n = total number of values

p = percentile rank

Example 4–7: A teacher gives a 20-point test to 10 students. The scores are shown here. Find the percentile rank of a score of 12. Also find the value corresponding to the 25th percentile.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Solution:

Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Then substitute into the formula.

$$\text{Percentile rank} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} .100\%$$

Since there are six values below a score of 12, the solution is

$$\text{Percentile rank} = \frac{6 + 0.5}{10} .100\% = 65\text{th percentile}$$

$$\text{and } \mathrm{c} = \frac{10 \ . \ 25}{100} = 2.5 \Rightarrow c = 3 \ \text{(third order)}$$

Hence, the value 5 corresponds to the 25th percentile.

(Note: If c is not a whole number, round it up to the next whole number as in this example.)

Thus, a student whose score was 12 did better than 65% of the class.

Example 4–8: Using the data set in the previous Example, find the value that corresponds to the 60th percentile.

Solution

Arrange the data in order from smallest to largest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Substitute in the formula.

$$c = \frac{n.p}{100} = \frac{10.60}{100} = 6$$

If *c* is a whole number, use the value halfway between the *c* and *c* +1 values when counting up from the lowest value. In this case, the 6th and 7th values.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

6th value 7th value

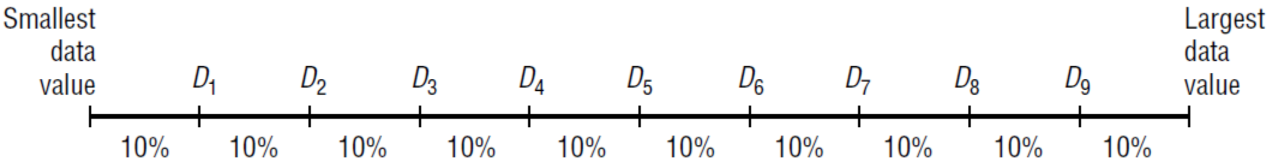
The value halfway between 10 and 12 is 11. Find it by adding the two values and dividing by 2.

$$\frac{10+12}{2} = 11$$

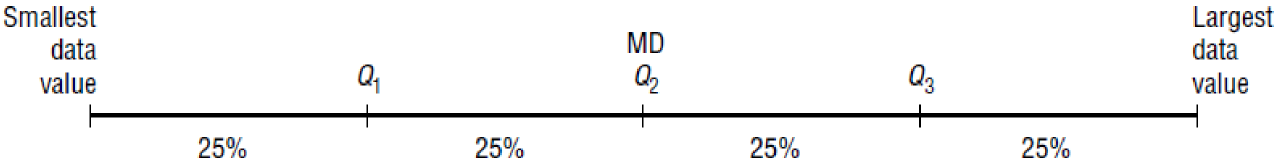
Hence, 11 corresponds to the 60th percentile. Anyone scoring 11 would have done better than 60% of the class.

Deciles and Quartiles:

Deciles divide the distribution into 10 groups, as shown. They are denoted by D_1, D_2,…, D_{10} etc.



Quartiles divide the distribution into four groups, separated by Q_1, Q_2, Q_3.



Case 1: Ungrouped Data

Example 4–9: Find Q_1, Q_2, and Q_3 for the data set 15, 13, 6, 5, 12, 50, 22, 18.

Solution

Step 1 Arrange the data in order.

5, 6, 12, 13, 15, 18, 22, 50

Step 2 Find the median (Q_2).

5, 6, 12, 13, 15, 18, 22, 50

↑

MD

$$Q_2 = MD = \frac{13+15}{2} = 14$$

Step 3 Find the median of the data values less than 14.

5, 6, 12, 13

$$\underset{Q_1}{\uparrow}$$

$$Q_1 = \frac{6+12}{2} = 9$$

So Q_1 is 9.

Step 4 Find the median of the data values greater than 14.

15, 18, 22, 50

$$\underset{Q_3}{\uparrow}$$

$$Q_3 = \frac{18+22}{2} = 20$$

Here Q_3 is 20. Hence, $Q_1 = 9$, $Q_2 = 14$, and $Q_3 = 20$.

Interquartile range (IQR)

$$\text{IQR} = Q_3 - Q_1$$

Midhinge:

$$\text{Midhinge} = \frac{Q_1 + Q_3}{2}$$

Case 2: Grouped Data

Using the same method of calculations as in the median, we can get Q_1 and Q_3 equation as follows:

$$Q_1 = L_{Q_1} + \left(\frac{\frac{n}{4} - F}{f_{Q_1}} \right) i , \qquad Q_3 = L_{Q_3} + \left(\frac{\frac{3n}{4} - F}{f_{Q_3}} \right) i$$

Example 4–10: Based on the grouped data below, find the interquartile range (IQR).

Time to travel to work	*f*
1 – 10	8
11 – 20	14
21 – 30	12
31 – 40	9
41 – 50	7

Solution: Construct the cumulative frequency distribution

Height (in cm)	*f*	*cf*
1 – 10	**8**	**8**
11 – 20	**14**	**22**
21 – 30	**12**	**34**
31 – 40	**9**	**43**
41 – 50	**7**	**50**

$$\text{Class } Q_1 = \frac{n}{4} = \frac{50}{4} = 12.5 \rightarrow \text{ class } \mathrm{Q}_1 \text{ is the } 2^{nd} \text{ class}$$

$$Q_1 = L_{Q_1} + \left(\frac{\frac{n}{4} - F}{f_{Q_1}} \right) i$$

$$= 10.5 + \left(\frac{12.5 - 8}{14} \right) 10 = 13.7143$$

$$\text{Class } Q_3 = \frac{3n}{4} = \frac{3(50)}{4} = 37.5 \rightarrow \text{ class } \mathrm{Q}_3 \text{ is the } 4^{th} \text{ class}$$

$$Q_3 = L_{Q_3} + \left(\frac{\frac{3n}{4} - F}{f_{Q_3}} \right) i$$

$$= 30.5 + \left(\frac{37.5 - 34}{9} \right) 10 = 34.3889$$

IQR = Q_3 - Q_1 = 34.3889 - 13.7143 = 20.6746

Outliers

An **outlier** is an extremely high or an extremely low data value when compared with the rest of the data values.

$$x > Q_3 + 1.5(\text{IQR}) \qquad \text{or} \qquad x < Q_1 - 1.5(\text{IQR})$$

Example 4–11: Check the following data set for outliers.

5, 6, 12, 13, 15, 18, 22, 50

Solution

The data value 50 is extremely suspect. These are the steps in checking for an outlier.

Step 1 Find Q_1 and Q_3. From the previous example, Q_1 is 9 and Q_3 is 20.

Step 2 Find the interquartile range (IQR), which is Q_3 - Q_1.
$\text{IQR} = Q_3 - Q_1 = 20 - 9 = 11$

Step 3 Multiply this value by 1.5.
$1.5(11) = 16.5$

Step 4 Subtract the value obtained in step 3 from Q_1, and add the value obtained in step 3 to Q_3.
$9 - 16.5 = -7.5$ and $20 + 16.5 = 36.5$

Step 5 Check the data set for any data values that fall outside the interval from -7.5 to 36.5. The value 50 is outside this interval; hence, it can be considered an outlier.

3–4 Exploratory Data Analysis

A **boxplot** is a graph of a data set obtained by drawing a horizontal line from the minimum data value to Q_1, drawing a horizontal line from Q_3 to the maximum data value, and drawing a box whose vertical sides pass through Q_1 and Q_3 with a vertical line inside the box passing through the median or Q_2.

Example 4–12: The number of meteorites found in 10 states of the United States is 89, 47, 164, 296, 30, 215, 138, 78, 48, 39. Construct a boxplot for the data.

Solution

Step 1 Arrange the data in order:

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

Step 2 Find the median.

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

↑

Median

$$Q_2 = MD = \frac{78+89}{2} = 83.5$$

Step 3 Find Q_1.

30, 39, 47, 48, 78

↑

Q_1

Step 4 Find Q_3.

89, 138, 164, 215, 296

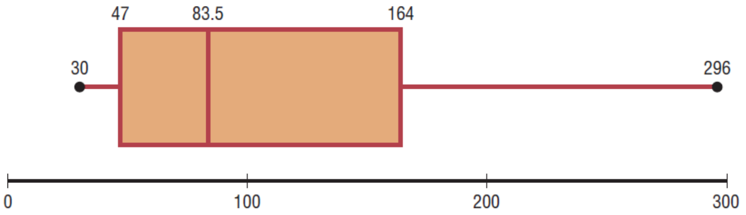
↑

Q_3

Step 5 Draw a scale for the data on the x axis.

Step 6 Located the lowest value, Q_1, median, Q_3, and the highest value on the scale.

Step 7 Draw a box.


47
83.5
164
30
296
0
100
200
300