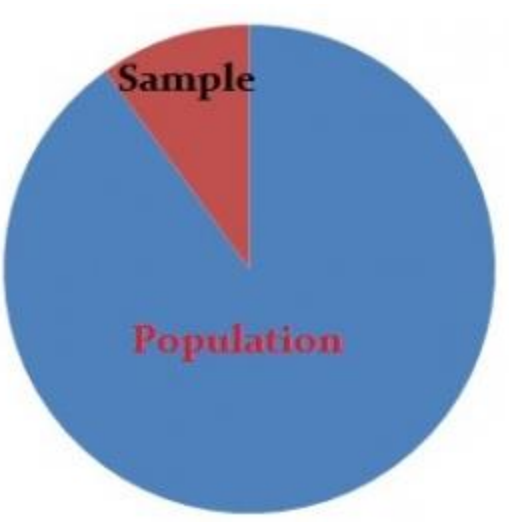
2 STATISTICS AND PROBABILITY

2.1 Discuss statistical concepts to include discrete and continuous variables, sampling, variance, skewness
2.2 Present information in a range of graphical and tabular forms, using pie charts, trend graphs, correlation diagrams (+/-), cumulative frequency curves, histograms and frequency tables with both discrete and continuous variables
2.3 Calculate the statistics for measuring and contrasting averages and dispersion of grouped data by calculating the mean, mode, median, weighted average, range, inter-quartile range and standard deviation
2.4 Calculate the number of possible outcomes on tests with no repetitions by using the Fundamental Principle of Counting, and Permutations and Combinations
2.5 Demonstrate an understanding of relative frequency and probability by using Information Technology simulations
2.6 Solve simple probability problems of one or two events including where two events are mutually exclusive and where two events are independent
2.7 Discuss findings, to include interpretation of results and distortions which may arise, and reasons for findings

Sampling

In statistics, quality assurance, and survey methodology, sampling is the selection of a subset (a statistical sample) of individuals from within a statistical population to estimate characteristics of the whole population. ... Results from probability theory and statistical theory are employed to guide the practice.



Samples are parts of a population. For example, you might have a list of information on 100 people (your “sample”) out of 10,000 people (the “population”). You can use that list to make some assumptions about the entire population’s behavior.

However, it’s not that simple. When you do stats, your sample size has to be ideal—not too large or too small. Then once you’ve decided on a sample size, you must use a sound technique to collect the sample from the population:

Probability Sampling uses randomization to select sample members. You know the probability of each potential member’s inclusion in the sample. For example, 1/100. However, it isn’t necessary for the odds to be equal. Some members might have a 1/100 chance of being chosen, others might have 1/50.

Non-probability sampling uses non-random techniques (i.e. the judgment of the researcher). You can't calculate the odds of any particular item, person or thing being included in your sample.

Common Types

The most common techniques you'll likely meet in elementary statistics or AP statistics include taking a sample with and without replacement. Specific techniques include:

- Bernoulli samples have independent Bernoulli trials on population elements. The trials decide whether the element becomes part of the sample. All population elements have an equal chance of being included in each choice of a single sample. The sample sizes in Bernoulli samples follow a binomial distribution. Poisson samples (less common): An independent Bernoulli trial decides if each population element makes it to the sample.
- Cluster sampes divide the population into groups (clusters). Then a random sample is chosen from the clusters. It's used when researchers don't know the individuals in a population but do know the population subsets or groups.
- In systematic sampling, you select sample elements from an ordered frame. A sampling frame is just a list of participants that you want to get a sample from. For example, in the equal-probability method, choose an element from a list and then choose every kth element using the equation k = N\n. Small "n" denotes the sample size and capital "N" equals the size of the population.
- SRS : Select items completely randomly, so that each element has the same probability of being chosen as any other element. Each subset of elements has the same probability of being chosen as any other subset of k elements.
- In stratified sampling, sample each subpopulation independently. First, divide the population into homogeneous (very similar) subgroups before getting the sample. Each population member only belongs to one group. Then apply simple random or a systematic method within each group to choose the sample. Stratified Randomization: a sub-type of stratified used in clinical trials. First, divide patients into strata, then randomize with permuted block randomization.

Primary data

An advantage of using primary data is that researchers are collecting information for the specific purposes of their study. In essence, the questions the researchers ask are tailored to elicit the data that will help them with their study. Researchers collect the data themselves, using surveys, interviews and direct observations.

Secondary data

There are several types of secondary data. They can include information from the national population census and other government information collected by Statistics Canada. One type of secondary data that's used increasingly is administrative data. This term refers to data that is collected routinely as part of the day-to-day operations of an organization, institution or agency. There are any number of examples: motor vehicle registrations, hospital intake and discharge records, workers' compensation claims records, and more.

Mean/Mode/Median

number of pupils (*frequency*) that received that mark.

$$\Rightarrow \text{the mean} = \frac{1\times1+2\times1+3\times1+4\times3+5\times5+6\times3+7\times2+8\times2+9\times1+10\times1}{1+1+1+3+5+3+2+2+1+1}$$

$$= \frac{110}{20} = 5{\cdot}5 \text{ marks}$$

When the frequency table is in the form

Variable	x_1	x_2	x_3	x_4	$\cdots$	x_n
Frequency	f_1	f_2	f_3	f_4	$\cdots$	f_n

then the mean, $\bar{x}$, is given by the formula

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \cdots + f_nx_n}{f_1 + f_2 + f_3 + \cdots + f_n}$$

This formula can be expressed in the more concise form $\bar{x} = \dfrac{\sum fx}{\sum f}$

Remember

The mean, $\bar{x}$, of a frequency distribution is given by the formula

$$\bar{x} = \frac{\sum fx}{\sum f}$$

where Σfx is the sum of all the variables multiplied by the corresponding frequencies and Σf is the sum of the frequencies.

Example 1.

The goals scored in a series of thirty matches were recorded as follows:
1, 3, 4, 3, 2, 1, 6, 5, 1, 2, 3, 2, 1, 2, 4, 3, 1, 4, 6, 2, 5, 2, 4, 6, 2, 1, 6, 2, 1, 5.

Make a frequency table of the data above and find the mean and mode of the distribution.

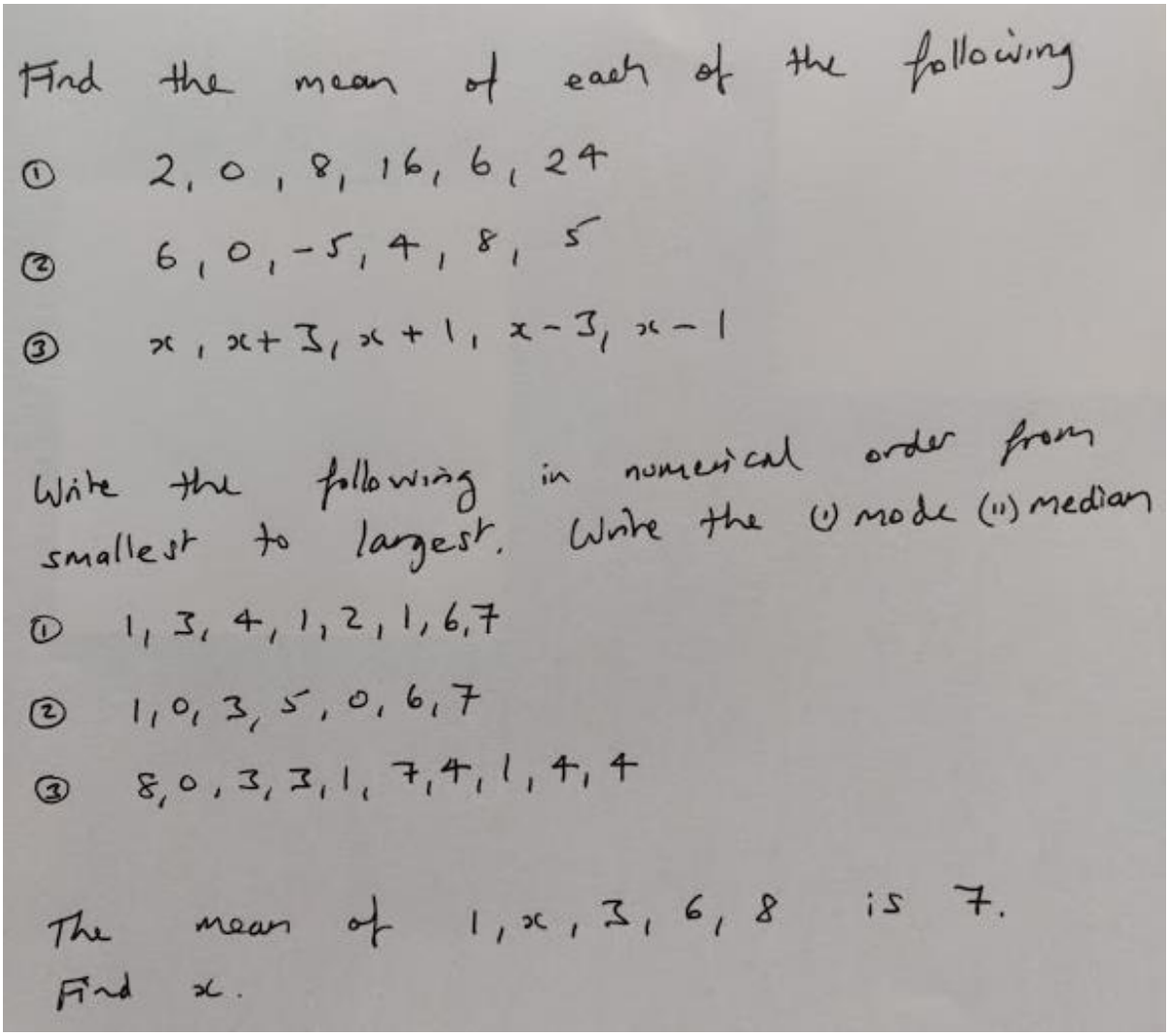
The frequency table is given below:

Goals scored	1	2	3	4	5	6
No. of matches	7	8	4	4	3	4

$$\text{Mean} = \frac{7\times1+8\times2+4\times3+4\times4+3\times5+4\times6}{7+8+4+4+3+4} = \frac{90}{30} = 3$$

$\Rightarrow$ Mean = 3

Mode = 2, since 2 occurs with the greatest frequency.



Weighted Mean

Formula

$$\bar{x}_w = \frac{\sum wx}{\sum w}$$

Example: Calculating grades

Suppose a class has quizzes, homework and three exams. The scores are weighted as follows:

Quiz	HW	Exam1	Exam2	Final
10%	10%	20%	20%	40%

Suppose your averages are as follows:

Quiz	HW	Exam1	Exam2	Final
80%	89%	79%	84%	87%

How is your grade calculated?

Multiply percentages by respective weights (written as decimals)

0.1(80)+0.1(89)+0.2(79)+0.2(84)+0.4(87)

= 84.3% final grade

Example: Price Increase

Type of Meat	Pork	Mutton	Beef	Poultry
% Price Increase	2%	8%	3%	5%

Mean Price Increase = $\frac{2+8+3+5}{4} = 4.5\%$

Type of Meat	Pork	Mutton	Beef	Poultry
% Price Increase	2%	8%	3%	5%
Weight	3	1	4	2

Weighted Mean = $\frac{6+8+12+10}{3+1+4+2} = \frac{36}{10} = 3.6\%$

Exercises

	Shoes	Hardware	Clothes	Food	Housing	Fuel
% rise in year 1	12	19	23	24	8	15
% rise in year 2	18	20	14	16	4	8
Assigned weights	1	2	4	6	4	3

The above table shows the percentage rise in prices of various items and their assigned weights. Which of the two years had the bigger weighted mean of the % price rises?

Students in an American University have to take three subjects: a major, a minor and a general. Their marks in these are weighted with weights 5, 3, 2 respectively.
Here are the marks of 2 students, Alice and Bernard:

	Major score	Minor score	General score
Alice	55	64	72
Bernard	58	60	70

(i) Which student had the higher mean score?

(ii) Which student had the higher WEIGHTED mean score?

Answers

1

% rise in Year 1	12	38	92	144	32	45	18.15
% rise in Year 2	18	40	56	96	16	24	12.5

Year 1 Weighted Mean	18.15
Year 2 Weighted Mean	12.5

Answer = Year 1 has bigger weighted mean.

2

Weights	5	3	2					
	Major	Minor	General					**Sums**
Alice	55	64	72	275	192	144	=>	**611**
Bernard	58	60	70	290	180	140	=>	**610**

Mean Scores	
Alice	63.67
Bernard	62.67

Weighted Mean Scores	
Alice	61.1
Bernard	61

I. Alice has highest mean score

II. Alice has highest weighted mean score.

2. Rewrite each of the following arrays of numbers in order of size and then write down (i) the mode, (ii) the median:

   (a) 1, 3, 4, 1, 2, 1, 6, 7 (b) 1, 0, 3, 5, 0, 6, 7

   (c) 8, 11, 2, 5, 8, 7, 8, 2, 5 (d) 8, 0, 3, 3, 1, 7, 4, 1, 4, 4

3. (i) The mean of 3, 7, 8, 10 and x is 6. Find x.

   (ii) The mean of 1, x, 3, 6 and 8 is 7. Find x.

4. The mean of four numbers is 15. If three of the numbers are 12, 10 and 13, find the fourth number.

5. The mean height of 6 men is 1·7 m and the mean height of 4 women is 1·6 m. Find

   (i) the total height of the 6 men

   (ii) the total height of the 4 women

   (iii) the mean height of the 6 men and 4 women.

6. The mean of six numbers is 12. When a seventh number is added the mean of the seven numbers is 14. Find the seventh number.

7. The following table gives the number of goals scored in 60 matches on a particular week-end:

Goals scored	1	2	3	4	5	6
No. of matches	14	16	8	8	6	8

   (i) Write down the mode of the distribution

   (ii) Calculate the mean.

8. The numbers 7, 5, 13, 5, 13, 4, 11, x, y have mean 8 and mode 5. Find the values of x and y.

9. The table below shows the number of goals scored in 100 football matches on a particular Saturday.

No. of goals scored	0	1	2	3	4	5
No. of matches	10	25	30	25	10	0

   (i) Write down the modal number of goals scored.

   (ii) Calculate the mean of the distribution.

10. If the mean of the frequency distribution below is 2, find the value of x.

Variable	0	2	3	4
Frequency				

Estimating the Mean from a Grouped Frequency Distribution

When dealing with a large number of variables, such as the ages of people in a certain district, it is often more convenient to arrange the data in **groups** or **classes**. Thus, when recording the ages of people, the results could be grouped (0–9) years, (10–19) years . . . etc.

The **grouped frequency distribution table** below shows the marks (out of 25) achieved by 50 students in a test.

Marks achieved	1–5	6–10	11–15	16–20	21–25
No. of students	11	12	15	9	3

While it is not possible to find the exact mean of a grouped frequency distribution, we can find an estimate of the mean by taking the **mid-interval value** of each class. The mid-interval value in the (1–5) class is found by adding 1 and 5 and dividing by 2, i.e. $\frac{1+5}{2} = 3$

Similarly, the mid-interval value of the (6–10) class is $\frac{6+10}{2} = 8$.

The table given above is reproduced again with the mid-interval values written in smaller print over each class interval.

Mid-interval values	3	8	13	18	23
Marks achieved	1–5	6–10	11–15	16–20	21–25
No. of students	11	12	15	9	3

$$\text{mean} = \frac{\sum fx}{\sum f}$$

$$= \frac{(11 \times 3) + (12 \times 8) + (15 \times 13) + (9 \times 18) + (3 \times 23)}{11 + 12 + 15 + 9 + 3}$$

$$= \frac{555}{50} = 11{\cdot}1$$

Test Questions 7A

1. Find the mean of each of these arrays of numbers:

(i) 2, 6, 10, 14, 18 (ii) 2, 0, 8, 16, 6, 24

(iii) 6, 0, −5, 4, 8, 5 (iv) $x, x+3, x+1, x-3, x-1$

11. The ages of children in a youth-club are given in the following table:

Ages (in years)	10–12	12–14	14–16	16–18	18–20
No. of children	12	24	18	12	4

(10–12 means greater than or equal to 10, but less than 12, etc.)

(i) What is the modal age group?

(ii) Use the mid-interval value of each class to estimate the mean of the distribution, giving your answer to the nearest half year.

12. The following table shows the weights of parcels posted by a distribution company on a particular day:

Weight (in kg)	0–2	2–4	4–6	6–8	8–10
No. of parcels	6	12	20	48	14

(i) State the modal class.

(ii) Calculate the mean of the distribution.

13. The following table gives the number of points scored by a rugby team in 40 matches:

Points scored	0–4	5–9	10–14	15–19	20–24
No. of matches	3	9	15	10	3

Estimate the mean number of points scored per match.

14. Fifty orange boxes were examined and the number of bad oranges in each box was recorded. The results are given in the following table:

No. of bad oranges per box	0–4	5–9	10–14	15–24
No. of boxes	32	8	7	3

Find the mean number of bad oranges per box.

A Cumulative Frequency Graph is a graph plotted from a cumulative frequency table. A cumulative frequency graph is also called an **ogive** or cumulative frequency curve..

Example 1
Draw a cumulative frequency graph for the frequency table below.

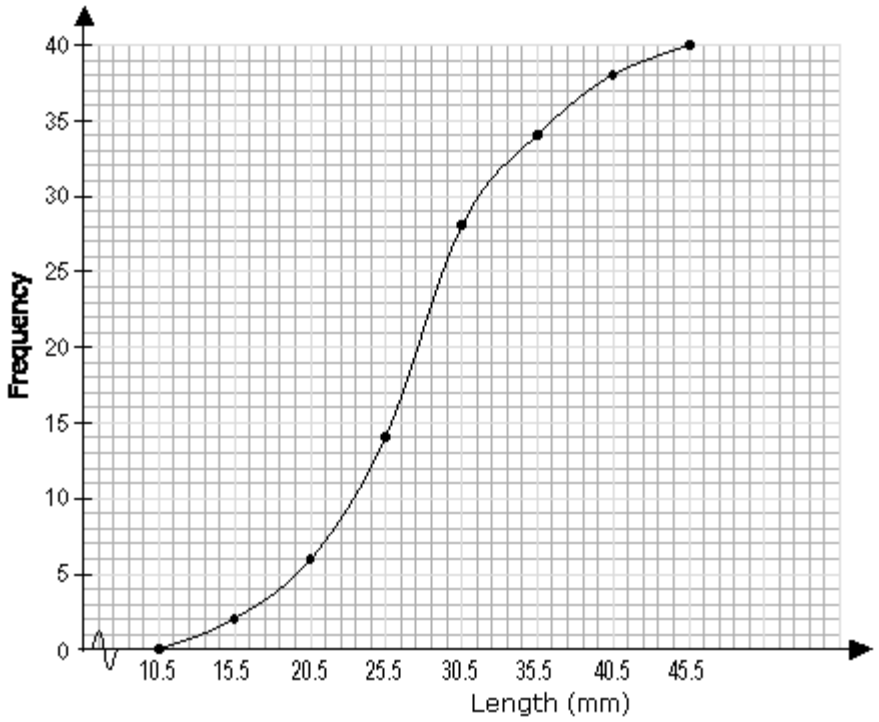
Length (*x* mm)	**Frequency**
11 – 15	2
16 – 20	4
21 – 25	8
25 – 30	14
31 – 35	6
36 – 40	4
41 – 45	2

Solution

We need to add a class with 0 frequency before the first class and then find the upper boundary for each class interval.

Length (*x* mm)	**Frequency**	**Upper Class Boundary**	**Length (*x* mm)**	**Cumulative Frequency**
6 – 10	0	10.5	$x \leq 10.5$	0
11 – 15	2	15.5	$x \leq 15.5$	2
16 – 20	4	20.5	$x \leq 20.5$	6
21 – 25	8	25.5	$x \leq 25.5$	14
25 – 30	14	30.5	$x \leq 30.5$	28
31 – 35	6	35.5	$x \leq 35.5$	34
36 – 40	4	40.5	$x \leq 40.5$	38
41 – 45	2	45.5	$x \leq 45.5$	40

And then plot the cumulative frequency against the upper class boundary of each interval and join the points with a smooth curve.



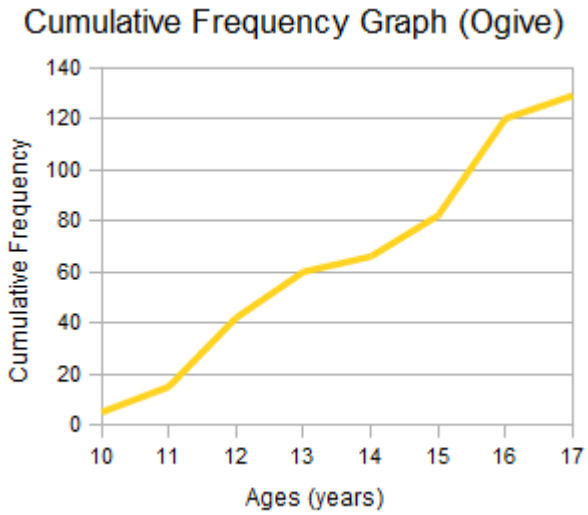
Example 2

Age (years)	Frequency
10	5
11	10
12	27
13	18
14	6
15	16
16	38
17	9

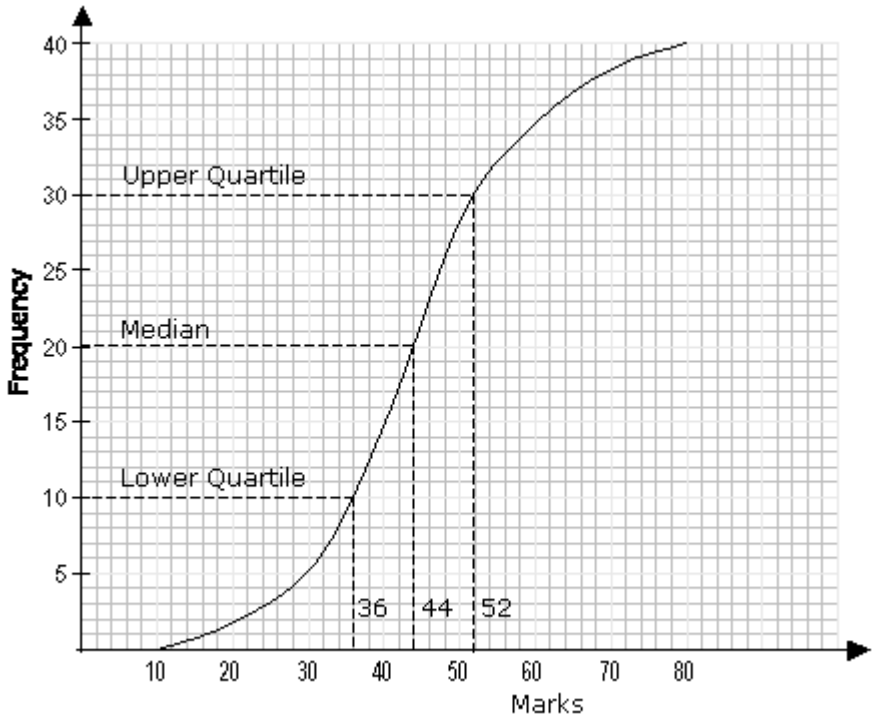
:

Age (years)	Frequency	Cumulative Frequency
10	5	5
11	10	5+10 = 15

12	**27**	**15+27 = 42**
13	**18**	**42+18 = 60**
14	**6**	**60+6 = 66**
15	**16**	**66+16 = 82**
16	**38**	**82+38 = 120**
17	**9**	**120+9 = 129**



Exam Results Example - Cumulative Frequency Curve (Ogive)



Percentile
A percentile is a certain percentage of a set of data.

Median
The median corresponds to the 50th percentile i.e. 50% of the total frequency.

$$\frac{50}{100}\times 40 = \frac{1}{2}\times 40 = 20$$

From the graph, 20 on the vertical axis corresponds to 44 on the horizontal axis. The median mark is 44.

Upper Quartile
The upper quartile corresponds to the 75th percentile i.e. 75% of the total frequency.

$$\frac{75}{100}\times 40 = \frac{3}{4}\times 40 = 30$$

From the graph, 30 on the vertical axis corresponds to 52 on the horizontal axis. The upper quartile is 52.

Lower Quartile
The lower quartile corresponds to the 25th percentile i.e. 25% of the total frequency.

$$\frac{25}{100}\times 40 = \frac{1}{4}\times 40 = 10$$

From the graph, 10 on the vertical axis corresponds to 36 on the horizontal axis. The lower quartile is 36.

Inter Quartile Range

Upper - Lower Quartile = 52 - 36 = 16

3. Teenagers at a disco were asked their ages. The results are shown in the table below:

Age (years)	13	14	15	16	17	18	19
No. of teenagers	4	9	22	19	11	9	6

Copy and complete the cumulative frequency table below.

Age years(≤)	13	14	15	16	17	18	19
No. of teenagers	4	13					

Draw a cumulative frequency curve and use it to estimate

(i) the median
(ii) the interquartile range of the ages of the teenagers.

4. The circumference of each of 150 young trees in a plantation was measured, with the following results:

Circumference (cm)	30–35	35–40	40–45	45–50	50–55	55–60
Frequency	12	25	44	48	18	3

(a) Copy and complete the following cumulative frequency table:

Circumference (≤)	35	40	45	50	55	60
Number of trees	12				147	

(b) Draw a cumulative frequency graph of the distribution.

From your graph estimate

(i) the median of the distribution (ii) the interquartile range
(iii) the number of trees with a circumference between 43 cm and 53 cm.

5. The following table gives the percentage increase in a representative sample of 120 grocery items over a four-year period.

Percentage increase	0–20	20–40	40–60	60–80	80–100
No. of items	10	22	62	20	6

Construct a cumulative frequency table and hence draw a cumulative frequency curve.

Use your curve to estimate

(i) the median percentage increase (ii) the upper quartile
(iii) the lower quartile (iv) the interquartile range
(v) the number of items which increased by more than 55%.

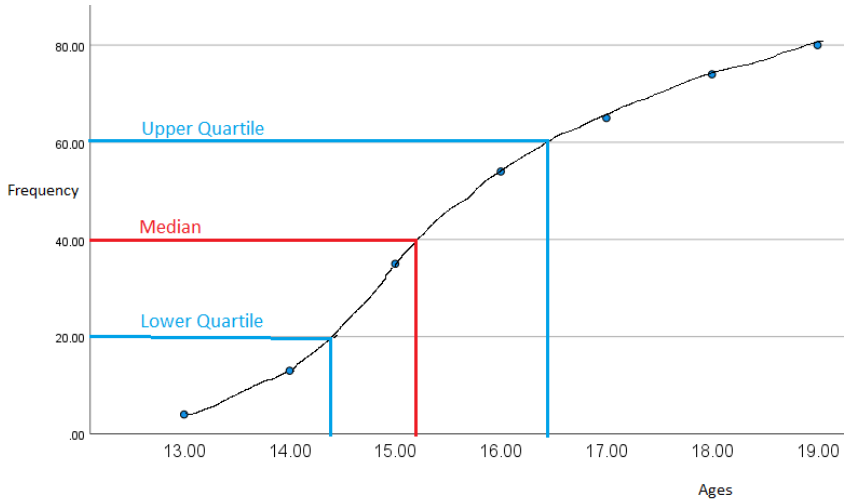
Ogive – Cumulative Frequency Curve

Question 3

Age (x)	13	14	15	16	17	18	19
No. of Teenagers (f)	4	9	22	19	11	9	6

Age	≤13	≤14	≤15	≤16	≤17	≤18	≤19
No. of Teenagers	4	13	35	54	65	74	80

Cumulative Frequency Curve (Ogive)



i. **Median** (middle value) = 15.2 years old

ii. Interquartile Range

$$Interquartile\ Range = Upper\ Qurtile - Lower\ Quartile$$
$$16.4 - 14.4 = 2\ years$$

Mean *(arithmetic average)*

x	52	126	330	304	187	162	114
f	4	9	22	19	11	9	6

sum of fx = 1275

sum of f = 80

Mean = 15.9375

Mode *(most common value or the value with the biggest frequency)*

Mode = 15 years

Range *(highest x value – lowest x value)*

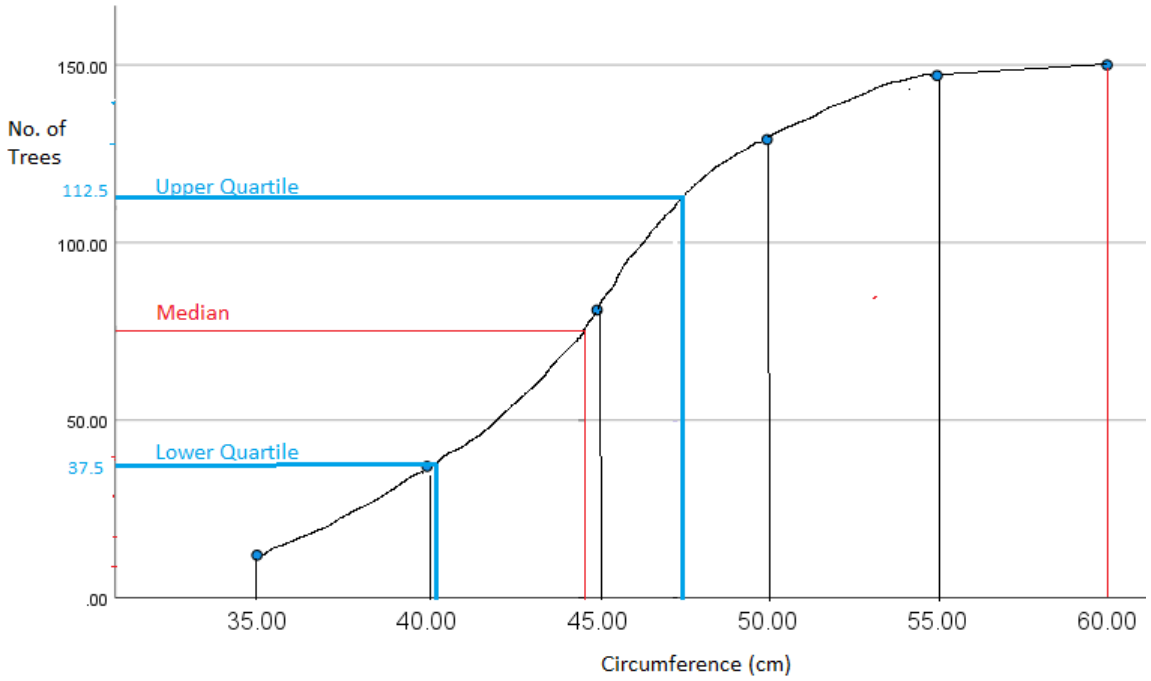
19 – 13 = 6

Range = 6 years

Question 4

Circumference (cm)	30_35	35_40	40_45	45_50	50_55	55_60
Frequency	12	25	44	48	18	3

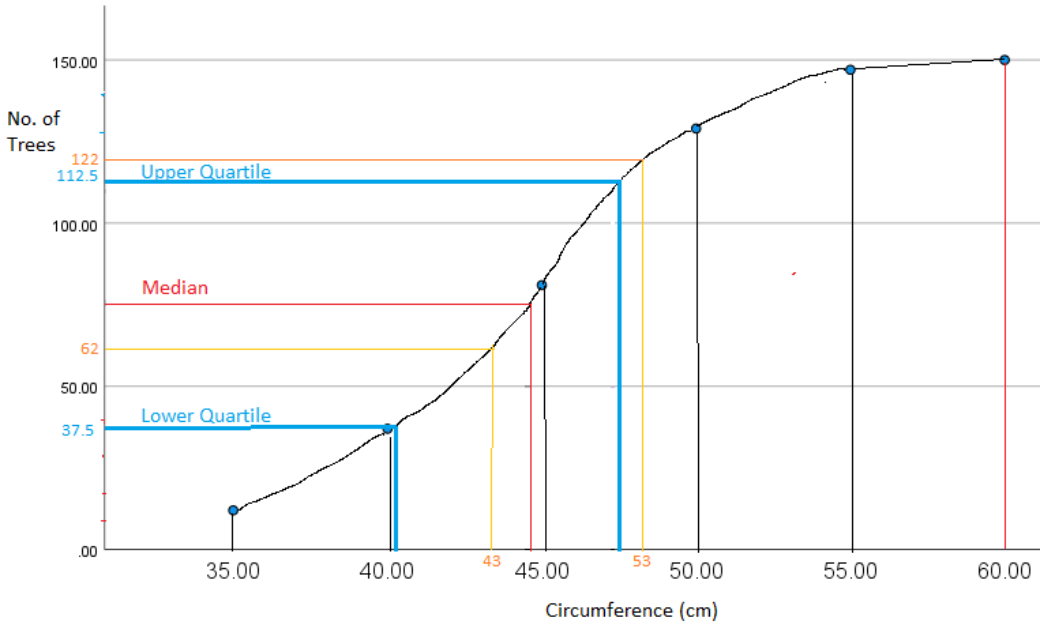
Circumference (cm)	≤35	≤40	≤45	≤50	≤55	≤60
Frequency	12	37	81	129	147	150



i. **Median** (middle value of 75) = 44 cm

ii. **Interquartile Range** $Interquartile\ Range = Upper\ Qurtile - Lower\ Quartile$

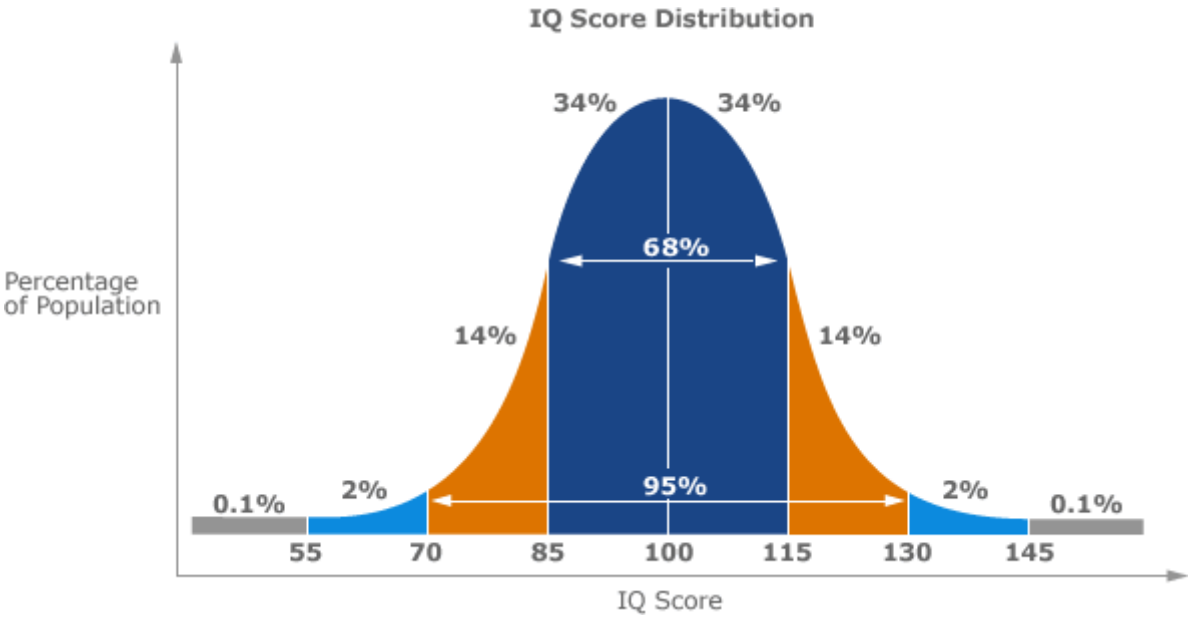
$47.5 - 41 = 6.5\ cm$



iii. 122 – 62 = 60 trees

Statistics Basics

The most common basic statistics terms you'll come across are the mean, mode and median. These are all what are known as "Measures of Central Tendency." Also important in this early chapter of statistics is the shape of a distribution. This tells us something about how data is spread out around the mean or median. Perhaps the most common distribution you'll see is the normal distribution, sometimes called a bell curve. Heights, weights, and many other things found in nature tend to be shaped like this:

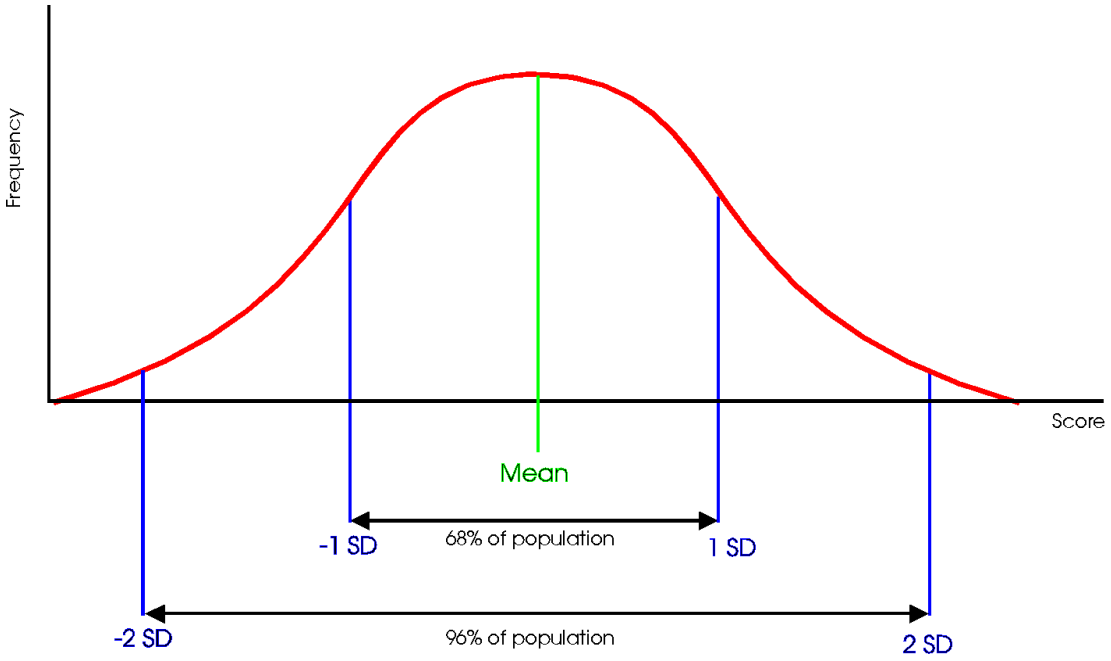


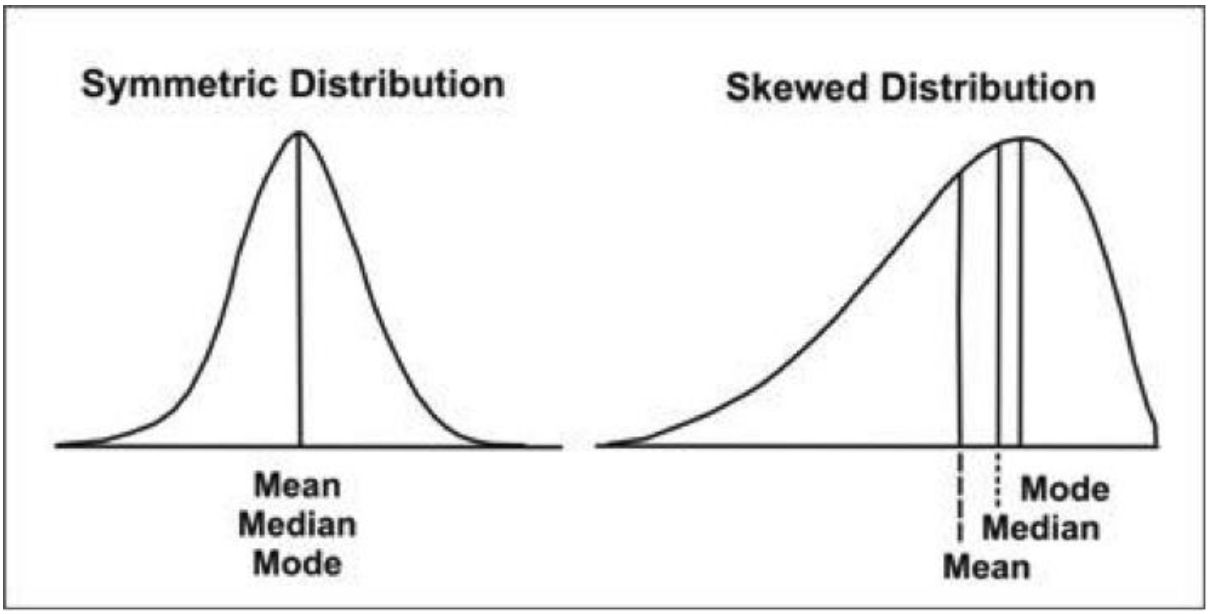
Example of a Normal Distribution

A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.

The Normal Distribution Curve





Measure of Central Tendency

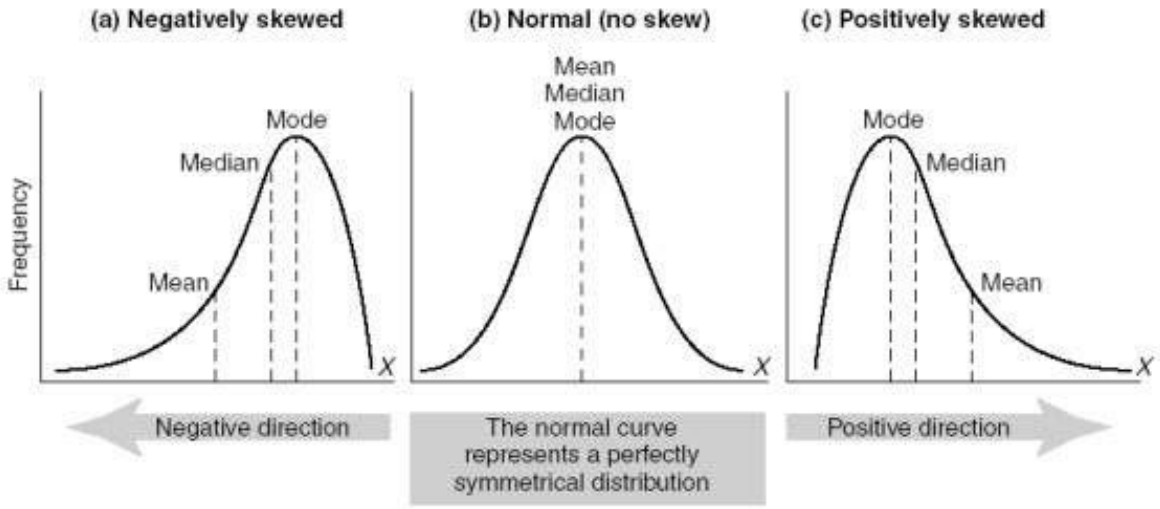
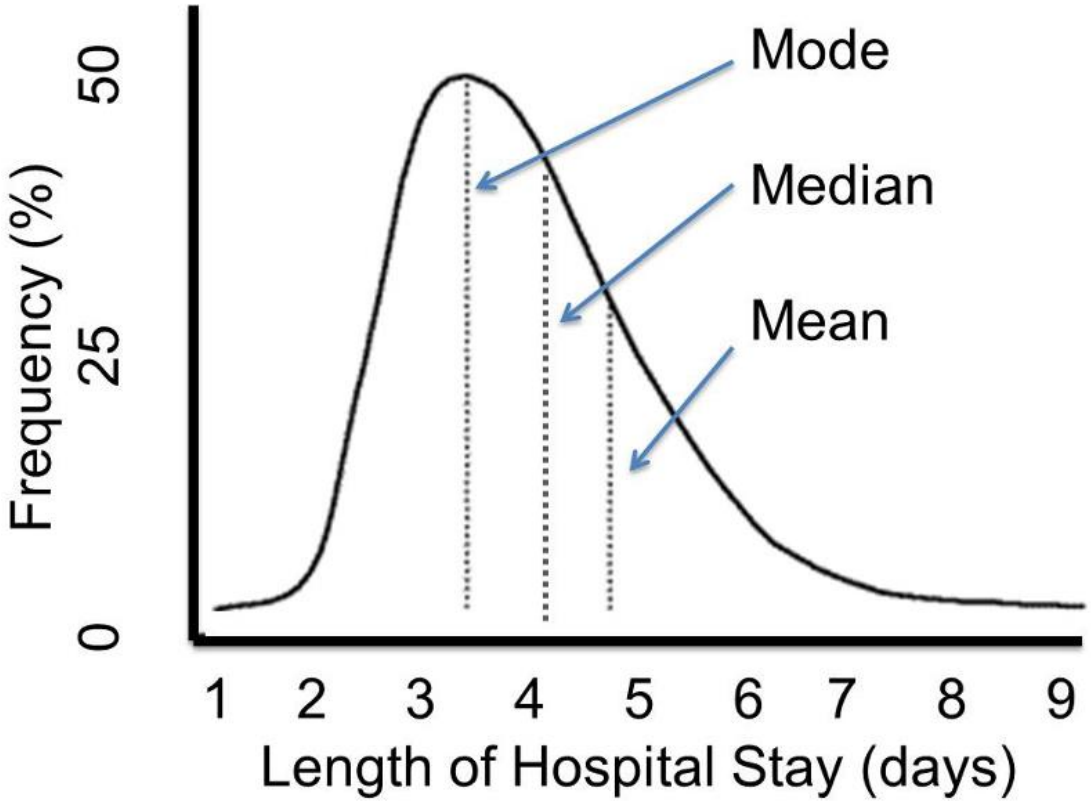


FIGURE 15.6 Examples of normal and skewed distributions



Mean, median, Mode and Range

Mean, median, and Mode are three kinds of "**averages**". The "**mean**" is the "**average**" you're used to, where you add up all the numbers and then divide by the number of numbers. The "**median**" is the "**middle**" value in the list of numbers. To find the median, your numbers have to be listed in numerical order from smallest to largest, so you may have to rewrite your list before you can find the median. The "**mode**" is the value that occurs most often. If no number in the list is repeated, then there is no mode for the list. The "**range**" of a list a numbers is just the difference between the largest and smallest values.

Exercise

Find the mean, median, mode, and range for the following list of values:
13, 18, 13, 14, 13, 16, 14, 21, 13

The mean is the usual average, so I'll add and then divide:
(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) ÷ 9 = 15

Note that the mean, in this case, isn't a value from the original list. This is a common result. You should not assume that your mean will be one of your original numbers.

The median is the middle value, so first I'll have to rewrite the list in numerical order:
13, 13, 13, 13, 14, 14, 16, 18, 21

There are nine numbers in the list, so the middle one will be the (9 + 1) ÷ 2 = 10 ÷ 2 = 5th number:
13, 13, 13, 13, 14, 14, 16, 18, 21

So the median is 14.
The mode is the number that is repeated more often than any other, so 13 is the mode.
The largest value in the list is 21, and the smallest is 13, so the range is 21 – 13 = 8.

Mean: 15
Median: 14
Mode: 13
Range: 8

What is an Average?

The word “average” is used in everyday life to describe where the middle number of a data set is. It’s the typical number you would expect to find in a series of numbers. In statistics, the average is called the “arithmetic mean,” usually just shortened to the mean. Both the average and the mean use the same formula:

Examples

Example 1: You earned €129, €139, €155 and €176 over the last 4 weeks. What is your average pay?

Step 1: Add up all of the numbers in the set. €129 + €139 + €155 + €176 = €599.

Step 2: Divide Step 1 by the total number of items in the set. There are 4 items in the set, so €599 / 4 = €149.75.

Example 2: You have semester grades of B, C, D, A, B and B. What is your average grade?

Step 1: Add up all of the numbers in the set. You have grades here, so you need to convert them on a 4.0

scale:
B = 3.0
C = 2.0
D = 1.0
A = 4.0
B = 3.0
B = 3.0
So we have: 3.0 + 2.0 + 1.0 + 4.0 + 3.0 + 3.0 = 16.0.

Step 2: Divide Step 1 by the total number of items in the set. There are 6 items in the set, so 16.0/6 = 2.66.

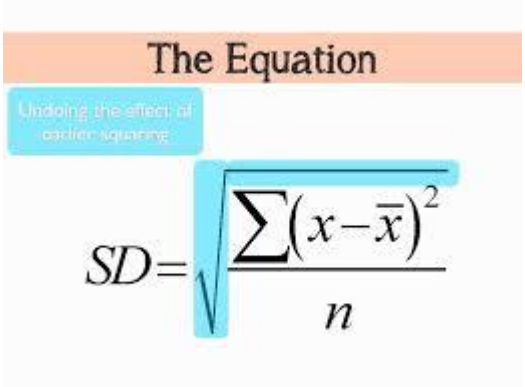
Bias in Statistics

Bias is the tendency of a statistic to overestimate or underestimate a parameter. To understand the difference between a statistic and a parameter, see this article. Bias can seep into your results for a slew of reasons including sampling or measurement errors, or unrepresentative samples.

Sampling error is the tendency for a statistic not to exactly match the population. Error doesn't necessarily mean that a mistake was made in your sampling; Sampling Variability could be a more accurate name. For example, let's say you have a population in the United States with an average height of 5 feet 9 inches. If you take a sample, even a fairly sizable sample of say, 10,000 people, it's unlikely that you'll get exactly 5 feet 9 inches. You might get very close, perhaps to within a fraction of an inch. If you repeat the experiment, you might get another very close result. For example, in experiment 1 you might get 5 feet 8.9 inches and in experiment 2 you might get 5 feet 9.1 inches. The tendency for statistics to get very close, but not exactly right, is called sampling error. Note: If the statistic is unbiased, the average of all statistics from all samples will average the true population parameter.

Standard Deviation

Standard Deviation a quantity expressing by how much the members of a group differ from the mean value for the group.



Formula for calculating Standard Deviation from a grouped frequency table.

$$\sigma = \sqrt{\frac{\sum f \cdot (x_i - \mu)^2}{\sum f}}$$

Variance

The sample variance, s2, is used to calculate how varied a sample is. A sample is a select number of items taken from a population. For example, if you are measuring American people's weights, it wouldn't be feasible (from either a time or a monetary standpoint) for you to measure the weights of every person in the population. The solution is to take a sample of the population, say 1000 people, and use that sample size to estimate the actual weights of the whole population. The variance helps you to figure out how spread out your weights are.

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

Grouped frequency table

$$S = \sqrt{\frac{\sum (x - \overline{x})^2 f}{n}}$$

$$= \sqrt{\frac{152}{30}}$$

$$= 2.25$$

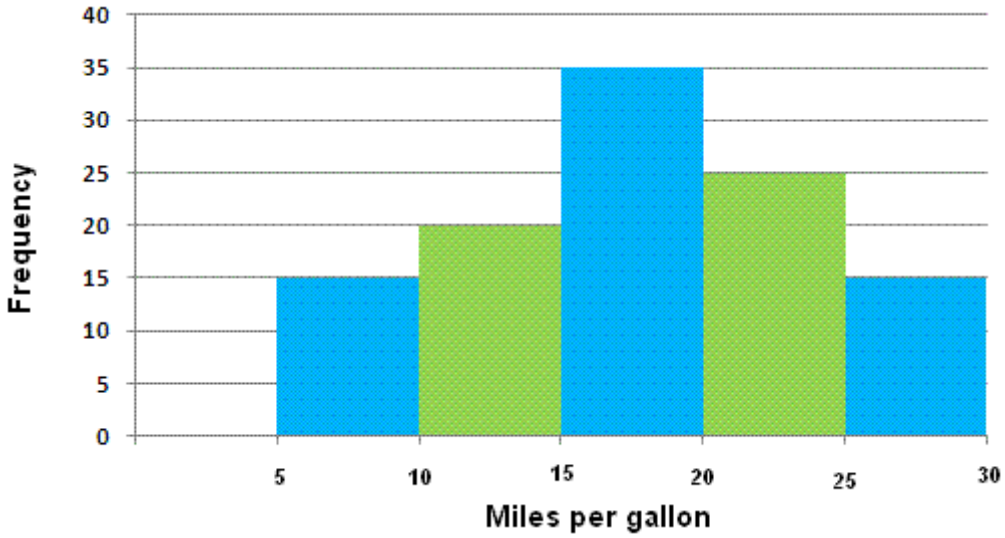
Histogram Exercise (Equal Widths)

The histogram below shows the efficiency level (in miles per gallons) of 110 cars.

a) How many cars have have an efficiency between 15 and 20 miles per gallon?

b) How many cars have have an efficiency more than 20 miles per gallon?

c) What percentage of cars have have an efficiency less than 20 miles per gallon?



Solutions

a) 35 cars

b) 25 + 15 = 40 cars

c) (15 + 20 + 35) / 110 = 0.636 = 63.6%

Histogram with Different Widths Example

When constructing a histogram with non-uniform (unequal) class widths, we must ensure that the areas of the rectangles are proportional to the class frequencies.

Remember that the histogram differs from a bar chart in that it is the area of the bar that denotes the value, not the height. This means that we would need to consider the widths in order to determine the height of each rectangle.

Example

The following frequency distribution gives the masses of 48 objects measured to the nearest gram. Draw a histogram to illustrate the data.

Mass (g)	10 – 19	20 – 24	25 – 34	35 – 50	51 – 55
Frequency	6	4	12	18	8

Solution:

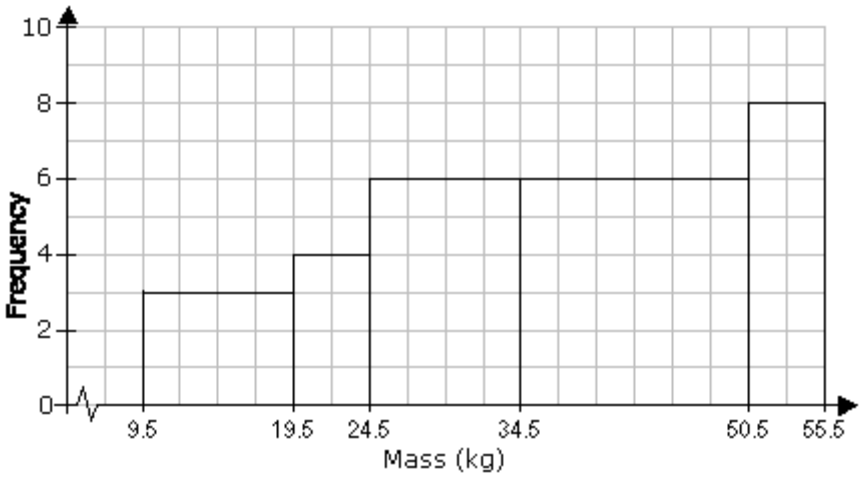
Evaluate each class widths.

Mass (g)	10 – 19	20 – 24	25 – 34	35 – 50	51 – 55
Frequency	6	4	12	18	8
Class width	10	5	10	15	5

Since the class widths are not equal, we choose a convenient width as a standard and adjust the heights of the rectangles accordingly. We notice that the smallest width size is 5. We can choose 5 to be the standard width. The other widths are then multiples of the standard width.

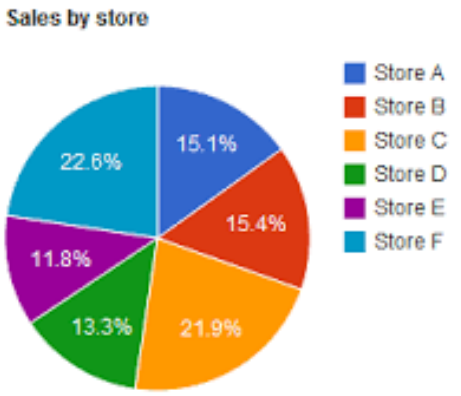
The table below shows the calculations of the heights of the rectangles.

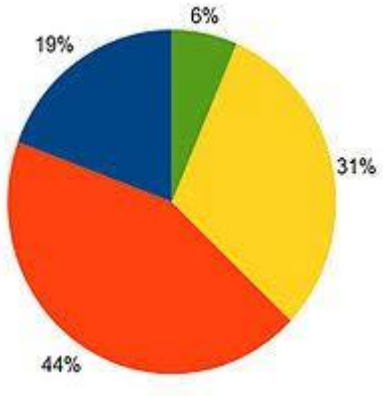
Mass (g)	10 – 19	20 – 24	25 – 34	35 – 50	51 – 55
Frequency	6	4	12	18	8
Class widths	10	5	10	15	5
	2 × standard	standard	2 × standard	3 × standard	standard
Rectangle's height in histogram	6 ÷ 2 = 3	4	12 ÷ 2 = 6	18 ÷ 3 = 6	8



Pie Chart

When we represent data in a pie chart we must remember we are using 360 degrees.



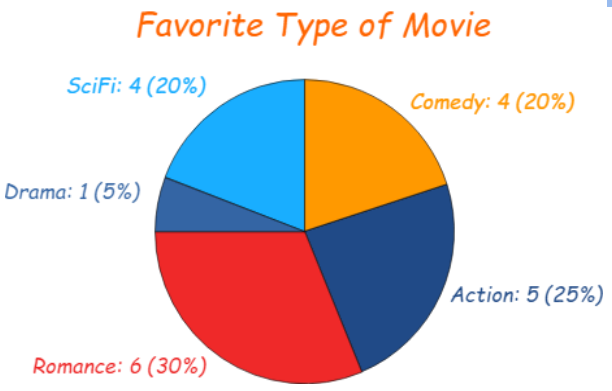


Example

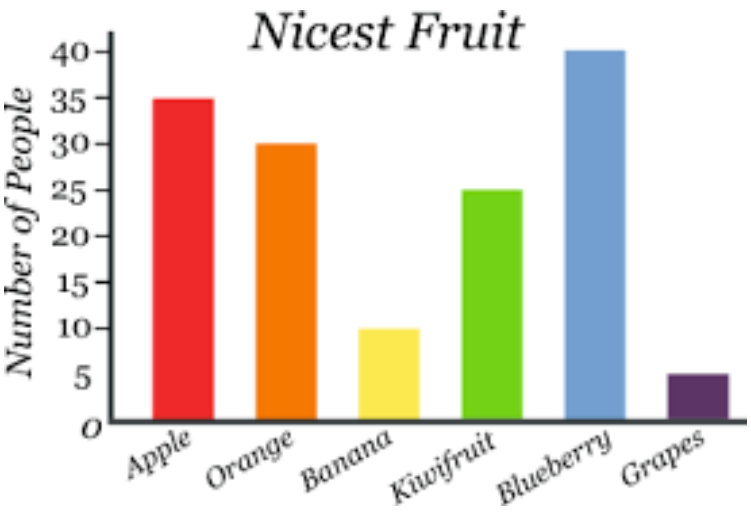
Table: Favorite Type of Movie

Comedy	Action	Romance	Drama	SciFi
4	5	6	1	4

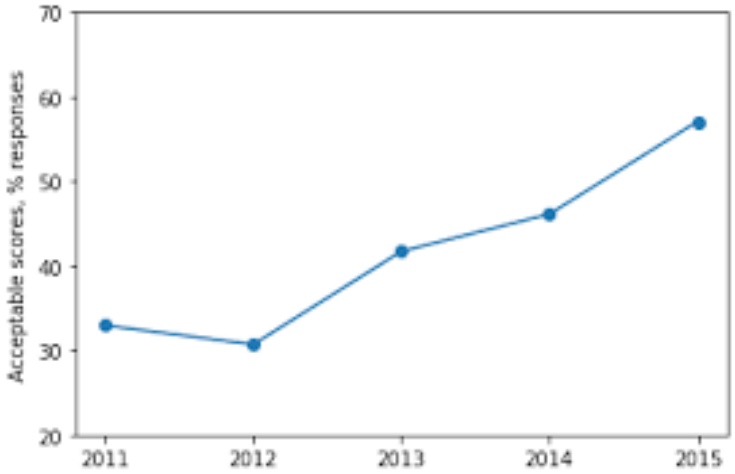
Comedy	Action	Romance	Drama	SciFi	TOTAL
4	5	6	1	4	20
20%	25%	30%	5%	20%	100%
4/20 × 360° = 72°	5/20 × 360° = 90°	6/20 × 360° = 108°	1/20 × 360° = 18°	4/20 × 360° = 72°	360°



Bar Chart



Trend Graph



Exercise

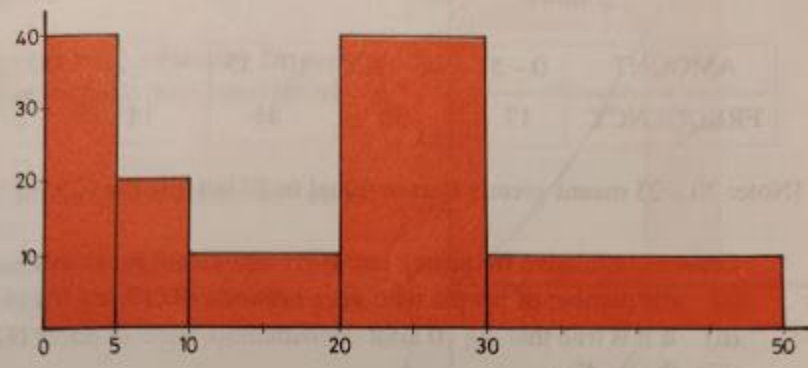
Represent the following data of student grades in a pie chart, bar chart and trend graph.

A	B	C	D
4	12	10	2

Histograms

Question

The age-distribution of the people on our road is shown on this histogram:



If there are 40 people in the 20-30 year age-group, find

(i) the number of people in the 0-5 year age-group

(ii) the total number of people on our road

(iii) the mean age (using mid-interval values).

Answer

(i) 20

(ii) 100

Mid-interval value	2.5	7.5	15	25	40	
x	0_5	5_10	10_20	20_30	30_50	
f	20	10	10	40	20	
	50	75	150	1000	800	2075
	20	10	10	40	20	100
					Mean =	20.75

(iii) 20.75

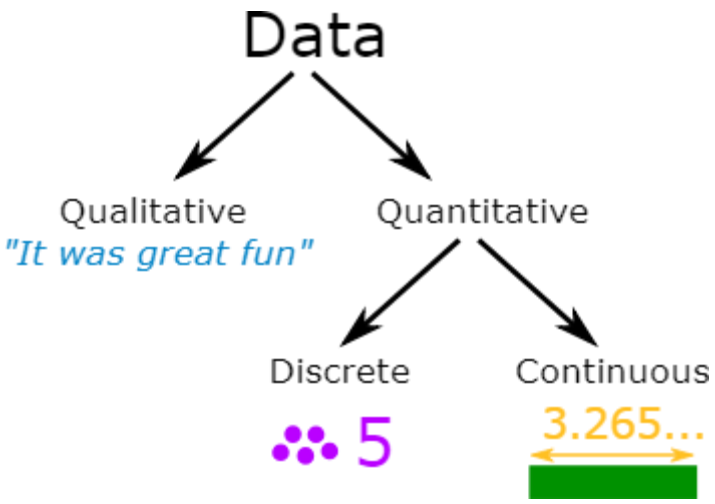
Data
Data is a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things. Data can be Descriptive (like "high" or "fast") or Numerical (numbers). Numerical Data can be Discrete or Continuous:

Population - Sample Population
In stats, a sample is a part of a population. A population is a whole, it's every member of a group. A population is the opposite to a sample, which is a fraction or percentage of a group. Sometimes it's possible to survey every member of a group. A classic example is the Census in Ireland. Note: if you do manage to survey everyone, it actually is called a census:

Primary Data
Data observed or collected directly from first-hand experience. A survey or questionnaire you might organise.

Secondary Data
Published data and data collected in the past or by other parties is called secondary data. For example data gathered by the CSO (Central Statistics Office - www.cso.ie)

Qualitative vs **Quantitative**
Data can be qualitative or quantitative. Qualitative data is descriptive information (it describes something). Quantitative data is numerical information (numbers). Quantitative data can be Discrete or Continuous:



Discrete Data
Discrete Data can only take certain values. Discrete data is counted. Discrete data can only take certain values (like whole numbers)

Examples:
- The number of students in a class. We can't have half a student!
- The result of rolling a die - Only has the values 1, 2, 3, 4, 5 or 6

Continuous Data
Continuous data is measured. Continuous data can take any value (within a range)

Examples:
- A person's height: could be any value (within the range of human heights), not just certain fixed heights

- Time in a race: you could even measure it to fractions of a second
- A dog's weight
- The length of a leaf

Continuous Variables

A variable is a quantity that has a changing value; the value can vary from one example to the next. A continuous variable is a variable that has an infinite number of possible values. In other words, any value is possible for the variable. A continuous variable is the opposite of a discrete variable, which can only take on a certain number of values. A continuous variable doesn't have to have every possible number (like -infinity to +infinity), it can also be continuous between two numbers, like 1 and 2. For example, discrete variables could be 1,2 while the continuous variables could be 1,2 and everything in between: 1.00, 1.01, 1.001, 1.0001…

Examples of continuous variables: (Heights, weights are examples of continuous variables)

- Time it takes a computer to complete a task. You might think you can count it, but time is often rounded up to convenient intervals, like seconds or milliseconds. Time is actually a continuum: it could take 1.3 seconds or it could take 1.333333333333333… seconds.
- A person's weight. Someone could weigh 180 pounds, they could weigh 180.10 pounds or they could weigh 180.1110 pounds. The number of possibilities for weight are limitless.
- Income. You might think that income is countable (because it's in dollars) but who is to say someone can't have an income of a billion dollars a year? Two billion? Fifty nine trillion? And so on…
- Age. So, you're 25 years-old. Are you sure? How about 25 years, 19 days and a millisecond or two? Like time, age can take on an infinite number of possibilities and so it's a continuous variable.
- The price of gas. Sure, it might be $4 a gallon. But one time in recent history it was 99 cents. And give inflation a few years it will be $99. not to mention the gas stations always like to use fractions (i.e. gas is rarely $4.47 a gallon, you'll see in the small print it's actually $4.47 9/10ths

Discrete Variables

A variable is a quantity that has changing values. A discrete variable is a variable that can only take on a certain number of values. In other words, they don't have an infinite number of values. If you can count a set of items, then it's a discrete variable. The opposite of a discrete variable is a continuous variable. Continuous variables can take on an infinite number of possibilities.

Examples of discrete variables:

- Number of quarters in a purse, jar, or bank. Discrete because there can only be a certain number of coins (1,2,3,4,5…). Coins don't come in amounts of 2.3 coins or 10 1/2 coins, so it isn't possible for there to be an infinite number of possibilities. In addition, a purse or even a bank is restricted by size so there can only be so many coins.
- The number of cars in a parking lot. A parking lot can only hold a certain number of cars.
- Points on a 10-point rating scale. If you're graded on a 10-point scale, the only possible values are 1,2,3,4,5,6,7,8,9, and 10.
- Ages on birthday cards. Birthday cards only come in years…they don't come in fractions. So there are a finite amount of possibilities (presumably, about one hundred).

Standard Deviation

Example 1
Calculate the mean and standard deviation of the following:

9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4

Mean (μ) = 9+2+5+4+12+7+8+11+9+3+7+4+12+5+4+10+9+6+9+420 = 140/20 = 7 Mean=7

$(x_i - \mu)^2$ => 4, 25, 4, 9, 25, 0, 1, 16, 4, 16, 0, 9, 25, 4, 9, 9, 4, 1, 4, 9

= 4+25+4+9+25+0+1+16+4+16+0+9+25+4+9+9+4+1+4+9 = 178

Variance

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

Variance = (1/20) × 178 = 8.9

Standard Deviation

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

Standard Deviation σ = √(8.9) = 2.983

Example 2

Occurance(x):(in separated lines)	Frequency of Occurance(f)
1	5
2	12
3	8
4	3
5	0
6	0
7	1

Occurance(X)	Frequency(f)	Freq*X	(X-mean)	(X-mean)2	f*(X-mean)2
1	5	5	-1.483	2.199	10.993
2	12	24	-0.483	0.233	2.797
3	8	24	0.517	0.268	2.14
4	3	12	1.517	2.302	6.906
5	0	0	2.517	6.337	0
6	0	0	3.517	12.371	0
7	1	7	4.517	20.405	20.405
Total ->	29	72	-	-	43.241

Mean = 2.483
Standard Deviation = 1.221
Variance = 1.491

Standard Deviation from list of numbers

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

σ = **lower case sigma**
$\sum$ = **capital sigma**
$\overline{x}$ = **x bar**

Standard Deviation from Frequency Distribution Table

$$\sqrt{\frac{\sum(X-\bar{X})^2}{N}}$$

$$\text{Standard Deviation (population)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \text{mean})^2}{n}}$$

$$\text{Standard Deviation (sample)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \text{mean})^2}{n-1}}$$

Frequency Distribution of a Discrete Variable

Since, a discrete variable can take some or discrete values within its range of variation, it will be natural to take a separate class for each distinct value of the discrete variable as shown in the following example relating to the daily number of car accidents during 30 days of a month.

3 4 4 5 5 3

4 3 5 7 6 4

4 3 4 5 5 5

5 5 3 5 6 4

5 4 4 6 5 6

Table No. 2: Showing frequency distribution for daily number of car accidents during a month.

Number of car accidents	Frequency
3	5

4	9
5	11
6	4
7	1
Total	30

Frequency Distribution of a Continuous Variable

For a continuous variable if we take a class for each distinct value of the variable, the number of classes will become unduly large, thus defeating the purpose of tabulation. In fact, since a continuous variable can assume an infinite number of values within its range of variation, the classification or sub-division of such data is necessarily artificial. Some guidelines that should be followed while dividing continuous data into classes are as follows:

1. The classes should be mutually exclusive, i.e., non-overlapping. No two classes should contain the same interval of values of the variable.
2. The classes should be exhaustive, i.e., they must cover the entire range of the data.
3. The number of classes and the width of each class should neither be too small nor too large. In other words, there should be relatively fewer classes if the difference between the least value of the variable and its highest value is small and relatively more classes if the same difference is large. This difference between the least value of the variable and the greatest value of the variable is called the range of the variable or the data set.
4. The classes should, preferably, be of equal width.

Let us consider the following example regarding daily maximum temperatures in $^{\circ}C$ in a city for 50 days.

28 28 31 29 35 33 28 31 34 29

25 27 29 33 30 31 32 26 26 21

21 20 22 24 28 30 34 33 35 29

23 21 20 19 19 18 19 17 20 19

18 18 19 27 17 18 20 21 18 19

Minimum Value= 17

Maximum Value=35

Range=35-17=18

Number of classes=5 (say)

$\therefore$ width of each class=4

Table No. 3: Showing frequency distribution of temperature in a city for 50 days.

Class Intervals(Temperatures in $^{\circ}C$)	Frequency
17-20	17
21-24	7
25-28	10
29-32	9
33-36	7
Total	50

Standard Deviation

Random Sample

$$\sigma = \sqrt{\frac{\sum(x - \overline{x})^2}{N}}$$

Calculate the standard deviation of the following numbers: 4, 2, 5, 8, 6.

Example of Frequency Distribution Table

Frequency Distribution

$$\sigma = \sqrt{\frac{\sum f \cdot (x_i - \mu)^2}{\sum f}}$$

Calculate the standard deviation from the following:

	f	$x - \overline{x}$	$(x - \overline{x})^2$	$(x - \overline{x})^2 f$
100	1	11.2	125.44	125.44
98	1	9.2	84.64	84.64
95	2	6.2	38.44	76.88
94	1	5.2	27.04	27.04
85	3	-3.8	14.44	43.32
77	1	-11.8	139.24	139.24
74	1	-14.8	219.04	219.04
totals	10			715.60

Standard Deviation Practice Problems (with answers)

1. Consider the following three data sets A, B and C.

A = {9,10,11,7,13}

B = {10,10,10,10,10}

C = {1,1,10,19,19}

a. Calculate the mean of each data set.
b. Calculate the standard deviation of each data set.
c. Which set has the largest standard deviation?

2. The frequency table of the monthly salaries of 20 people is shown below.

Salary (in $)	Number of people with this salary
3500	5
4000	8
4200	5
4300	2

a. Calculate the mean of the salaries of the 20 people.
b. Calculate the standard deviation of the salaries of the 20 people.

ANSWERS:

1.

a. mean of Data set A = (9+10+11+7+13)/5 = 10

mean of Data set B = (10+10+10+10+10)/5 = 10

mean of Data set C = (1+1+10+19+19)/5 = 10

b.

Standard Deviation Data set A

$= \sqrt{[\ ((9-10)^2+(10-10)^2+(11-10)^2+(7-10)^2+(13-10)^2)/5\ ]} = 2$

Standard Deviation Data set B

$= \sqrt{[\ ((10-10)^2+(10-10)^2+(10-10)^2+(10-10)^2+(10-10)^2)/5\ ]} = 0$

Standard Deviation Data set C

$= \sqrt{[\ ((1-10)^2+(1-10)^2+(10-10)^2+(19-10)^2+(19-10)^2)/5\ ]} = 8.05$

c. Data set C has the largest standard deviation.

2.

a. Mean= $3955
b. standard deviation= 282 (rounded to the nearest unit)

Exercise – Student Ages Data

A survey was carried out in a Further Education college where students were asked what age they were on the first day of college. The results were counted and presented in a grouped frequency table as shown below:

Age	**18_28**	**28_38**	**38_48**	**48_58**	**58_68**	**68_78**
No. of Students	**131**	**172**	**152**	**32**	**12**	**1**

You are required to answer the following:

1. What is the modal class?
2. Calculate the mean.
3. Calculate the standard deviation.

Solution

Mean

Mid interval value	23	33	43	53	63	73		
Age	**18_28**	**28_38**	**38_48**	**48_58**	**58_68**	**68_78**		
No. of Students	**131**	**172**	**152**	**32**	**12**	**1**	**Sum of f ->**	500
f by mid interval	3013	5676	6536	1696	756	73	**Sum of fx ->**	17750

Mean(μ)= 35.5

Mean = $\frac{\sum fx}{\sum f}$

Standard Deviation

x	**f**	**Mean**	**(x-M)^2**	**f(x-M)^2**
23	131	35.5	156.25	20468.8
33	172	35.5	6.25	1075
43	152	35.5	56.25	8550
53	32	35.5	306.25	9800
63	12	35.5	756.25	9075
73	1	35.5	1406.25	1406.25
	500			50375

St. Dev(σ)= 10.0374

Standard Deviation formula $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{\sum f}}$ *(μ represents the Mean)*

Answers

1. Modal class = 28 to 38 years old
2. Mean = 35.5 years old
3. Standard deviation = 10.0374 years old

Microsoft Excel

A	B	C	D	E	F	G	H	I	J	K
1	A									
2										
3	x	7	8	9	10	11	12	13		
4	f	1	3	5	7	5	3	1		
5										
6	Mean => f X x	7	24	45	70	55	36	13	250	<= sum c
7		1	3	5	7	5	3	1	25	<= sum c
8										
9	Standard deviation =>(x-M)	-3	-2	-1	0	1	2	3		
10	(x-M)^2	9	4	1	0	1	4	9	28	<= sum c
11	f(x-M)^2	9	12	5	0	5	12	9	52	<= sum c M)^2
12										
13	Standard Deviation=>	1.44222051								
14										
15	**FORMULA**									
16	x	7	8	9	10	11	12	13		
17	f	1	3	5	7	5	3	1		
18										
19	Mean => f X x	=C16*C17	=D16*D17	=E16*E17	70	55	36	13	=SUM(C19:I19)	<= sum c
20		=C17	=D17	=E17	7	5	3	1	=SUM(C20:I20)	<= sum c
21										
22	Standard deviation =>(x-M)	=C16-M6	=D16-M6	=E16-M6	0	1	2	3		
23	(x-M)^2	=C22^2	=D22^2	=E22^2	0	1	4	9	=SUM(C23:I23)	<= sum c
24	f(x-M)^2	=C23*C17	=D23*D17	=E23*E17	0	5	12	9	=SUM(C24:I24)	<= sum c M)^2
25										
26	Standard Deviation =>	=SQRT(M24)								