Chapter 3 - Descriptive Statistics

Characteristics of Data:

1. Center: A representative or average value that indicates where the middle of the data set is located.

2. Variation: A measure of the amount that the data values vary.

3. Distribution: The nature or shape of the spread of the data over the range of values (such as bell-shaped, uniform, or skewed).

4. Outliers: Sample values that lie very far away from the vast majority of the other sample values.

5. Time: Changing characteristics of the data over time.

The Mnemonic is "Computer Viruses Destroy Or Terminate."

Centre: A measure of centre is a value at the centre or middle of a data set. It has 3 ways mean, median and mode.

Mean: The arithmetic mean, or the mean, of a set of data, is the measure of the centre found by adding the data values and dividing the total by the number of data values.

Mean = $\frac{\Sigma x}{n}$

Σx = denotes the sum of a set of data values
x = is the variable usually used to represent the individual data values.
n = represents the number of data values in a sample.
N = represents the number of data values in a population.

$\bar{x} = \frac{\sum x}{n}$ is the mean of a set of *sample* values.

$\mu = \frac{\sum x}{N}$ is the mean of all values in a *population*.

Eg: The Chapter Problem refers to word counts from 186 men and 210 women. Find the mean of these first five word counts from men: 27,531; 15,684; 5,638; 27,997; and 25,433.

The mean is computed by using Formula 3-1. First, add the data values, then divide by the number of data values:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{27{,}531 + 15{,}684 + 5{,}638 + 27{,}997 + 25{,}433}{5} = \frac{102{,}283}{5}$$
$$= 20{,}456.6$$

Since $\bar{x} = 20{,}456.6$ words, the mean of the first five word counts is 20,456.6 words.

Median: The median of a data set is the measure of the centre that is the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude. The median is often denoted by $\tilde{x}$ (pronounced "x-tilde").

To find the median, first sort the values (arrange them in order), then follow one of these two procedures:

1. If the number of data values is odd, the median is the number located in the exact middle of the list.

2. If the number of data values is even, the median is found by computing the mean of the two middle numbers.

Eg: 1. Find the median for this sample of data values used in Example 1: 27,531, 15,684, 5,638, 27,997, and 25,433.

First sort the data values, as shown below:

5,638 15,684 25,433 27,531 27,997

Because the number of data values is an odd number (5), the median is the number located in the exact middle of the sorted list, which is 25,433. The median is therefore 25,433 words.

2. Median Repeat Example 2 after including the additional data value of 8,077 words. That is, find the median of these word counts: 27,531, 15,684, 5,638, 27,997, 25,433, and 8,077.

First, arrange the values in order:

5,638 8,077 15,684 25,433 27,531 27,997

Because the number of data values is an even number (6), the median is found by computing the mean of the two middle numbers, which are 15,684 and 25,433.

$$\text{Median} = \frac{15{,}684 + 25{,}433}{2} = \frac{41{,}117}{2} = 20{,}558.5$$

The median is 20,558 words.

Mode: **The mode of a data set is the value that occurs with the greatest frequency.**

A data set can have one mode, more than one mode, or no mode.

- **When two data values occur with the same greatest frequency, each one is a mode and the data set is bimodal.**
- **When more than two data values occur with the same greatest frequency, each is a mode and the data set is said to be multimodal.**
- **When no data value is repeated, we say that there is no mode.**

Eg: Find the mode of these word counts: 18,360 18,360 27,531 15,684 5,638 27,997 25,433.

The mode is 18,360 words because it is the data value with the greatest frequency.

In Example, the mode is a single value. Here are two other possible circumstances:

Two modes: The values of 0, 0, 0, 1, 1, 2, 3, 5, 5, 5 have two modes: 0 and 5.

No mode: The values of 0, 1, 2, 3 and 5 have no mode because no value occurs more than once.

Midrange: The midrange of a data set is the measure of centre that is the value midway between the maximum and minimum values in the original data set. It is found by adding the maximum data value to the minimum data value and then dividing the sum by 2, as in the following formula:

$$\text{midrange} = \frac{\text{maximum data value} + \text{minimum data value}}{2}$$

Eg: Find the midrange of these values from Example 1: 27,531, 15,684, 5,638, 27,997, and 25,433.

The midrange is found as follows:

$$\text{midrange} = \frac{\text{maximum data value} + \text{minimum data value}}{2}$$

$$= \frac{27{,}997 + 5{,}638}{2} = 16{,}817.5$$

The midrange is 16,817.5 words.

Round-Off Rule for the Mean, Median, and Midrange:

1. **Carry one more decimal place than is present in the original set of values.**

2. **(Because values of the mode are the same as some of the original data values, they can be left as is without any rounding.)**

Mean from a Frequency Distribution:

mean from frequency distribution: $\bar{x} = \dfrac{\Sigma(f \cdot x)}{\Sigma f}$

First multiply each frequency and class midpoint, then add the products.

Σf: sum of frequencies

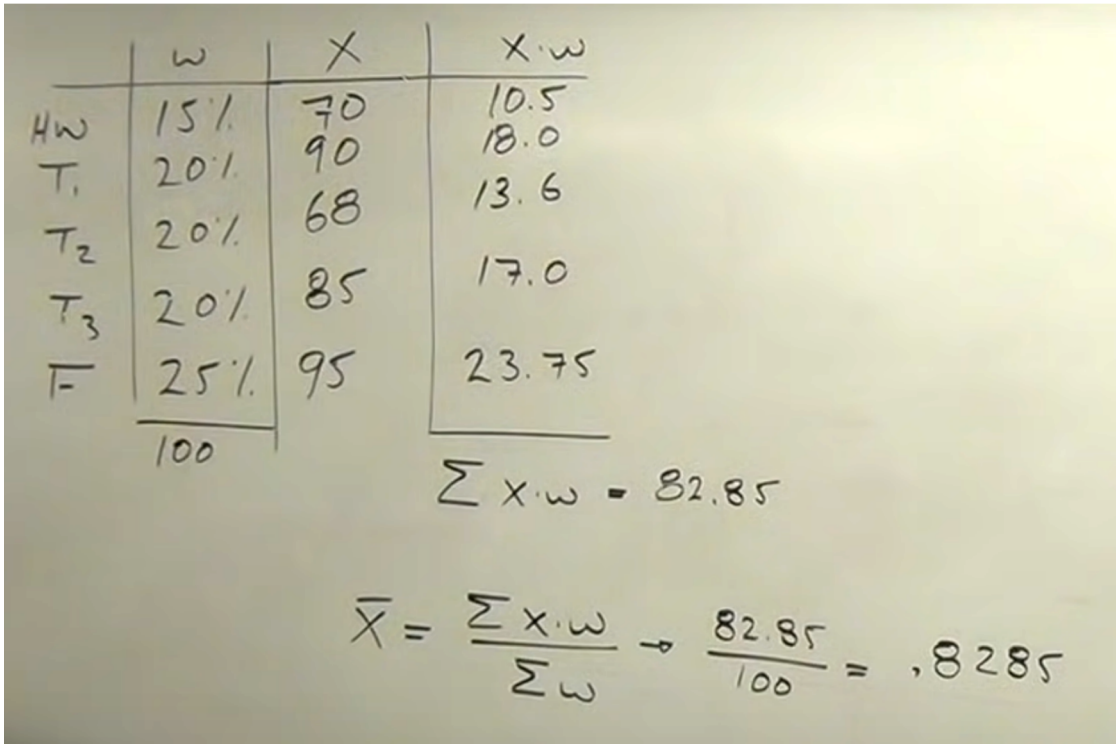
Table 3-1 Finding the Mean from a Frequency Distribution

Word Counts from Men	Frequency f	Class Midpoint x	$f \cdot x$
0–9,999	46	4,999.5	229,977.0
10,000–19,999	90	14,999.5	1,349,955.0
20,000–29,999	40	24,999.5	999,980.0
30,000–39,999	7	34,999.5	244,996.5
40,000–49,999	3	44,999.5	134.998.5
Totals:	$\Sigma f = 186$		$\Sigma(f \cdot x) = 2{,}959{,}907$
			$\bar{x} = \dfrac{\Sigma(f \cdot x)}{\Sigma f} = \dfrac{2{,}959{,}907}{186} = 15{,}913.5$

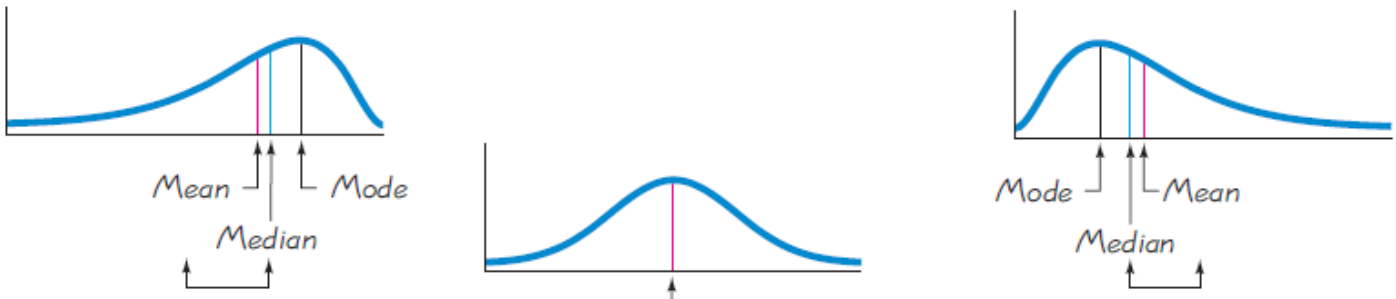
Weighted Mean:

weighted mean: $\bar{x} = \frac{\Sigma(w \cdot x)}{\Sigma w}$

W = Weight, Hw = Homework, T = Tests, F = Final



Skewness: A distribution of data is skewed if it is not symmetric and extends more to one side than to the other. (A distribution of data is symmetric if the left half of its histogram is roughly a mirror image of its right half.)



(a) Skewed to the Left (Negatively Skewed): The mean and median are to the *left* of the mode (but their order is not always predictable).

(b) Symmetric (Zero Skewness): The mean, median, and mode are the same.

(c) Skewed to the Right (Positively Skewed): The mean and median are to the *right* of the mode (but their order is not always predictable).

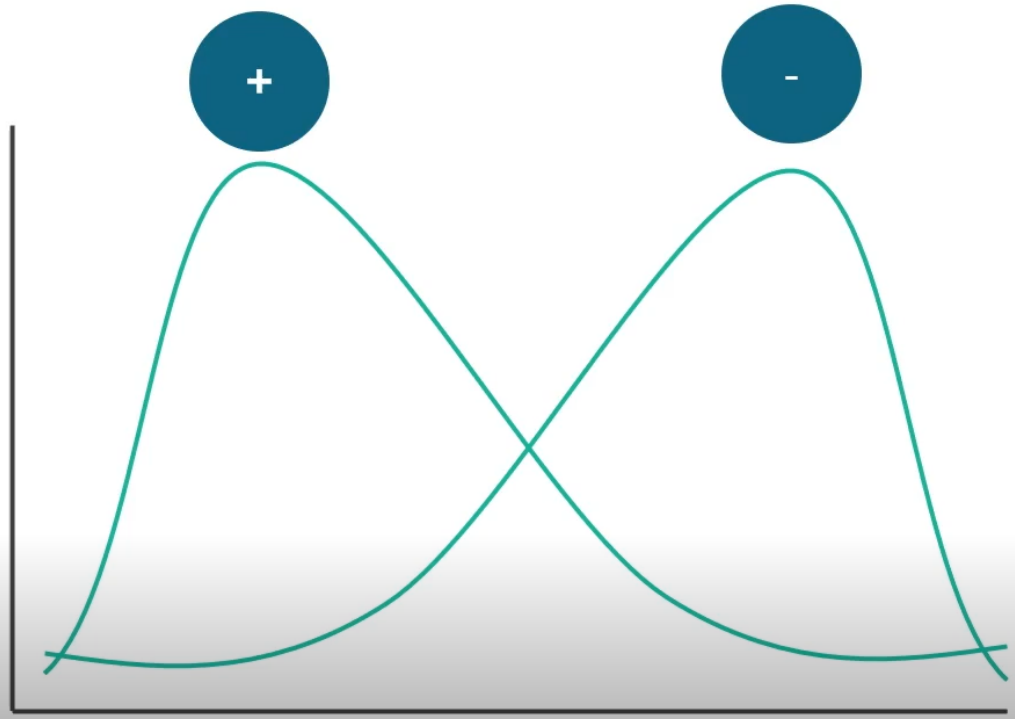
Figure 3-2 Skewness

Bowley's Coefficient of Skewness

= (Mean –Mode)/ Standard Deviation

= 3 * (Mean –Median)/ Standard Deviation

Based on negative or positive deviation, the tail will take a left or right direction.



Measures of Variation: How the Data is Spread

Ways to measure Variation:

1. **Range: The range of a set of data values is the difference between the maximum data value and the minimum data value.**

   Range = (maximum data value) - (minimum data value)

   Because the range uses only the maximum and the minimum data values, it is very sensitive to extreme values and isn't as useful as other measures of variation that use every data value, such as the standard deviation. However, because the range is so easy to compute and understand, it is used often in statistical process control.

2. **Standard Deviation: A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.**

   Or

   Measures the average distance your data values are from the mean.

   1. Never negative and Never Zero unless all entries are the same.
   2. Greatly affected by outliers

   Sample standard deviation is denoted by 's'.

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}} \quad \textbf{or} \quad s = \sqrt{\frac{n\Sigma (x^2) - (\Sigma x)^2}{n(n-1)}}$$

Problem: India has 1 satellite used for military and intelligence purposes, Japan has 3, and Russia has 14. Find the Standard Deviation of the sample values of 1, 3, and 14.

X	$x - \bar{x}$	$(x - \bar{x})^2$
1	1 - 6 = -5	25
3	3 - 6 = -3	9
14	14 - 6 = 8	64
N = 3		$\sum(x - \bar{x})^2$ = 98

Mean ($\bar{x}$) = 1 + 3 + 14 / 3 = 6

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

s = $\sqrt{\frac{98}{3-1}} = \sqrt{\frac{98}{2}} = \sqrt{49} = 7$

Other formula:

x	x^2
1	1
3	9
14	196
N = 3 X = 18	Σx^2 = 206

$$s = \sqrt{\frac{n\Sigma (x^2) - (\Sigma x)^2}{n(n-1)}} = \sqrt{\frac{3 * 206 - (18)^2}{3(3-1)}} = \sqrt{\frac{618 - 324}{6}} = \sqrt{\frac{324}{6}} = \sqrt{49} = 7$$

SOLUTION Shown below is the computation of the standard deviation of 1 satellite, 3 satellites, and 14 satellites using Formula 3-5.

$n = 3$ (because there are 3 values in the sample)

$\Sigma x = 18$ (found by adding the sample values: $1 + 3 + 14 = 18$)

$\Sigma x^2 = 206$ (found by adding the squares of the sample values, as in $1^2 + 3^2 + 14^2 = 206$)

Using Formula 3-5, we get

$$s = \sqrt{\frac{n(\Sigma x^2) - (\Sigma x)^2}{n(n-1)}} = \sqrt{\frac{3(206) - (18)^2}{3(3-1)}} = \sqrt{\frac{294}{6}} = 7.0 \text{ satellites}$$

Standard Deviation of a Population:

population standard deviation $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

Variance: The variance of a set of values is a measure of variation equal to the square of the standard deviation.

Sample variance: s^2 square of the standard deviation (s).

Population Variance: σ^2 square of the population standard deviation (σ).

s = *sample* standard deviation

s^2 = *sample* variance

σ = *population* standard deviation

σ^2 = *population* variance

- **Closely grouped data will have A small Standard Deviation**
- **Spread-out data will have a large Standard Deviation**

If a data set is Normally Distributed, we can use the empirical value.

- **About 68% of all values fall within 1 standard deviation of the mean.**
- **About 95% of all values fall within 2 standard deviations of the mean.**
- **About 99.7% of all values fall within 3 standard deviations of the mean.**
- **If a data value lies within 2 standard deviations of the mean it is considered as usual.**
- **A data value outside 3 standard deviations from the mean is very rare.**

Problem: Empirical Rule IQ scores have a bell-shaped distribution with a mean of 100 and a standard deviation of 15. What percentage of IQ scores are between 70 and 130?

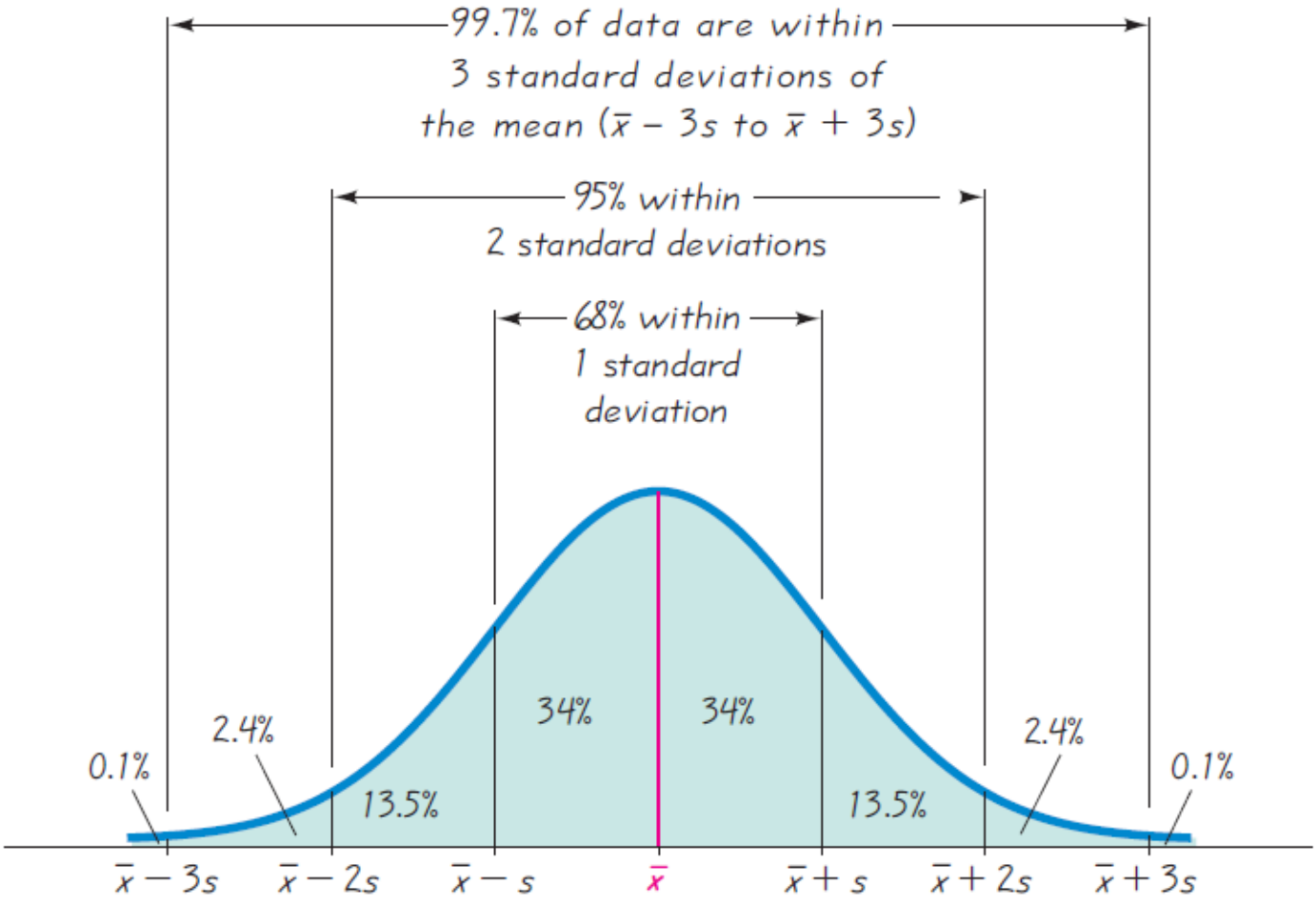
Solution:
The key to solving this problem is to recognize that 70 and 130 are each exactly 2 standard deviations away from the mean of 100, as shown below.

2 standard deviations = 2s = 2(15) = 30

Therefore, 2 standard deviations from the mean is

100 - 30 = 70
or 100 + 30 = 130

The empirical rule tells us that about 95% of all values are within 2 standard deviations of the mean, so about 95% of all IQ scores are between 70 and 130.



Coefficient of variation:

When comparing variation in two different sets of data, the standard deviations should be compared only if the two sets of data use the same scale and units and they have approximately the same mean. If the means are substantially different, or if the samples use different scales or measurement units, we can use the coefficient of variation, defined as follows.

The coefficient of variation (or CV) for a set of the nonnegative sample or population data, expressed as a per cent describes the standard deviation relative to the mean, and is given by the following:

Sample	Population
$CV = \frac{s}{\bar{x}} \cdot 100\%$	$CV = \frac{\sigma}{\mu} \cdot 100\%$

Problem - Heights and Weights of Men: Compare the variation in heights of men to the variation in weights of men, using these sample results obtained from Data Set 1 in Appendix B: for men, the heights yield $\bar{x}$ = 68.34 in. and s = 3.02 in; the weights yield $\bar{x}$ = 172.55 lb and s = 26.33 lb. Note that we want to compare variation among heights to variation among weights.

Solution: We can compare the standard deviations if the same scales and units are used and the two means are approximately equal, but here we have different scales (heights and weights) and different units of measurement (inches and pounds), so we use the coefficients of variation:

$$\text{heights:} \quad CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{3.02 \text{ in.}}{68.34 \text{ in.}} \cdot 100\% = 4.42\%$$

$$\text{weights:} \quad CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{26.33 \text{ lb}}{172.55 \text{ lb}} \cdot 100\% = 15.26\%$$

Although the standard deviation of 3.02 in. cannot be compared to the standard deviation of 26.33 lb, we can compare the coefficients of variation, which have no units. We can see that heights (with CV = 4.42%) have considerably less variation than weights (with CV = 15.26%).

Measures of Relative Standing and Boxplots -

Part 1: Basics of z Scores, Percentiles, Quartiles, and Boxplots

z Scores: A z score (or standardized value) is the number of standard deviations that a given value x is away from the mean. The z score is calculated by using one of the following:

SAMPLE: $z = \frac{x - \bar{x}}{s}$

POPULATION: $z = \frac{x - \mu}{\sigma}$

Allows the comparison of the variation in two different Samples/Population.

Find out who is relatively taller (Use Z score for Men's Populations)?

1. **Sachin's Height = 76 inches**

   Mean for Men's Height = 71.5 inches

   Standard Deviation of Men's Height = 2.1 inches

   Z = (76 - 71.5) / 2.0 = 2.142

2. **Dravid's Height = 86 inches**

   Mean for Cricket players Height = 80 inches

   Standard deviation of Cricket players height = 3.3 inches

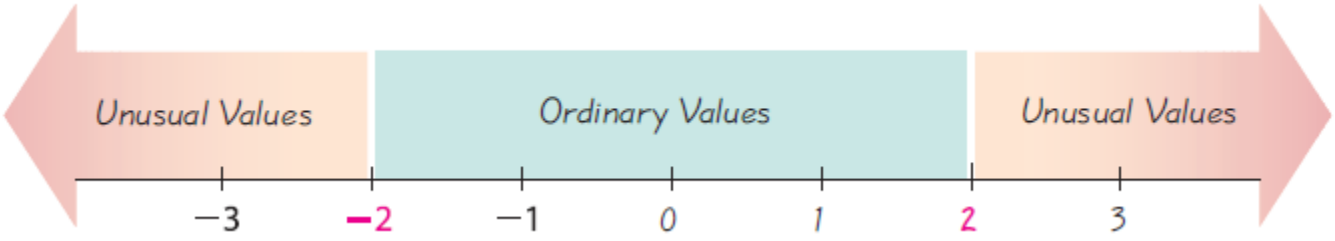
   Z = (86 - 80) / 3.3 = 1.818

Conclusion: Sachin was relatively taller than Dravid in the whole of men's height. Dravid is Taller among cricket players.

z Scores, Unusual Values, and Outliers:

Ordinary values: $-2 \leq z\text{ score} \leq 2$

Unusual values: $z\text{ score} < -2$ *or* $z\text{ score} > 2$



Quartiles: **Quartiles are measures of location, denoted Q1, Q2 and Q3, which divide a set of data into four groups with about 25% of the values in each group.**

1st Quartile Q1 = Separates the bottom 25% of the sorted values from the top 75%. Bottom 25% of sorted data.

2nd Quartile Q2 - (Median) = Same as the median; separates the bottom 50% of the sorted values from the top 50%. Bottom 50% of sorted data.

3rd Quartile Q3 = Separates the bottom 75% of the sorted values from the top 25%. Bottom 75% of sorted data.

Eg:

1. 1,3,6,10,15,21,28,36

Q 2 = Median = 12.5

Q1 = 1,3,6,10 = 3 + 6 / 2
= 4.5

Q3 = 15,21,28,36 = 28 + 21 / 2
= 24.5

2. 1,3,6,10,15,21,28,36,39

Q2 = Median = 15

Q1 = 4.5

Q3 = 32

Percentile - **Percentiles are measures of location, denoted P1, P2, … , P99, which divides a set of data into 100 groups with about 1% of the values in each group. (Separates the data into 100 Parts)**

Percentile of x: $\frac{No\ of\ data\ value\ less\ than\ x}{Total\ number\ of\ Values} * 100$

Eg:

1. **You scored 87/100 on a test and 39 people scored lower than you And there are 54 people in the class?**

   Percentile of 87 = 39/54 * 100 = 72.22 percentile.

Inter Quartile Range (IQR): Q3 - Q1 = Middle 50% of the data.

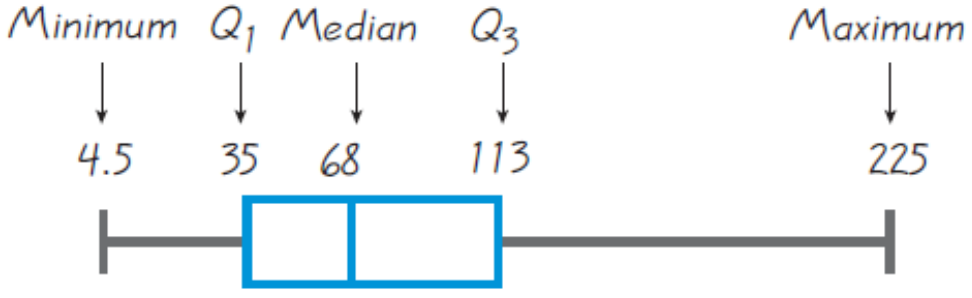
Boxplot:

For a set of data, the 5-number summary consists of the minimum value, the first quartile Q1 the median Q2(or second quartile), the third quartile Q3 and the maximum value.

A boxplot (or box-and-whisker diagram) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile Q1 the median Q2, and the third quartile Q3.

Procedure for Constructing a Boxplot:

1. Find the 5-number summary consisting of the minimum value, Q1, the median, Q3 and the maximum value.
2. Construct a scale with values that include the minimum and maximum data values.
3. Construct a box (rectangle) extending from Q1 to Q3 and draw a line in the box at the median value Q2.
4. Draw lines extending outward from the box to the minimum and maximum data values.



Eg:

1. **1, 4, 5, 5, 7, 9, 12, 13, 13, 15, 21**

   Min: 1

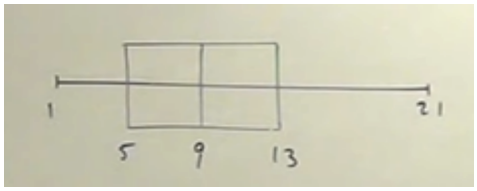
   Q1: 5

   Median(Q2) : 9

   Q3: 13

   Max: 21

   Boxplot -



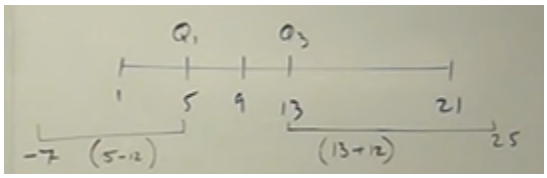
   Outlier:

   Find IQR : Q3 - Q1 = 13 - 5 = 8

   1.5 * IQR = 1.5 * 8 = 12

   Q1 - 1.5 * IQR = -7

   Q3 + 1.5 * IQR = 25.



   There is nothing less than -7 and nothing more than 25 so no outliers.

Ch. 3: Descriptive Statistics

$$\bar{x} = \frac{\Sigma x}{n} \quad \text{Mean}$$

$$\bar{x} = \frac{\Sigma f \cdot x}{\Sigma f} \quad \text{Mean (frequency table)}$$

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}} \quad \text{Standard deviation}$$

$$s = \sqrt{\frac{n(\Sigma x^2) - (\Sigma x)^2}{n(n - 1)}} \quad \text{Standard deviation (shortcut)}$$

$$s = \sqrt{\frac{n[\Sigma (f \cdot x^2)] - [\Sigma (f \cdot x)]^2}{n(n - 1)}} \quad \text{Standard deviation (frequency table)}$$

$$\text{variance} = s^2$$