



# Section 4a: Statistics

**Maths Literacy**, Workshop Series 2010



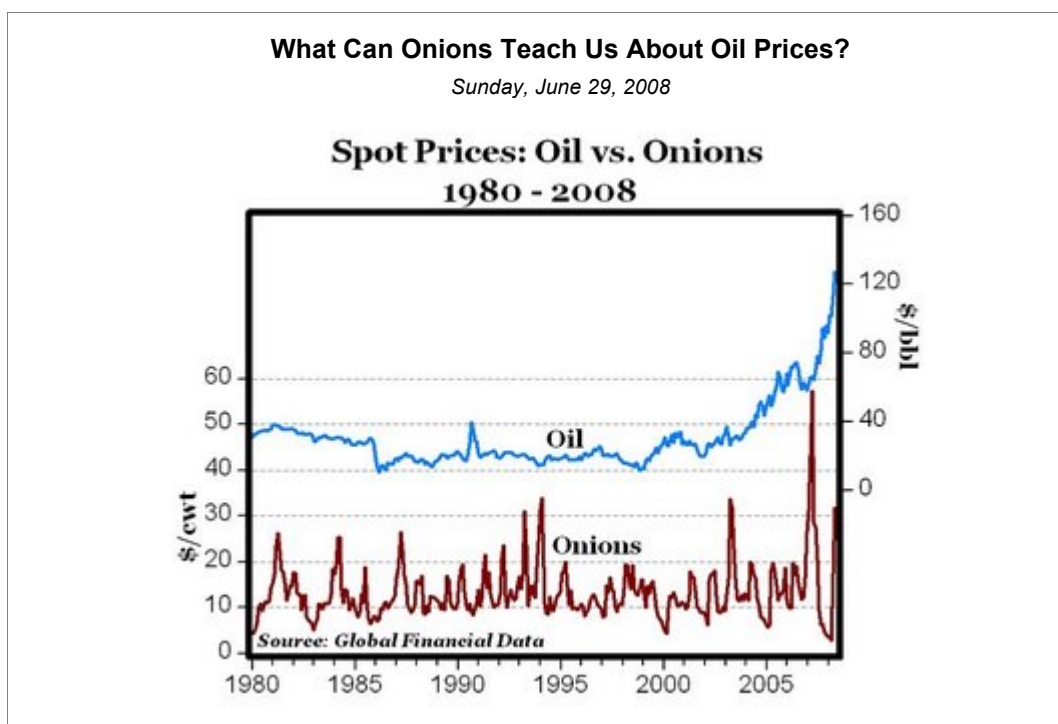
# Section 4a: Statistics

## 1. Introduction

Almost daily we are confronted with statistical information from news reports ("4 out of 5 people earn less than R100 a day"), and even from general conversations ("half the time I don't know what you're talking about").

Statistical literacy is necessary for you to understand material presented in publications such as newspapers, television, and the internet. [Numeracy](#) is a prerequisite to being statistically [literate](#).

Can you interpret the following information correctly?



<http://mjperry.blogspot.com/2008/06/what-onions-teach-us-about-oil-prices.html>

(2009-02-26)

At first it seems like *the volatility* (price change) in onions were lower than that of oil. But is this true? Let's have a look.

1. Estimate the price of oil per barrel in 2000.
2. Estimate the price of onions per cwt crate in 2000.
3. Estimate the percentage increase in the oil price from 2000 to "mid" 2005.
4. Estimate the percentage increase in the price of onions from 2000 to "mid" 2005.
5. Compare your answers in question 3 and 4. Which commodity had the biggest price increase?

Check your answers with those given at the end of the section.

You can also visit the following web site to test yourself on interpreting data

<http://hospitality.hud.ac.uk/studyskills/usingData/InterpretingData/index.htm>

(2009-02-26)

Enjoy the following statistical joke:

Two statisticians were travelling in an airplane from LA to New York. About an hour into the flight, the pilot announced that they had lost an engine, but don't worry, there are three left.

However, instead of 5 hours it would take 7 hours to get to New York. A little later, he announced that a second engine failed, and they still had two left, but it would take 10 hours to get to New York.

Somewhat later, the pilot once again announced that a third engine had failed. But do not fear, he announced, because the plane could fly on a single engine, although now it would take 18 hours to get to New York.

At this point, one statistician turned to the other and said, "We better not loose that last engine, or we'll be up here forever!"

<http://www.onlinemathlearning.com/math-jokes-statistics.html>

(2009-02-26)

## Learning outcomes

At the end of this section you should be able to analyze data by making use of descriptive statistics to make comparisons and draw conclusions.



### START UP ACTIVITY 4.1:

What is the average oil price?

Take a look at the history of oil prices:

Crude oil prices behave much as any other commodity with wide price swings in times of shortage or oversupply. The crude oil price cycle may extend over several years responding to changes in demand as well as OPEC and non-OPEC supply.

The U.S. petroleum industry's price has been heavily regulated through production or price controls throughout much of the twentieth century. In the post World War II era U.S. oil prices at the wellhead averaged \$24.98 per barrel adjusted for inflation to 2007-dollars. In the absence of price controls the U.S. price would have tracked the world price averaging \$27.00. Over the same post war period the median for the domestic and the adjusted world price of crude oil was \$19.04 in 2007-prices.

<http://www.wtrq.com/prices.htm>

(03-03-2009)



1. Explain what averaged \$24.98 per barrel mean.
2. Explain what median \$19.04 per barrel mean.
3. For the first seven months of 2008 the average oil price was \$120 per barrel. The next five months the average oil price was \$80 per barrel. What was the average oil price per barrel in 2008?

## Organising data in frequency distribution tables

In this section we will explain how to organise data. The raw data are first summarised and then presented in a statistical table. The data that we have to summarise can be either discrete or continuous.

- **Discrete data**

Numerical discrete data occur when the measurements are integers that correspond to a count of some sort. Examples are the number of children per family, the number of students smoking in a class. Data are discrete if only a countable number of distinct values are possible.

- **Continuous data**

Measurements that result from the process of measuring rather than counting are continuous rather than discrete, such as time. Unlike a discrete variable, a continuous variable is not limited to particular values such as integers. For example, the sprinting time of an athlete may be measured to the nearest tenth of a second, but it could also be measured to the nearest hundredth of a second.

The main advantage of using tables is that the major data characteristics should become immediately clear to the reader. Tables are also known as frequency distribution tables.

For the rest of this section we will differentiate between *ungrouped data* and *grouped data*.

- Ungrouped data are data which has not yet been classified by class intervals. For example the ages of the 5 students in my class are:  
20          24 22          22          25
- Grouped data refer to data in which we only get information about the 'class interval' into which our variable falls. For example there are 5 students between 20 and 25 years old in my class. In this case I do not know the exact age of each student.

## Ungrouped data

Imagine that you conducted a survey, assessing member usage of a certain gymnasium. Your job was to distribute the questionnaires to members of the gym during a 1-day period and to collect the completed questionnaires the next day. Prior to a complete analysis of the 20 questions that were included on the questionnaire, your manager also wants to get a glimpse of the members' ages.



A total of 100 questionnaires were returned by the members. You wrote down the ages of the 100 members with the results listed in the table below.

40	8	15	29	58	42	49	24	24	15
48	36	33	52	17	35	38	41	12	28
11	22	66	40	34	65	23	49	30	34
16	57	52	53	15	25	31	45	19	34
52	47	47	39	10	48	53	36	18	40
64	13	37	44	44	33	30	26	37	33
21	25	24	60	26	40	52	40	26	42
33	47	43	19	46	52	40	33	35	23
39	27	61	21	23	26	44	39	14	16
69	38	35	37	50	12	25	23	19	12

Observations in this form are referred to as raw data.

### FREQUENCY DISTRIBUTION

As you glance over the data in the table, you realise that your boss will not be able to read anything in it, unless you organize it in some systematic way.

You decide to list all the ages from the lowest to the highest and write the frequency of their occurrences next to them.

Frequency distribution table:

Age $x$	Frequency $f$	Age $x$	Frequency $f$	Age $x$	Frequency $f$	Age $x$	Frequency $f$
8	1	23	4	39	3	55	0
9	0	24	3	40	6	56	0
10	1	25	3	41	1	57	1
11	1	26	4	42	2	58	1
12	3	27	1	43	1	59	0
13	1	28	1	44	3	60	1
14	1	29	1	45	1	61	1
15	3	30	2	46	1	62	0
16	2	31	1	47	3	63	0
17	1	32	0	48	2	64	1
18	1	33	5	49	2	65	1
19	3	34	3	50	1	66	1
20	0	35	3	51	0	67	0
21	2	36	2	52	5	68	0
22	1	37	3	53	2	69	1
		38	2	54	0		
							$\sum f = 100$



By doing this, you have constructed an ungrouped frequency distribution of scores. The table lists all the categories and the number of elements in the data set that belongs to each category. The number of elements in a category is called the frequency of that category.

Check that the sum of the frequencies equals the number of observations in the raw data (100 in the example above).

### CUMULATIVE FREQUENCY DISTRIBUTION

Sometimes it is desirable to know the number of observations which fall below a particular value or in some range of values. In order to obtain this, we must form a cumulative frequency distribution. This is where the frequencies for each observation are progressively added and the cumulative totals placed in another column.

Let us complete the cumulative frequency distribution table for the ages of the 100 members.

Cumulative frequency distribution table:

Age $x$	$f$	Cumulative frequency $F$	Age $x$	$f$	Cumulative frequency $F$
8			23	4	25
9	1	1	24	3	28
10	0	1	25	3	31
11	1	2	26	4	35
12	1	3	27	1	36
13	3	6	28	1	37
14	1	7	29	1	38
15	1	8	30	2	40
16	3	11	31	1	41
17	2	13	32	0	41
18	1	14	33	5	46
19	1	15	34	3	49
20	3	18	35	3	52
21	0	18	36	2	54
22	2	20	37	3	57
	1	21	38	2	59

Age $x$	$f$	Cumulative frequency $F$	Age $x$	$f$	Cumulative frequency $F$
39	3	62	55	0	92
40	6	68	56	0	92
41	1	69	57	1	93
42	2	71	58	1	94
43	1	72	59	0	94
44	3	75	60	1	95
45	1	76	61	1	96
46	1	77	62	0	96
47	3	80	63	0	96



48	2	82	64	1	97
49	2	84	65	1	98
50	1	85	66	1	99
51	0	85	67	0	99
52	5	90	68	0	99
53	2	92	69	1	100
54	0	92			
				$\sum f = 100$	

The cumulative frequency of the 'last' class interval must equal the number of observations in the dataset (100 in the example above).

One might want to know the "halfway" score in the distribution, i.e. such that 50% of the ages are below that score and 50% of the ages are above that score. Lucky for us, our problem add up to a cumulative frequency of 100.

So the "halfway" score in our case is also the 50-th score. Looking at our table the age of 35 is our "halfway" score. This means that 50% of the members are older than 35 years, and 50% are younger than 35.

Note that in this example the ages are widely spread out, a number of ages have a frequency of zero, and there is no immediate clear indication of any patterns. Under these circumstances it is customary to group the scores into class intervals and then obtain a frequency distribution of grouped data. We will discuss this in part 1.4.



#### START UP ACTIVITY 4.2

- The following list of scores is given in a statistics examination.

63	81	35	92	84	87	83	92	88	63
68	76	46	81	83	76	58	57	56	56
77	75	98	57	77	61	62	39	34	61
94	79	54	39	45	54	74	48	77	48

- Set up a frequency and cumulative frequency distribution table.
  - Find the "halfway" score.
- A survey was conducted at a movie theater over a period of two months. Its purpose was to get an overview of the type of films being offered. Three categories were identified: X, for all ages, A, for adults only, P, for parental guidance. The following data have been recorded:

X	P	P	X	X
---	---	---	---	---





A	A	P	A	P
A	P	A	A	P
P	X	X	A	X

- 2.1.** Set up a frequency and cumulative frequency distribution table.
- 3.** Students were asked to rate their experience in a statistics class on a scale from 1 to 5 with 1 = very negative to 5 = very positive. The following data have been recorded:

1	5	4	1	2	3	4	3	5	3
4	3	3	3	4	5	3	5	2	1
3	5	2	4	3	2	5	3	3	2
5	2	5	5	2	5	1	4	4	5
4	1	5	2	1	1	1	3	2	3

- 3.1.** Set up a frequency and cumulative frequency distribution table.
- 3.2.** Find the “halfway” rate.

## Grouped data

The data may be further reduced by grouping the observations into class intervals to form a grouped frequency distribution. A class interval is a range of values in which some of the observation may lie.

The reason for grouping is that some of the scores have such low frequency counts associated with them that it is not justified in maintaining these scores as separate and distinct entities. However, on the negative side, grouping results does lead to the loss of information. For example, individual scores lose their identity when we group them into class intervals.

An important question is: “How many class intervals do we choose?”

Obviously, the interval lengths should not be so large so that we lose too much information provided by our original measurement. For example, if we were to divide the previously collected ages into two classes, those below 39 and those above, practically all the information about the original ages would be lost. On the other hand, the class intervals should not be so small so that the purpose served by grouping is defeated.

### NUMBER OF CLASS INTERVALS

The number of classes depends on the number of observations in the data set. Larger numbers of observations require larger numbers of classes.

We can determine a suitable number of classes by using Sturge's formula:



### Number of class intervals

$$c = 1 + 3,3(\log n)$$

where  $c$  = the number of classes

$n$  = the number of observations in the data set.

Always round the answer up to the next highest integer.

Once the number of classes has been decided on, we can determine the class width (the size) of each class intervals. All classes are of equal width and must not overlap.

### Class width

$$\text{Approximate class width} = \frac{\text{highest value} - \text{lowest value}}{\text{number of classes}}$$

Always round the answer up to the next highest integer.

The class boundaries are the end points of the different class intervals, and which will separate one class from another. The boundaries of the various classes must be established so as to include the entire range (from the minimum value to the maximum value) of observations in the data set.

### Class boundaries

**Step 1A:** When you have discrete data, take the lower boundary of the first class as the lowest value in the data set. Start with a closed interval.

**Step 1B:** When you have continuous data, take the largest integer less than or equal to the smallest data value as the lower boundary of the first interval. Start with a closed interval.

**Step 2:** Find the upper boundary of the first class by adding the class width to the lower boundary. For the upper boundaries, make use of an open ended interval to avoid overlapping of class intervals.

**Step 3:** The lower boundary of the second class is equal to the upper boundary of the first class.

**Step 4:** Determine the upper boundary of the second class by adding the class width to its lower boundary.

**Step 5:** Make use of closed left end points and open right endpoints for all classes

**Step 6:** Repeat until there are  $c$  class intervals.

### REMARK

The rule requiring equal class widths does not apply when the data are spread over a wide range but are highly concentrated in a small part of the range and relatively few numbers elsewhere.

Using smaller widths where the data are highly concentrated and larger widths where they are not concentrated, helps to reduce the loss of information due to grouping.



Let us look at the table of ages of 100 members in a gymnasium again. How many class intervals must we have? What does the class intervals look like?

- **Number of class intervals:**

$$c = 1 + 3,3(\log n) = 1 + 3,3(\log 100) = 7,6$$

Round this number up to 8.

- **Class width**

$$\text{Approximate class width} = \frac{\text{highest value} - \text{lowest value}}{\text{number of classes}}$$

$$= \frac{69 - 8}{8} = 7,625$$

Round this number of to 8.

- **Class boundaries**

Since the data are discreet, the lower boundary starts at 8. Add the class width of 8, to get the upper boundary of 16. Class intervals are given below.

[8,16)  
[16,24)  
[24,32)  
[32,40)  
[40,48)  
[48,56)  
[56,64)  
[64,72)

### FREQUENCY DISTRIBUTION

In practice it is not unusual to construct a grouped frequency distribution directly from the raw data. It should always be verified that  $\sum f$  is equal to the number of observations in the raw data.

Let us complete the grouped frequency distribution for the ages of the 100 members in a gymnasium.

Frequency distribution table:

Class interval	Frequency $f$
[8,16)	11
[16,24)	14
[24,32)	16
[32,40)	21
[40,48)	18
[48,56)	12
[56,64)	4
[64,72)	4
	$\sum f = 100$



You will note that by grouping you have obtained an immediate picture of the distribution of ages of members. For example, you immediately see that the ages range from 8 to 72 years. Also, the majority of members in your sample are 32 to 40 years old. It is also apparent that the number of older members tends to taper off.

### CUMULATIVE FREQUENCY DISTRIBUTION

It is often desirable to rearrange the data from a frequency distribution into a cumulative frequency distribution, which is a distribution that shows the cumulative frequency below the upper real limit of the corresponding class interval. Besides the interpretation of the frequency distribution, a cumulative frequency distribution is of great value in obtaining the median and the various percentile ranks of scores, as we shall see later.

Let us complete the cumulative frequency distribution table for the ages of the 100 members.

Cumulative frequency distribution table:

Class interval $x$	Frequency $f$	Cumulative frequency $F$
[8,16)	11	11
[16,24)	14	25
[24,32)	16	41
[32,40)	21	62
[40,48)	18	80
[48,56)	12	92
[56,64)	4	96
[64,72)	4	100
	$\sum f = 100$	

A cumulative frequency distribution shows the total number of observations in an interval and in all intervals preceding that one.

#### EXAMPLE 4.1

Construct a cumulative frequency distribution table for the sample below.

The weights of 28 boys					
74,4	85,1	77,2	86,3	91,6	77,5
69,6	73,9	60,8	87,7	76,5	90,2
88,8	72,1	62,3	74,9	68,1	109,4
99,5	86,2	88,3	89,4	73,7	
69,5	95,8	77,7	89,6	108,9	



**SOLUTION**

- **Number of class intervals**

$$\begin{aligned}c &= 1 + 3,3(\log 28) \\ &= 1 + 3,3(1,4472) \\ &\approx 5,776\end{aligned}$$

The number rounds up to 6.

- **Class width**

$$\begin{aligned}\text{Approximate class width} &= \frac{\text{highest value} - \text{lowest value}}{\text{number of classes}} \\ &= \frac{109,4 - 60,8}{6} \approx 8,1\end{aligned}$$

The number rounds up to 9.

- **Class boundaries**

Since the data are continuous, the lower boundary starts at 60. Add the class width of 9, to get the upper boundary of 69.

[60,69)  
[69,78)  
[78,87)  
[87,96)  
[96,105)  
[105,114)

Cumulative frequency distribution table:

Class interval $x$	Frequency $f$	Cumulative frequency $F$
[60,69)	3	3
[69,78)	10	13
[78,87)	3	16
[87,96)	9	25
[96,105)	1	26
[105,114)	2	28
	$\sum f = 28$	

The cumulative frequency of the 'last' class interval must equal the number of observations in the dataset, which is 28 in this example.

One might like to know the "halfway" score in the distribution, i.e. such that 50% of the weights are below that score and 50% of the weights are above that score. The frequencies



add up to 28. So the 14-th score will be at 50%.  $\left[\frac{14}{28} = 0,5 = 50\%\right]$  From the table we can see that this score (the 14-th) falls in the interval 78 – 87 kilogram.



### ASSESSMENT ACTIVITY 4.3

1. Use the data in activity 2.2 question 1 and prepare a grouped frequency cumulative frequency distribution table. Find the “halfway” score for the test.
2. The data below represents the weights in kilogram for 35 grade 1 children.

15,3	17,8	13,5	12,0	11,7	16,3	14,2
10,3	14,9	16,8	14,6	10,9	15,2	15,8
13,4	16,1	18,3	15,0	11,7	14,2	15,6
15,6	12,7	15,9	14,1	16,9	12,4	19,7
15,3	11,9	18,2	16,6	12,4	16,3	17,3

- 2.1. Prepare a grouped frequency and cumulative frequency distribution table.
- 2.2. Find the “halfway” weight.

## Graphical representation of frequency distributions

We want to use a graph to represent the numerical data that we have grouped into frequency distribution tables as histograms. For a histogram, adjacent bars are used to represent the individual classes of the frequency table. The width of the bars is equal to the class width and the height of the bars corresponds to the frequency of each class. (See section 3 part 6 on Histograms.)

It is immediately apparent from the histogram what the shape or the distribution of the data in a frequency table looks like.

### How to construct a histogram:

- Step 1:** Mark the class intervals on the horizontal axis and the frequencies of the classes on the vertical axis.
- Step 2:** Label both axes clearly.
- Step 3:** Draw a vertical bar on each class interval where the height of the bar equals the frequency of each class interval.

### EXAMPLE 4.2

Draw a histogram for the data given in the previous example, by using the frequency table that was constructed there:



Class interval	Frequency $f$
[60,69)	3
[69,78)	10
[78,87)	3
[87,96)	9
[96,105)	1
[105,114)	2
	$\sum f = 28$

**SOLUTION:**

The weights of 28 boys

Frequency

60    69    78    87    96    105    114

Weight in kg

From the histogram it is immediately clear that the majority of the boys weigh between 69 and 96 kilogram and that very few boys weigh between 96 and 114 kilogram.

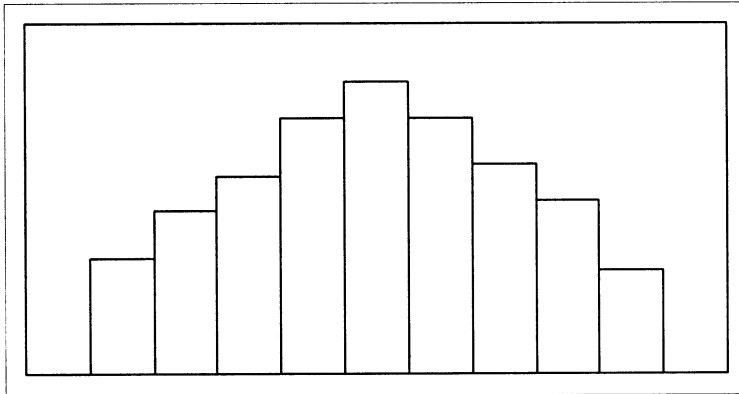
**SHAPES OF FREQUENCY DISTRIBUTION**

The most common shapes are:

- Symmetric
- Skew

A symmetric frequency distribution is identical on both sides of its central point.





If you draw a vertical line at some point in the histogram such that the shape to the left and to the right of the vertical line are mirror images of each other, we say that it is symmetric.

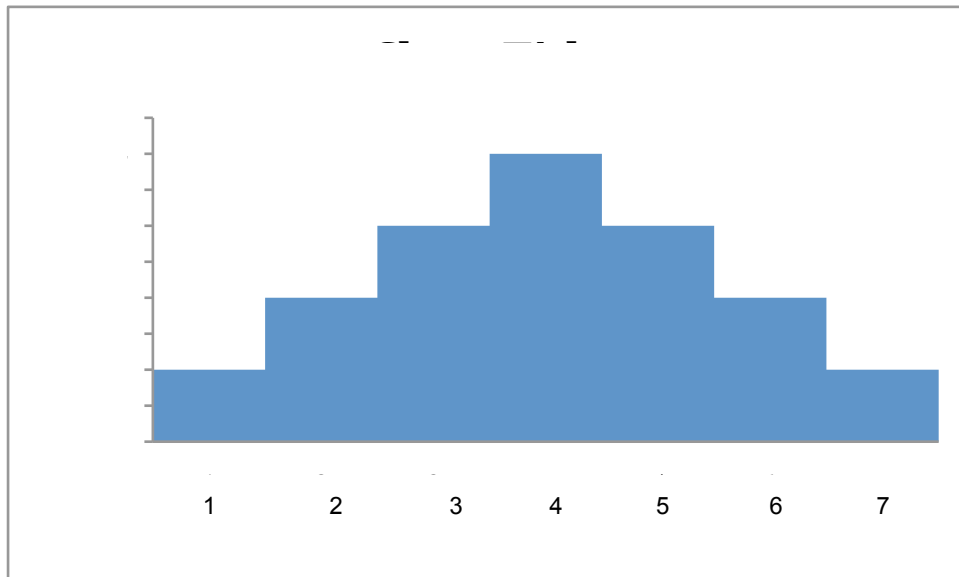
#### EXAMPLE 4.3

Consider the following data set:

1	2	2	3	3	3	4	4	4	4
5	5	5	6	6	7				

Draw a histogram for the given data set.

**SOLUTION:**



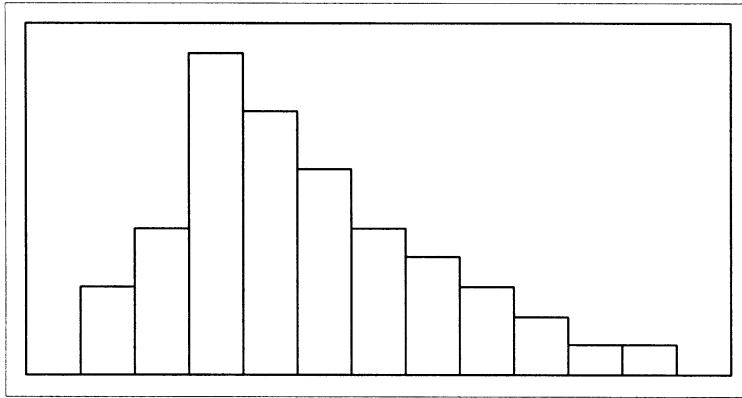
The histogram displays a symmetrical distribution of the data.

For a skew frequency distribution, the tail on the one side is longer than the tail on the other side.





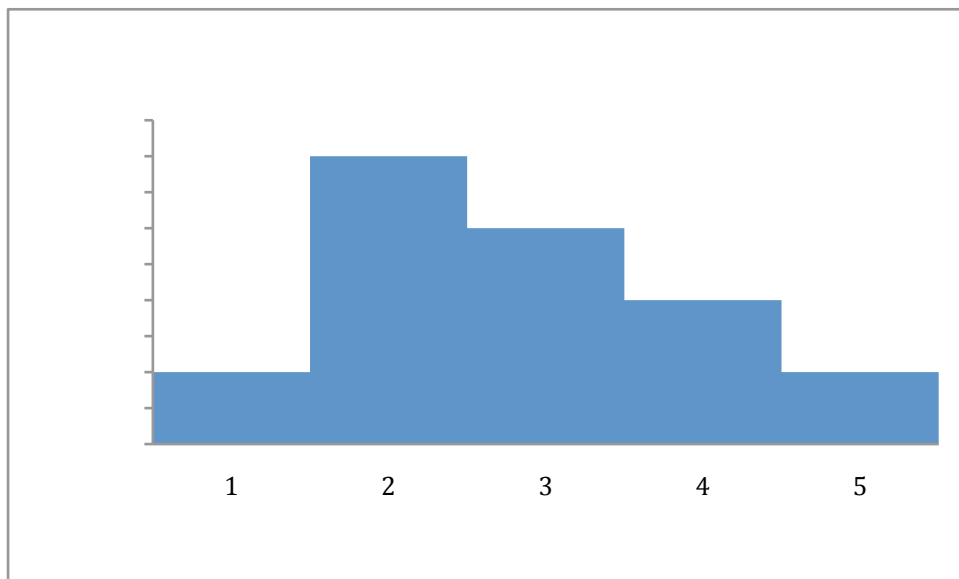
A frequency distribution that is skew to the right has a longer tail on the right hand side.

**EXAMPLE 4.4**

Consider the following data set:

1    2    2    2    2    3    3    3    4    4    5

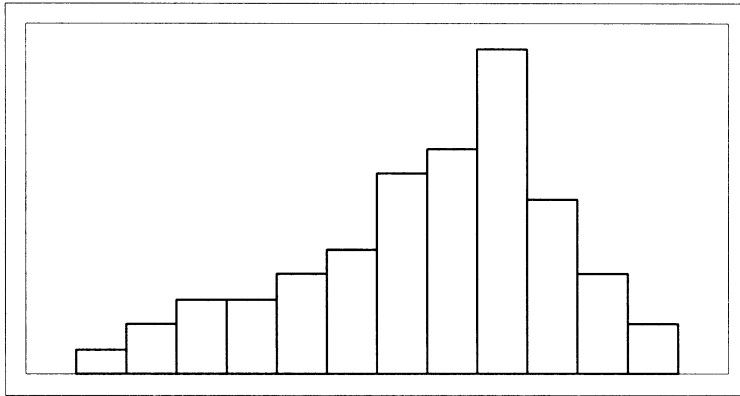
Draw a histogram for the given data set.

**SOLUTION:**

The histogram for the data is not symmetrical. It is skew to the right. The left-hand side seems to be "chopped off" compared to the right side.

A frequency distribution that is skew to the left has a longer tail on the left hand side.

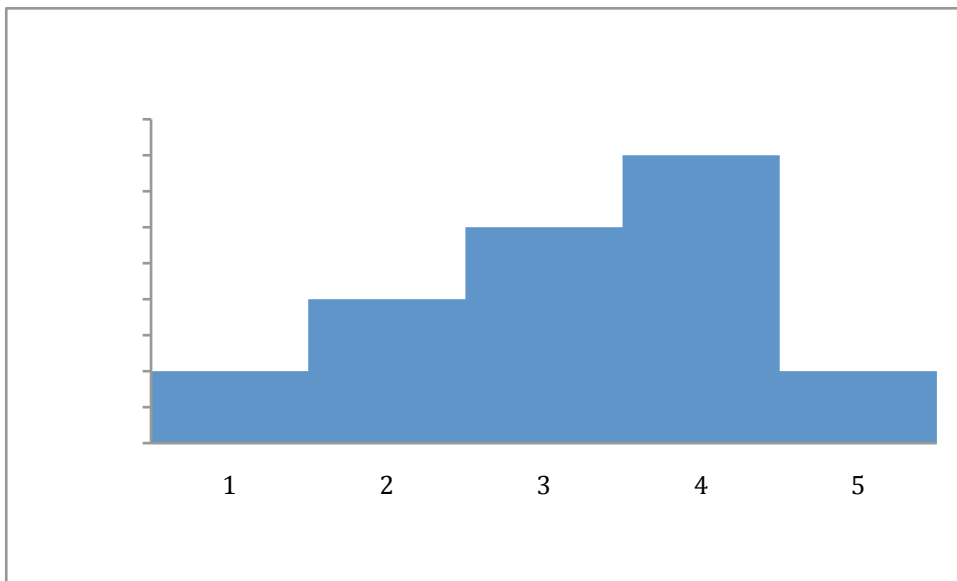


**EXAMPLE 4.5**

Consider the following data set:

1    2    2    3    3    3    4    4    4    4    5

Draw a histogram for the given data set.

**SOLUTION**

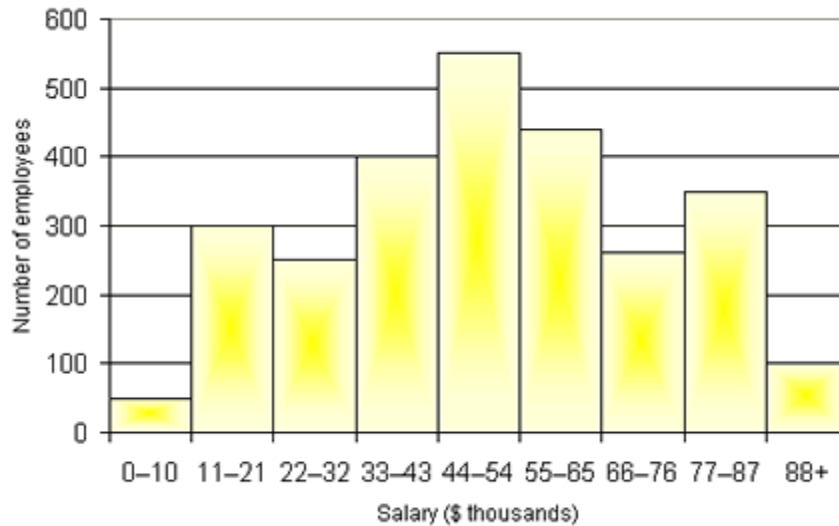
The histogram for this set of data is also not symmetrical. The right-hand side seems to be "chopped off" now, compared to the left hand side.

The purpose of drawing a frequency distribution is to help us to understand the data better. We want to see an overall pattern. The overall pattern can be described in terms of centre, spread and shape of the distribution. We'll discuss this in the following sections.



**EXAMPLE 4.6**

Let's interpret the following histogram.



<http://www.statcan.gc.ca/edu/power-pouvoir/ch9/images/histo1.gif>

(04-03-2009)

1. The centre is a value that divides the observations so that about half of the observations are larger than the centre and about half of the observations are smaller than the centre. The centre is at about 44-54 thousand dollars. That is, the salary for a typical employee is, generally speaking, about 44-54 thousand dollars.
2. The spread (or range) of the salaries is from 0 to 88+ thousand dollar.
3. The histogram has roughly a symmetrical shape.
4. Estimate the number of employees that earns more than 66 thousand dollars. This involves all classes to the right of 66 thousand dollars. This is:  
 $250 + 350 + 100 = 700$  employees  
 Therefore 700 employees earn more than 66 thousand dollars.
5. Estimate the number of employees that earns more than 22 thousand dollars but less than 44 thousand dollars. This involves all classes between 22 and 44 thousand dollars. This gives  
 $250 + 400 = 650$  employees  
 Therefore 650 employees earn more than 22 but less than 44 thousand dollars.

**ASSESSMENT ACTIVITY 4.4**

1. The data below show the number of deaths per age group.

Age Group	Number of Deaths	Proportion of Deaths (%)
0 to 4	3	0.9



5 to 15	11	3.3
16 to 17	10	3.0
18 to 20	39	11.8
21 to 25	42	12.7
26 to 29	29	8.7
30 to 39	54	16.3
40 to 49	40	12.1
50 to 59	39	11.8
60 to 69	28	8.4
70+	37	11.1
<b>Total</b>	<b>332</b>	<b>100</b>

<http://www.tacsafety.com.au/jsp/content/NavigationController.do?arealD=12&tierID=1&navID=4B348A89&navLink=null&pageID=173>

(03-03-2009)

- 1.1. Draw a histogram for the data. Describe the shape of the distribution.
2. Use the data in activity 2.2, question 1, to draw a histogram.

## Measures of central tendency

How do we make sense out of the data that are presented in tabular format?

On many occasions it is convenient to describe a set of numbers by using one single representative number. Most sets of data show a tendency to group or cluster about a certain central point. The term average is often associated with these measures. However, the term average has so many popular meanings, that many statisticians prefer to refer to a measure of central tendency.

Since the center of a distribution may be defined in different ways, there are a number of different measures of central tendency. In this section we will discuss the three most frequently used measures of central tendency: the *arithmetic mean*, the *median* and the *mode* for ungrouped data, and also for grouped data.

### MEASURES OF LOCATION FOR UNGROUPED DATA

#### A. ARITHMETIC MEAN

The most commonly used measure of location is the *arithmetic mean*, or simply called the *mean*. It is defined as the sum of the observations divided by the number of observations. The mean is denoted by  $\bar{x}$ , and calculated according to the formula:



**Formula for arithmetic mean**

$$\text{mean} = \frac{\text{sum of sample observations}}{\text{number of sample observations}}$$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

$$= \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum x}{n}$$

where  $x_1$  = the first observation in the sample

$x_2$  = the second observation in the sample

$\vdots$

$x_n$  = the last observation in the sample

$n$  = the sample size

**EXAMPLE 4.7**

A manager of a restaurant recorded the number of customers who have been served at the restaurant between 18:00 and 11:00, for a particular week. The results were as follows:

52      67      85      62      93      105      49

Find the average number of customers served per night for that week.

**SOLUTION:**

$$\bar{x} = \frac{\text{sum of sample observations}}{\text{number of sample observations}}$$

$$\bar{x} = \frac{52 + 67 + 85 + 62 + 93 + 105 + 49}{7}$$

$$\bar{x} \approx 73,29$$

The mean of a list of integers is not necessarily an integer. "The average customer per night is 73,29 " is a jarring way of making a statement that is more appropriately expressed by "the average number of customers in the collection of nights examined is 73,29".

One of the most important properties of the mean requires an understanding of the concept deviation. Deviation is the distance and direction of a score from a reference point (usually the mean). If the mean is subtracted from all the scores, the sum of these differences (deviations) is zero.

Symbolically, this is written as  $\sum (x - \bar{x}) = 0$ .



The importance of this property of the mean will become more apparent when measures of dispersion are discussed.

Tabulating data in frequency form is often very convenient. It is therefore useful to be able to calculate a mean directly from such a table. In this case the calculation of the mean is found from the following formula:

**Formula for arithmetic mean in table form:**

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

where  $f$  = the frequency of the class interval  
 $x$  = the score in the class interval

**EXAMPLE 4.8**

The annual salaries of the 15 employees of a small hypothetical business are given below.

Annual salary $x$	Frequency $f$
180 000	1
144 000	1
84 000	1
60 000	1
54 000	3
42 000	1
36 000	3
30 000	4
	$\sum f = 15$

Find the mean annual salary per employee.

**SOLUTION:**

Let us complete the table for  $fx$ .

Annual salary $x$	Frequency $f$	$fx$
180 000	1	180 000
144 000	1	144 000
84 000	1	84 000
60 000	1	60 000
54 000	3	162 000



42 000	1	42 000
36 000	3	108 000
30 000	4	120 000
	$\sum f = 15$	$\sum fx = 900000$

Note that the entries in the third ( $fx$ ) column were obtained by multiplying the corresponding entries from the first two columns.

$$\begin{aligned}\text{Mean} &= \frac{\sum fx}{\sum f} \\ \bar{x} &= \frac{\sum fx}{\sum f} \\ \bar{x} &= \frac{900000}{15} \\ \bar{x} &= 60000\end{aligned}$$

The mean salary per employee is R60 000.

Not a bad average salary. But be careful when using this number. After all, only four of the fifteen employees make that much money. The other eleven employees earn less than R60 000 each.

We will now discuss the median. This will give us a better idea of what the average employee is earning.

## B. MEDIAN

The median of a finite list of numbers can be found by arranging all the numbers from the lowest value to the highest value and picking the middle one. 50% of the observations in a data set are smaller than (or equal to) this value and 50% of the observations are larger than (or equal to) this value.

*If there are  $n$  observations in the sample (arranged from the smallest to the largest) and  $n$  is an odd number, the median is the value in the middle of the list, i.e., for odd  $n$ , the median is the  $\left(\frac{n+1}{2}\right)$ -th observation.*

*If  $n$  is even, the median is not unique, and there is no observation in the middle position. The median is then the average of the  $\frac{n}{2}$ -th observation and the  $\left(\frac{n}{2}+1\right)$ -th observation.*



**EXAMPLE 4.9**

A manager of a restaurant has recorded the number of customers who were served at the restaurant between 18:00 and 11:00 during a particular week. The results were as follows:

52    67    85    62    93    105    49

Find the median number of customers served for that week.

**SOLUTION:**

Arrange the data in order of size:

49    52    62    67    85    93    105

In this case,  $n = 7$  (which is odd), therefore the median number is the

$\left(\frac{n+1}{2}\right)$ -th observation. This is the 4-th score. In our case it is 67 customers. The median number of customers served during that week is 67.

**EXAMPLE 4.10**

A study was done on survival time for 16 patients following a new treatment for cancer. The time in months was recorded as follows:

24    20    22    19    21    18    25    16  
20    23    19    45    18    17    18    21

Find the median survival time.

**SOLUTION:**

Arrange the data in order of size:

16    17    18    18    18    19    19    20    20    21    21  
22    23    24    25    45

In this case,  $n = 16$  (which is even), therefore the median survival time is the

the average of the  $\frac{n}{2}$ -th observation and the  $\left(\frac{n}{2}+1\right)$ -th observation. In our case it turns out to be the 8-th and the 9-th score which is 20 in both cases.

Median survival time =  $\frac{20+20}{2} = 20$  months.

If the data are in the form of a frequency distribution, the same basic formulae for calculation apply.

**EXAMPLE 4.11**

The manager of a shoe store was interested to know how many shoes of each size were sold during a week. The results were as follows:





Size $x$	Frequency $f$
3	5
4	8
5	9
6	18
7	20
8	17
9	6
10	1
11	1

Find the median shoe size sold for the week.

**SOLUTION:**

To facilitate matters, first construct a cumulative frequency table:

Size shoe $x$	Frequency $f$	Cumulative frequency
3	5	5
4	8	13
5	9	22
6	18	40
7	20	60
8	17	77
9	6	83
10	1	84
11	1	85
	$\sum f = 85$	

Since  $n = 85$  (which is odd), the median shoe size is the  $\left(\frac{n+1}{2}\right)$ -th observation. There are 85 observations, which are already sorted into numerical order. We have to find the 43-rd observation. The smallest number is at the top, where we should start counting. So the 43-th observation is a size 7 shoe. The median shoe size sold for the week, was a size 7.

The difference between the median and the mean is illustrated in the following example for which we already found the mean: See Example 4.8.

**EXAMPLE 4.12**

The annual salaries of 15 employees of a hypothetical small business are given below:

Annual salary $x$	Frequency $f$
----------------------	------------------



180 000	1
144 000	1
84 000	1
60 000	1
54 000	3
42 000	1
36 000	3
30 000	4
	$\sum f = 15$

Find the mean and the median annual salary per employee.

**SOLUTION:**

Let us complete the table including the cumulative frequency. Place the smallest amount at the top to simplify the problem.

Annual salary $x$	Frequency $f$	Cumulative frequency
30000	4	4
36000	3	7
42 000	1	8
54 000	3	11
60 000	1	12
84 000	1	13
144 000	1	14
180 000	1	15
	$\sum f = 15$	

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

$$\bar{x} = \frac{900000}{15}$$

$$\bar{x} = 60000$$

The mean salary per employee is R60 000, as was shown in Example 4.8.

Since  $n = 15$  (which is odd), the median salary is the  $\left(\frac{n+1}{2}\right)$ -th observation. There are 15 observations, which are already sorted into numerical order. We have to find the 8-th observation. From the table, the 8-th observation is R42 000. The median annual salary per employee is R42 000.

Comparing the mean and the median for a set of data can give you an idea of how widely the values in your dataset are spread apart. In this case, there's a somewhat substantial gap between the top earner of the employees and the rest of the employees.



In a sense, the median is the amount that the typical employee earns. By contrast, the mean is not at all typical.

A reporter who writes that the "average employee" earns R42 000 a year, using the median, provides a far more accurate picture. (Eight employees earn less than R42 000, and seven employees earn more than R42 000.)

When you want a statistic that tells you something about the employee in the middle (or a typical employee), you rather use the median. Thus an important characteristic of the median is its insensitivity to extreme scores.

In a later section we will discuss the [standard deviation](#). This value tells us how widely the values in a data set are spread apart. A large [standard deviation](#) tells you that the data are fairly diverse, while a small [standard deviation](#) tells you the data are tightly bunched together.

### C. MODE

The mode is the observation that occurs most frequently in the dataset. It is not really a measure of average, because it records the most frequent observation, which can be far from the centre of all the observations.

We can find this value by simply looking at the frequency of every value in the dataset, and choosing the one with the largest frequency. So the mode of the list (1, 2, 3, 5, 3, 6, 3) is 3.

If each observation occurs the same number of times, there is no mode. The list (2, 6, 7, 5, 4, 1, 9) has no mode. If two or more observations occur the same number of times, there is more than one mode. The list (2, 2, 7, 5, 3, 5, 6, 9) has the two modes 2 and 5.

We can also find the mode from a frequency distribution table. The observation with the largest frequency is the mode.

#### EXAMPLE 4.13

The marital status of 60 female lecturers on campus was recorded as follows:

Marital status $x$	Frequency $f$
Single	9
Married	24
Divorced	19
Widow	8

Find the mode marital status for the data.

#### SOLUTION:

The observation with the largest frequency (24) is married. So the mode marital status is "married".



The mode also makes sense for non-numerical data. It makes no sense to refer to the mean, or median, degree enrolled for. But it makes sense to say that the most frequent degree enrolled for is "B.Com".



### ASSESSMENT ACTIVITY 4.5

1. The following data give the number of tables produced per day at a certain factory, over a period of 10 days:

24 32 27 23 35 33 29 21 23 25

Calculate the

- 1.1. mean
  - 1.2. median and
  - 1.3. mode for these data.
2. Use the data in activity 2.2, question 1, to calculate the
- 2.1. mean
  - 2.2. median and
  - 2.3. mode for those data.
3. Calculate the mean, median and mode for each of the following sets.
- 5    7        3        4        7        2        7  
111 5        4        7        2        4
- 3.1. In which of the above sets of data is the mean a poor measure of central tendency? Why?
4. The table below shows the number of registered vehicles per province.

Province	Number registered in Apr 2005	Number registered in Apr 2006
GA	2,900,841	3,121,079
KZ	1,056,054	1,133,329
WC	1,298,539	1,381,231
EC	526,369	563,496
FS	460,886	479,449
MP	463,554	498,536
NW	435,993	459,520
LI	340,242	367,738
NC	165,592	174,966
RSA	7,648,070	8,179,344

<http://www.arrivealive.co.za/documents/stats/2006Part1A.pdf>  
(03-03-2009)



- 4.1. Calculate the mean for April 2005.
  - 4.2. Calculate the median for April 2005.
  - 4.3. Explain the big difference in these two answers. Which one would be the best measure of central tendency?
  - 4.4. Calculate the percentage increase in registered vehicles from April 2005 to April 2006.
5. Find the mode type of film in question 2 of activity 2.2.
  6. Find the mode rate in question 3 of activity 2.2.

## Measures of location for grouped data

### A. ARITHMETIC MEAN

If the data are given in the form of a frequency table, the values of the raw data are unknown and we cannot obtain the sum of the individual values.

Therefore, it is not possible to calculate the exact mean of the original data since information is lost when the data are grouped. The mean can only be estimated from a grouped frequency distribution, but this approximation is generally considered as very good.

Therefore another formula is needed to calculate the arithmetic mean for grouped data.

The approximate mean can be calculated by the formula:

#### Formula for mean

$$\bar{x} = \frac{\sum f_i m_i}{n}$$

where  $f_i$  = the frequency of the  $i^{\text{th}}$  class interval  
 $m_i$  = the class midpoint of the  $i^{\text{th}}$  class interval  
 $n$  = sample size

To start, the midpoint of each interval is used to represent all scores within that interval. Thus, the assumption is made that the scores in an interval are evenly distributed.

#### Class midpoints

$$\text{Class midpoint} = m = \frac{\text{lower boundary of that class} + \text{upper boundary of that class}}{2}$$



**EXAMPLE 4.14**

Calculate the mean for the following grouped data.

Class interval $x$	Frequency $f$
[75 – 79)	3
[80 – 84)	4
[85 – 89)	8
[90 – 94)	10
[95 – 99)	15
[100 – 104)	20

**SOLUTION:**

The usual method for such a calculation is given in the table below. Note that the midpoint of each class interval can be found by calculating the mean of the endpoints of the class interval.

Class interval $x$	Frequency $f$	Midpoint $m_i$	$fm_i$
[75 – 79)	3	77	231
[80 – 84)	4	82	328
[85 – 89)	8	87	696
[90 – 94)	10	92	920
[95 – 99)	15	97	1455
[100 – 104)	20	102	2040
	$\sum f = 60$		$\sum fm_i = 5670$

$$\bar{x} = \frac{\sum f_i m_i}{n}$$

$$\bar{x} = \frac{5670}{60} = 94,5$$

The mean for the grouped data is 94,5.

**EXAMPLE 4.15**

The following table gives information on the amount (in Rand) of the telephone bills for a sample of 50 families for a specific month.

Amount of telephone bills $x$	Frequency $f$
[150 – 300)	5
[300 – 450)	16
[450 – 600)	11



[600 – 750)	10
[750 – 900)	8

Calculate the mean amount of the telephone bills for that month.

**SOLUTION:**

Amount of telephone bills $x$	Frequency $f$	Midpoint $m_i$	$fm_i$
[150 – 300)	5	225	1125
[300 – 450)	16	375	6000
[450 – 600)	11	525	5775
[600 – 750)	10	675	6750
[750 – 900)	8	825	6600
	$\sum f = 50$		$\sum fm_i = 26250$

$$\bar{x} = \frac{\sum f_i m_i}{n}$$

$$\bar{x} = \frac{26250}{50} = 525$$

The mean amount of the telephone bills for that month was R525.

## B. MEDIAN

It is not possible to calculate the exact value of the median of the original data once the data are grouped into classes. The median can only be estimated from a grouped frequency distribution.

To calculate the approximate median for grouped data, first find the class interval that contains the median by calculating  $\frac{n}{2}$ . The first cumulative frequency that is equal to or just larger than  $\frac{n}{2}$ , will indicate this class interval.

Using this class interval, the median can be calculated by the following formula:

### Formula for median

$$\text{Median} = l + \frac{(u_i - l_i)(\frac{n}{2} - F_{i-1})}{f_i}$$

where  $l$  = lower class boundary of the interval containing the median

$u_i$  = upper class boundary of the interval containing the median

$F_{i-1}$  = cumulative frequency of the previous class interval (the one just before the class interval that contains the median)



$f_i$  = frequency of the class interval containing the median  
 $n$  = sample size

**EXAMPLE 4.16**

The following table gives information on the amount (in Rand) of the telephone bills for a sample of 50 families for a month.

Amount of telephone bills $x$	Frequency $f$
[150 – 300)	5
[300 – 450)	16
[450 – 600)	11
[600 – 750)	10
[750 – 900)	8

Calculate the median amount of the telephone bills for that month.

**SOLUTION:**

Amount of telephone bills $x$	Frequency $f$	Cumulative Frequency
[150 – 300)	5	5
[300 – 450)	16	21
[450 – 600)	11	32
[600 – 750)	10	42
[750 – 900)	8	50
	$\sum f = 50$	

$$\frac{n}{2} = \frac{50}{2} = 25$$

The cumulative frequency just larger than 25 is 32. The median is contained in the class interval [450 – 600).

Thus:

$$l_i = 450$$

$$u_i = 600$$

$$F_{i-1} = 21$$

$$f_i = 11$$

$$\text{Median} = l_i + \frac{(u_i - l_i)(\frac{n}{2} - F_{i-1})}{f_i}$$





$$\begin{aligned}
 &= 450 + \frac{(600 - 450)(25 - 21)}{11} \\
 &= 450 + \frac{600}{11} \\
 &= 450 + 54,5455 \\
 &= R504,55
 \end{aligned}$$

The median amount of the telephone bills for that month was R504,55. Compare your answer with the mean found in Example 4.15.

**EXAMPLE 4.17**

Calculate the median for the following grouped data.

Class interval $x$	Frequency $f$
[75 – 79)	3
[80 – 84)	4
[85 – 89)	8
[90 – 94)	10
[95 – 99)	15
[100 – 104)	20

**SOLUTION:**

Class interval $x$	Frequency $f$	Cumulative Frequency
[75 – 79)	3	3
[80 – 84)	4	7
[85 – 89)	8	15
[90 – 94)	10	25
[95 – 99)	15	40
[100 – 104)	20	60
	$\sum f = 60$	

$$\frac{n}{2} = \frac{60}{2} = 30$$

The cumulative frequency just larger than 30 is 40. The median is contained in the class interval [95 – 99)

Thus:

$$\begin{aligned}
 l_i &= 95 \\
 u_i &= 99 \\
 F_{i-1} &= 25
 \end{aligned}$$



$$f_i = 15$$

$$\begin{aligned} \text{Median} &= l_i + \frac{(u_i - l_i)(\frac{n}{2} - F_{i-1})}{f_i} \\ &= 95 + \frac{(99 - 95)(30 - 25)}{15} \\ &= 95 + \frac{20}{15} \\ &= 96,3 \end{aligned}$$

The median for the grouped data is 96,3.

Compare your answer with the mean found in Example 4.14.

### C. MODE

We can estimate the mode since information is lost when the data are grouped. The class interval with the largest frequency is defined as the modal class.

#### Formula for mode

$$\text{Mode} = L + \frac{d_1}{d_1 + d_2}(c)$$

where  $L$  = lower limit of the modal class  
 $d_1$  = frequency of the modal class minus frequency of the previous class  
 $d_2$  = frequency of the modal class minus frequency of the following class  
 $c$  = width of the class interval

#### EXAMPLE 4.18

Calculate the mode for the following grouped data.

Class interval $x$	Frequency $f$
[75 – 79)	3
[80 – 84)	4
[85 – 89)	8
[90 – 94)	10
[95 – 99)	15
[100 – 104)	20
	$\sum f = 60$

#### SOLUTION

From the table we can see that the modal class is [100 – 104). (It has the highest frequency.)



Thus:

$$L = 100$$

$$d_1 = 5$$

$$d_2 = 20 \quad (\text{if the modal class is not followed by another class, then } d_2 \text{ is taken}$$

$$c = 4 \quad \text{as the frequency of the modal class)}$$

$$\text{Mode} = L + \frac{d_1}{d_1 + d_2}(c)$$

$$\text{Mode} = 100 + \frac{5}{5 + 20}(4) = 100,8$$

Compare your answer with the mean and median obtained in Examples 4.14 and 4.17.

#### EXAMPLE 4.19

The following table gives information on the amount (in Rand) of the telephone bills for a sample of 50 families during a specific month.

Amount of telephone bills $x$	Frequency $f$
[150 – 300)	5
[300 – 450)	16
[450 – 600)	11
[600 – 750)	10
[750 – 900)	8
	$\sum f = 50$

Calculate the mode amount of the electric bills for that month.

#### SOLUTION:

From the table it is immediate that the modal class is [300 – 450). (It has the highest frequency, namely 16.)

Thus:

$$L = 300$$

$$d_1 = 11$$

$$d_2 = 5$$

$$c = 150$$

$$\text{Mode} = L + \frac{d_1}{d_1 + d_2}(c)$$

$$\text{Mode} = 300 + \frac{11}{11 + 5}(150) \approx 403$$

The mode amount of the telephone bills for that month was R403. This is the amount that appears most frequently during that month.



Compare your answer to the mean and median in Examples 4.15 and 4.16.



### ASSESSMENT ACTIVITY 4.6

1. Use activity 2.3 question 1 and calculate the
  - 1.1. mean
  - 1.2. median and
  - 1.3. mode scores for the data given.
  - 1.4. Compare your answers with those in activity 2.5 no.2.
2. Use activity 2.3 question 2 and calculate the
  - 2.1. mean
  - 2.2. median and
  - 2.3. mode weights for the data given.
3. Use activity 2.4 question 1 and calculate the
  - 3.1. mean
  - 3.2. median and
  - 3.3. mode ages for the data given.

## Relationship between the mean, median and mode

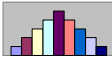
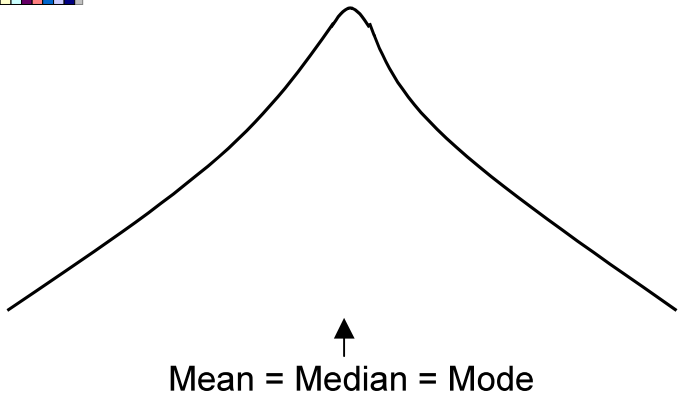
	Advantages	Disadvantages
<b>Mean</b>	It is the most familiar and commonly used measure of central tendency.	Because the calculation of the mean includes each value in the data set, it is greatly affected by any extreme values. Extreme values are values that are very small or very large relative to the majority of the values in a data set.
<b>Median</b>	The advantage of using the median, is that it is not influenced by extreme values. When extreme values are present, it is appropriate to use the median rather than the mean to describe the data set. It's probably most useful when using ordinal data. For ordinal data we rank the data in different categories according to graded order (greater than, less than, equal to). For example: if you have to evaluate your lecturer, you may use the categories "very satisfied," "moderately satisfied," "very dissatisfied."	A disadvantage of the median is that its calculation does not include all the values in the data set.



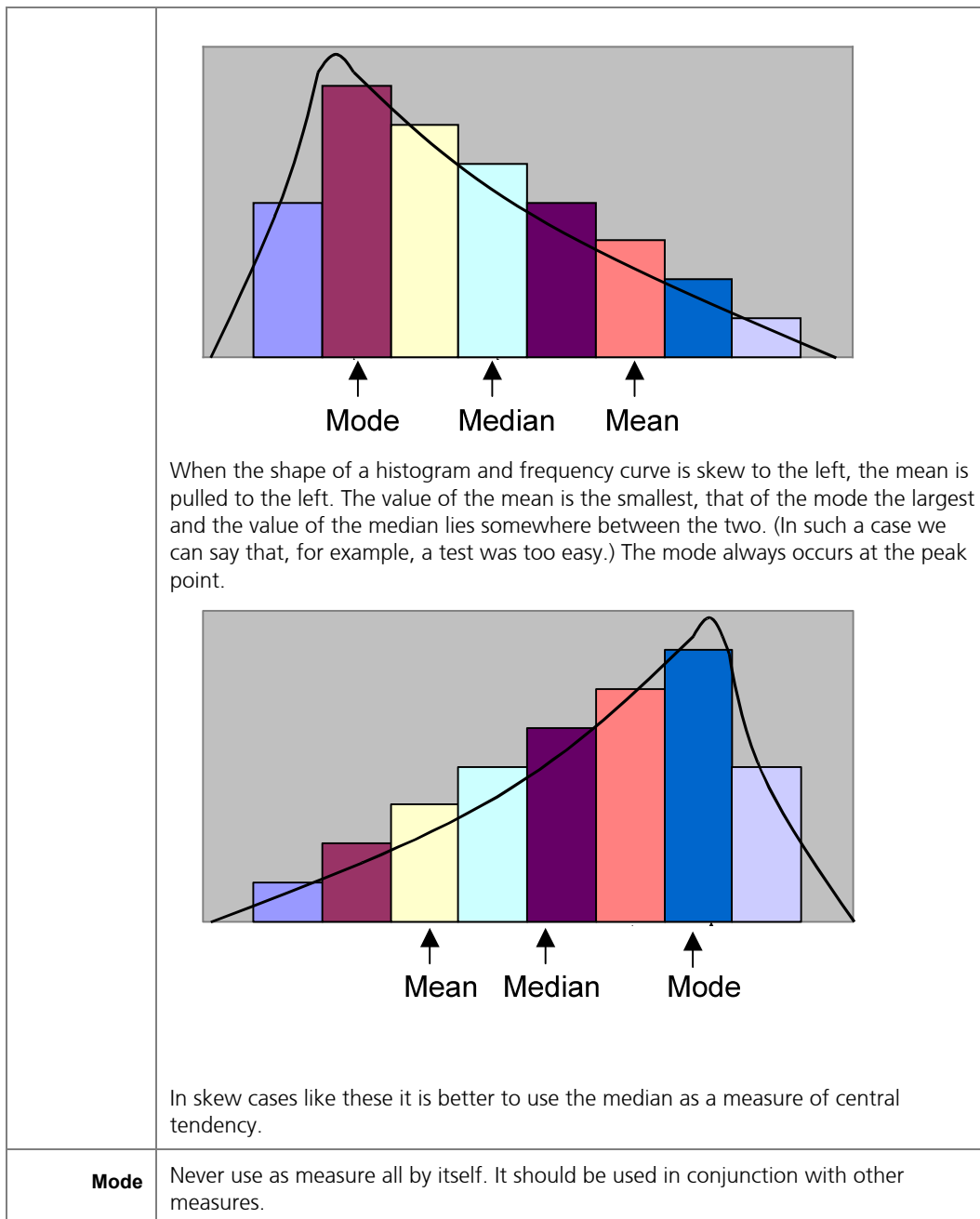
<b>Mode</b>	The mode is also not affected by extreme values. It's probably most useful when using nominal data. For nominal data we use classification according to specified categories of the data for example your marital status, gender or religion.	A disadvantage is that the mode does not include all the values in the data set; it is only those with the highest frequency that are taken into account. It is rarely used as a single measure for describing the central tendency for a set of data.
-------------	---	--

### THE MEAN, MEDIAN, MODE AND SKEWNESS

Two of the shapes that a frequency distribution curve can assume, are symmetric and skew.

<b>Mean</b>	<p>When the shape of a frequency curve is symmetric, the values of the mean, median and mode are the same and they lie at the centre of the distribution.</p>   <p style="text-align: center;"><b>Mean = Median = Mode</b></p> <p>In such a case we use the mean.</p>
<b>Median</b>	<p>When the shape of a frequency curve is skew, the mean is pulled away from the median and the three measures of central tendency differ. The mean tends to be pulled in the direction of the tail. As the distribution becomes more and more skewed, the differences among the measures of central tendency become greater.</p> <p>When the shape of a histogram and frequency curve is skew to the right, the mean is pulled to the right. The value of the mean is the largest, that of the mode the smallest and the value of the median lies somewhere between the two. (In such a case we can say that, for example, a test was too difficult.)</p>





### ASSESSMENT ACTIVITY 4.7

On the basis of the following measures of central tendency, state whether or not there is evidence of skewness and, if so, its direction:

1.  $\bar{x} = 52$       median = 55      mode = 62
2.  $\bar{x} = 63$       median = 55      mode = 49



3.  $\bar{x} = 55$       median = 55      mode = 55
4. For question 2 in activity 2.6 answer the following:
- 4.1. Indicate whether or not there is evidence of skewness and, if so, its direction.
- 4.2. Which measure of central tendency is the preferred one?

## Percentiles

Suppose you took a test and received a score of 87. By itself the score doesn't mean a lot, but if you know that your score of 87 put you at the 60-th percentile, you have a better understanding of your performance. Now you know that 60% of the students who took the test received a score lower (or equal to) yours and 40% received a higher score. Knowing the percentile rank of a score, allows you to compare it with other scores in a sensible way.

The median divides the data set into two halves: a top half and a bottom half (provided the data are sorted from the smallest to the highest value). The top half consists of those data elements above the median, whereas the bottom half consists of those data elements below the median.

Percentiles divide the data set into 100 equal parts. Each group represents 1% of the data set. There are 99 percentiles indicated by  $P_1, P_2, \dots, P_{99}$ .

$P_{50}$  is the same as the median.

## Obtaining the percentile rank for a given score

### UNGROUPED DATA

#### Formula for percentile rank of score $X$

$$\text{Percentile rank of score } X = \frac{L}{n} \times 100$$

where,  $L$  = cumulative frequency of the interval preceding the one containing  $X$   
 $X$  = given score  
 $n$  = sample size

#### EXAMPLE 4.20

The manager of a shoe store was interested to know how many shoes of each size were sold during a particular week. The results were as follows:

Find the percentile rank for size 7 shoe.

Shoe size $x$	Frequency $f$	Cumulative Frequency
3	5	5



4	8	13
5	9	22
6	18	40
7	20	60
8	17	77
9	6	83
10	1	84
11	1	85
$\sum f = 85$		

**SOLUTION:**

$$L = 40$$

$$n = 85$$

$$\begin{aligned} \text{Percentile rank} &= \frac{L}{n} \times 100 \\ &= \frac{40}{85} \times 100 \\ &\approx 47\% \end{aligned}$$

Compare your answer with Example 4.11. What do you notice?

**EXAMPLE 4.21**

The annual salaries of 15 employees of a hypothetical small business are given below.

Annual salary $x$	Frequency $f$	Cumulative Frequency
180 000	1	1
144 000	1	2
84 000	1	3
60 000	1	4
54 000	3	7
42 000	1	8
36 000	3	11
30 000	4	15
$\sum f = 15$		

Find the percentile rank for the salary of R42 000.

**SOLUTION:**

$$L = 7$$

$$n = 15$$

$$\begin{aligned} \text{Percentile rank} &= \frac{L}{n} \times 100 \\ &= \frac{7}{15} \times 100 \end{aligned}$$





$$\approx 47\%$$

Compare your answer with Example 4.12. What do you notice?

### GROUPED DATA

#### Formula for percentile rank of score $X$ in grouped data

$$\text{Percentile rank} = \frac{L + \left(\frac{X - X_2}{c}\right)f_i}{n} \times 100$$

where,  $L$  = cumulative frequency of the interval preceding the one containing  $X$

$X$  = given score

$X_2$  = score at the lower limit of the interval containing  $X$

$c$  = width of the interval

$f_i$  = frequency of the interval containing  $X$

$n$  = sample size

#### EXAMPLE 4.22

The following table gives information on the amounts (in Rand) of the telephone bills for a sample of 50 families during a particular month.

Calculate the percentile rank of a bill of R504.

Amount of telephone bill $x$	Frequency $f$	Cumulative Frequency
[150 – 300)	5	5
[300 – 450)	16	21
[450 – 600)	11	32
[600 – 750)	10	42
[750 – 900)	8	50
	$\sum f = 50$	

#### SOLUTION:

$$L = 21$$

$$X = 504$$

$$X_2 = 450$$

$$c = 150$$

$$f_i = 11$$

$$n = 50$$



$$\begin{aligned} \text{Percentile rank} &= \frac{L + \left(\frac{X - X_2}{c}\right)f_i}{n} \times 100 \\ &= \frac{21 + \left(\frac{504 - 450}{150}\right)11}{50} \times 100 \\ &\approx 50\% \end{aligned}$$

Compare your answer with Example 4.16. What do you notice?

## Finding the score for a given percentile rank

We explain this section by considering the following example:

The table below gives information on the amount (in Rand) of the telephone bills for a sample of 50 families during a particular month.

Amount of telephone bill $x$	Frequency $f$	Cumulative Frequency
[150 – 300)	5	5
[300 – 450)	16	21
[450 – 600)	11	32
[600 – 750)	10	42
[750 – 900)	8	50
	$\sum f = 50$	

Suppose we are interested in the score (telephone bill) lying at the 57-th percentile.

First we must determine the cumulative frequency corresponding to the 57-th percentile. Once this is done, we locate the interval that contains the cumulative frequency, and then establish the amount in question.

### Formula to find cumulative frequency, or cum $f$

$$\text{cum } f = \frac{\text{percentile rank}}{100} \times n$$

where,  $n$  = sample size

Find cum  $f$  for the 57-th percentile:

$$\begin{aligned} \text{cum } f &= \frac{\text{percentile rank}}{100} \times n \\ &= \frac{57}{100} \times 50 \\ &= 28,5 \end{aligned}$$

From the table we can see that this score falls in the interval [450 – 600).



Substitute the answer in the following formula to solve the problem.

**Formula for calculating the score for a given percentile rank for grouped data**

$$\text{Score at given percentile} = X + \frac{c(X_1 - X_2)}{f_i}$$

where,  $X_1$  = cumulative frequency of the score

$X_2$  = cumulative frequency of the interval preceding the interval containing cum  $f$

$X$  = score at the lower limit of the interval containing cum  $f$

$c$  = width of interval

$f_i$  = frequency within the interval containing cum  $f$

We can now solve our problem:

$$X_1 = 28,5$$

$$X_2 = 21$$

$$X = 450$$

$$c = 150$$

$$f_i = 11$$

$$\begin{aligned} \text{Score at given percentile} &= X + \frac{c(X_1 - X_2)}{f_i} \\ &= 450 + \frac{150(28,5 - 21)}{11} \\ &\approx R552 \end{aligned}$$

Score (or telephone bill) at the 57-th percentile is R552.

From Example 4.22 we know that the 50-th percentile equals R504. Let us use this information to test our formula.

**EXAMPLE 4.23**

The following table gives information on the amount (in Rand) of the telephone bills for a sample of 50 families during a particular month.

Calculate the score (telephone bill) at the 50-th percentile.

Amount of telephone bill $x$	Frequency $f$	Cumulative Frequency
[150 – 300)	5	5
[300 – 450)	16	21



[450 – 600)	11	32
[600 – 750)	10	42
[750 – 900)	8	50
	$\sum f = 50$	

**SOLUTION:**

Find cum  $f$  for the 50-th percentile:

$$\begin{aligned} \text{cum } f &= \frac{\text{percentile rank}}{100} \times n \\ &= \frac{50}{100} \times 50 \\ &= 25 \end{aligned}$$

Substitute the following into the formula:

$$X_1 = 25$$

$$X_2 = 21$$

$$X = 450$$

$$c = 150$$

$$f_i = 11$$

$$\begin{aligned} \text{Score at given percentile} &= X + \frac{c(X_1 - X_2)}{f_i} \\ &= 450 + \frac{150(25 - 21)}{11} \\ &\approx \text{R}504 \end{aligned}$$

The score at the 50-th percentile is R504.

It seems as if the formula is doing what it is supposed to be doing.

## Quartiles and deciles

Occasionally you will encounter terms like *deciles* and *quartiles*. These notions refer to specific divisions on the scale of percentile ranks.

A quartile is a percentile rank that divides a distribution into 4 equal parts. There are 3 quartiles in any scale of percentile ranks. Although one might occasionally speak of the bottom quartile, top quartile, etc., the term quartile technically refers to the three division points and not to the four divisions of the data. The definition specifies that at least 25% of the data will be less than or equal to the first quartile,  $Q_1$ , and at least 75% of the data will be less than or equal to the third quartile,  $Q_3$ . In the same way, at least 50% of the data are less than or equal to the second quartile  $Q_2$ , so  $Q_2$  is simply another term for the median. As a consequence,  $Q_1 = P_{25}$ ,  $Q_2 = P_{50}$ ,  $Q_3 = P_{75}$ .

A decile is a percentage rank that divides a distribution into 10 equal parts. The 9-th decile,  $D_9$ , is equal to the 90-th percentile,  $P_{90}$ . The 5-th decile,



$D_5$ , the 2-nd quartile,  $Q_2$ , the 50-th percentile,  $P_{50}$ , and the median are all the same.

Other equivalent terms, such as  $P_{25} = Q_1$ ,  $D_5 = P_{50} = Q_2$ ,  $P_{10} = D_1$ , etc., should be obvious by now.

## Ogive

The ogive is a graphical representation of the cumulative frequency distribution. It is especially useful when you want to display the total at any given time. Let's illustrate this in the following example.

### EXAMPLE:

You saved \$300 in both January and April and \$100 in each of February, March, May, and June. The ogive for the given data is given here:



Ogive of accumulated savings for one year.

An ogive displays a *running total*. Although each individual month's savings could be expressed in a bar chart, it is not at all transparent what the total amount of growth or loss at any given instance is, as is the case in an ogive.

<http://www.cliffsnotes.com/WileyCDA/CliffsReviewTopic/Ogive-Cumulative-Line-Graph-.topicArticleId-25951,articleId-25896.html>

(03-03-2009)

Ogives can be used for estimating the number of observations less than or equal to a particular value. The vertical axis on the left hand side is usually used to indicate the cumulative frequency, while the vertical axis on the right hand side is used to indicate the corresponding percentages.

The curve usually has an 'S' shape.

### How to construct an ogive

**Step 1:** Label the class intervals along the horizontal axis and the cumulative along the vertical axis on the left hand side. The corresponding percentages are labelled on the vertical axis on



the right hand side.

**Step 2:** Mark dots for the various classes at a height equal to the corresponding cumulative frequencies.

**Step 3:** Join the consecutive dots with a smooth curve.

Let's make use of the cumulative frequency distribution table for the ages of the 100 members in part 1.4.2 to draw an ogive.

The cumulative frequency distribution table was as follows:

Class interval $x$	Frequency $f$	Cumulative Frequency
[8,16)	11	11
[16,24)	14	25
[24,32)	16	41
[32,40)	21	62
[40,48)	18	80
[48,56)	12	92
[56,64)	4	96
[64,72)	4	100
	$\sum f = 100$	

Add an additional column for the cumulative frequency percentages. This is where each cumulative frequency is expressed as a percentage of the total frequency ( $\sum f$ ). For example, the cumulative frequency for the class interval [32,40) is:  $11 + 14 + 16 + 21 = 62$ . We also know that  $\sum f = 100$ .

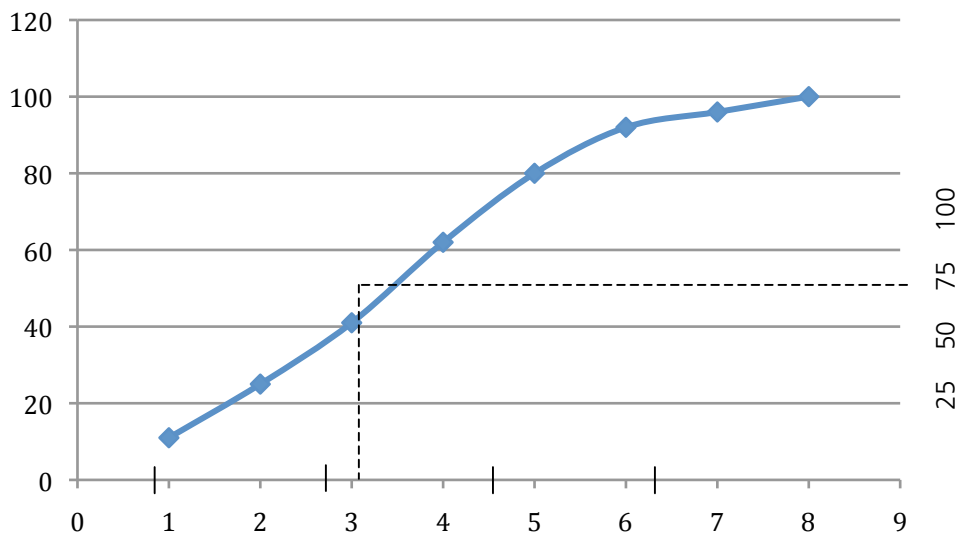
The cumulative frequency percentage is:  $\frac{62}{100} \times 100 = 62\%$

Class interval $x$	Frequency $f$	Cumulative frequency $F$	Cumulative frequency percentage
[8,16)	11	11	11
[16,24)	14	25	25
[24,32)	16	41	41
[32,40)	21	62	62
[40,48)	18	80	80
[48,56)	12	92	92
[56,64)	4	96	96
[64,72)	4	100	100
	$\sum f = 100$		

The ogive is now given below.

Ages of 100 members





To estimate the 'halfway' point in the distribution, i.e. such that 50% of the members have an age below that point and the other 50% are older, we draw a horizontal dotted line from the 50% point on the percentage scale until it hits the ogive curve, then project it vertically down to the horizontal axis. It seems as if 50% of our members are younger than 35 years old.



#### ASSESSMENT ACTIVITY 4.8

1. Suppose you wrote a statistics test. The teacher tells you that you scored at the 90th percentile. Will you be pleased with your performance?
2. Use activity 2.2 question 1 to find the percentile rank for the score 56.
3. The scores for your statistics test were as follows:

42 47 50 62 65 70 72 72 78 80 82 84 84 85 86 88 88 90 94 96 98 98 99

Find the percentile rank for the score 70.

4. Use activity 2.3 question 2 to find:
  - 4.1. The percentile rank for 10,9 kilogram.
  - 4.2. Find  $P_{76}$ .
  - 4.3. Find  $Q_1$ .
  - 4.4. Find  $D_6$ .
  - 4.5. Find  $D_5$ . Compare this with the median that you have already calculated.
5. The rainfall figures (in mm) in a certain city over a period of 1 year were as follows:



12	35	14	23	23	26
25	12	32	24	35	12
13	14	21	19	41	16
16	13	20	15	52	8
2	8	11	8	23	12
9	7	8	5	18	9
16	23	3	2	12	16
45	11	4	4	9	5

- 5.1. Complete a grouped frequency and cumulative frequency table for the data.
  - 5.2. Draw a histogram for the data. Describe the shape of the distribution.
  - 5.3. What percentage of the year had a rainfall of at least 18 mm?
  - 5.4. What percentage of the year had rainfall of less than 26 mm?
  - 5.5. Which interval contains the median?
  - 5.6. Which interval contains the first quartile?
  - 5.7. Find the percentile rank for a rainfall of 41 mm.
  - 5.8. Find  $P_{90}$ .
  - 5.9. Find  $Q_3$ .
  - 5.10. Find  $D_5$ .
  - 5.11. Find the median rainfall for the year.
  - 5.12. Find the average rainfall for the year.
  - 5.13. Indicate whether or not there is evidence of skewness and, if so, its direction.
  - 5.14. Which measure of central tendency is the best to use?
6. Answer the following questions before you draw an ogive for the cumulative frequency table given in 5.1.
- 6.1. Complete the table for the cumulative frequency percentages.
  - 6.2. Find the cumulative frequency that corresponds to the percentage mark of 25%.
  - 6.3. Find the cumulative frequency that corresponds to the percentage mark of 50%.
  - 6.4. Find the cumulative frequency that corresponds to the percentage mark of 75%.
  - 6.5. Draw the ogive curve.
  - 6.6. Make use of your graph to estimate the rainfall below which 75% of the measurements had occurred. Compare your answer with 5.9. What is your conclusion?
  - 6.7. Make use of your graph to estimate the rainfall below which 50% of the measurements had occurred. Compare your answer with 5.10. What is your conclusion?

## Measure of dispersion

The measures of central tendency by themselves cannot sufficiently describe a data set. They can often be misleading in the sense that they don't give any indication of how spread out the data is.

The measurements that tell us how the data set is distributed, or how far each element is from some measure of central tendency, are referred to as measures of dispersion.





There are several ways in which dispersion can be measured. The most popular and most important one is the *standard deviation*, which is an average distance of all the elements from the mean.

There are several reasons for requiring a measure of dispersion of a data set. It will give us an indication of the reliability of the average value. Let us explain this by the following example:

Consider the following two sets of data for the number of mistakes made by two typists:

1<sup>st</sup> data set (for typist 1): 32 27 38 25 20 32 34 28 40 29  
2<sup>nd</sup> data set (for typist 2): 3 80 64 5 11 87 0 2 53 0

Which of the two typists would you employ?

Both data sets have the same mean, namely 30,5. Does this mean that both typists deliver work of the same quality?

From the first set we can see that the numbers of mistakes are all relatively small compared to some of those in the second data set. This means that the first typist was more consistent.

The measures of dispersion should give relatively small values for data that are closely grouped together, like the first data set, and larger values for data that are widely spread-out, like the second data set.

The second data set possesses much more spread or variability than the first, implying a difference in the distributions of the two data sets.

We now discuss three measures of dispersion: *range*, *variance* and *standard deviation*.

## Range

The range is the simplest measure of dispersion and is almost trivial to calculate. It is obtained by taking the difference between the largest and the smallest values in a data set.

$$\text{Range} = \text{largest score} - \text{smallest score}$$

This is not a reliable measure of dispersion, since it takes into account only the most extreme values. It is therefore not a very useful measure of dispersion because in most cases we want information about the distribution of all the values.

The presence of one or two extreme values may result in a very large range and consequently a misleading idea of the true character of the data.



**EXAMPLE 4.24**

Find the range for the following data set:

47    55    38    43    37    92    37    49    52    50

**SOLUTION:**

$$\begin{aligned}\text{Range} &= \text{largest score} - \text{smallest score} \\ &= 92 - 37 \\ &= 55\end{aligned}$$

## Measures of dispersion for ungrouped data

**VARIANCE**

Since we use only two numbers when calculating the range, we have seen that the range often provides a misleading picture of the true distribution of a data set. By making use of all the scores in a set and their individual deviations from the mean, we may obtain a more useful measure of dispersion. (A deviation is the distance between a score in the set and the mean score of the set.)

To study the deviation from the mean, find the difference between each score and the mean. These are called deviations, which indicate the amount by which scores deviate from the mean. For a particular score  $x$ , the deviation is:

$$\text{Deviation} = x - \bar{x}, \text{ where } \bar{x} \text{ is the mean of the scores.}$$

A positive deviation indicates that the score is larger than the mean, while a negative deviation indicates that the score is smaller than the mean. A deviation of zero indicates that the score is equal to the mean.

**EXAMPLE 4.25**

Find the deviation for each of the scores in the data set below:

3    5    2    3    2

**SOLUTION:**

First we calculate the mean:

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} \\ &= \frac{3+5+2+3+2}{5} \\ &= 3\end{aligned}$$



$$\text{Deviation} = x - \bar{x}$$

$$\text{Deviation for } 3 = 3 - 3 = 0$$

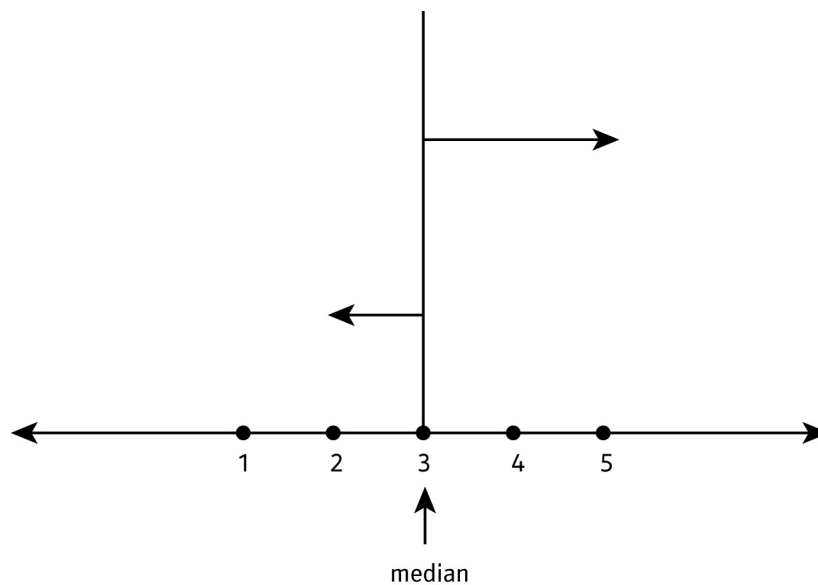
$$\text{Deviation for } 5 = 5 - 3 = 2$$

$$\text{Deviation for } 2 = 2 - 3 = -1$$

$$\text{Deviation for } 3 = 3 - 3 = 0$$

$$\text{Deviation for } 2 = 2 - 3 = -1$$

One can see the meaning of the positive and negative signs of the deviations on a number line:



Can you see that a positive deviation of 2 indicates that the score is to the right of the mean; while the negative deviation of -1, indicates that the score is to the left of the mean. A deviation of zero indicates that the score coincides with the mean.

For any data set, the sum of the deviations is always equal to zero:

$$\sum (x - \bar{x}) = 0 \quad \text{for all data sets}$$

Add all the deviations in the previous example to see that this is true.

The *variance* of the data set is denoted by  $s^2$ , and is calculated according to the following formula:

#### Formula for the variance of a data set

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

where  $\bar{x}$  = the mean



$x$  = scores in the sample  
 $n$  = the sample size

The variance is almost the mean of the squares of the deviations, i.e.  $\frac{\sum (x - \bar{x})^2}{n}$ . There is, however, a sound statistical reason (beyond the scope of this material) why we rather divide the sum of the squares of the deviations by  $n - 1$ , rather than  $n$ .

Although the following formula looks different from the one above, it gives exactly the same value for  $s^2$ . It is just easier to use this one.

#### Another formula for variance

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

This is the formula we shall be using to calculate the variance in the following example.

#### EXAMPLE 4.26

Consider the following two sets of data for the numbers of mistakes made by two typists:

1<sup>st</sup> data set: 32 27 38 25 20 32 34 28 40 29  
 2<sup>nd</sup> data set: 3 80 64 5 11 87 0 2 53 0

Calculate their variances.

#### SOLUTION:

1<sup>st</sup> data set:

$$\begin{aligned}\sum x^2 &= 32^2 + 27^2 + 38^2 + 25^2 + 20^2 + 32^2 + 34^2 + 28^2 + 40^2 + 29^2 = 9627 \\ (\sum x)^2 &= (32 + 27 + 38 + 25 + 20 + 32 + 34 + 28 + 40 + 29)^2 = 93025 \\ s^2 &= \frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n - 1} \\ s^2 &= \frac{9627 - \frac{1}{10}(93025)}{9} \\ s^2 &\approx 36,1\end{aligned}$$

2<sup>nd</sup> data set:

$$\sum x^2 = 3^2 + 80^2 + 64^2 + 5^2 + 11^2 + 87^2 + 0^2 + 2^2 + 53^2 + 0^2 = 21033$$



$$\begin{aligned} (\sum x)^2 &= (3 + 80 + 64 + 5 + 11 + 87 + 0 + 2 + 53 + 0)^2 = 93025 \\ s^2 &= \frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n - 1} \\ s^2 &= \frac{21033 - \frac{1}{10}(93025)}{9} \\ s^2 &\approx 1303,3 \end{aligned}$$

Both data sets have the same mean, namely 30,5. The scores of data set 2 are more spread out around the mean than the scores of data set 1.

This confirms our previous observation that the first typist was more consistent than the second one.

### STANDARD DEVIATION

Because the deviations are squared, the unit of measurement is also squared and therefore it differs from the unit of measurement of the original observations. For example, if the unit of measurement of the original observations is *gram*, the variance is measured in *gram*<sup>2</sup>.

The standard deviation is simply the square root of the variance, to align the measurement with that of the original data.

The standard deviation of a sample is denoted by *s*.

#### Formula for the standard deviation

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

An equivalent formula

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

The value of the standard deviation is large if the data are more spread-out, and small if the data is more concentrated around the mean.

For example, each of the three data sets {0, 0, 10, 10}, {0, 4, 6, 10} and {4, 4, 6, 6} has a mean of 5. Their standard deviations are 4, 3, and 2, respectively. The third set has a smaller standard deviation than the other two because its values are all close to 5.

The practical value of the standard deviation for a set of values is to have an understanding of how much variation there is from the "average" (mean).

*It is estimated that for a typical data set, roughly 90% of the data lie between (the mean - 3 × standard deviations) and (the mean + 3 × standard deviations).*



**EXAMPLE 4.27**

A study was done on the survival time for 16 patients receiving a new treatment for cancer. Their times in months were recorded as follows:

24    20    22    19    21    18    25    16  
20    23    19    45    18    17    18    21

1. Find the range.
2. Find the mean.
3. Find the standard deviation. Explain your findings.

**SOLUTION:**

1. Range = largest score – smallest score  
= 45 - 16  
= 29

2. Mean =  $\bar{x} = \frac{\sum x}{n}$   
 $\bar{x} = \frac{346}{16} \approx 21,6$

3. Let us make use of the formula:  $s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$  to calculate the variance.

Score $x$	Deviation form mean $(x - \bar{x})$	$(x - \bar{x})^2$
24	2,4	5,76
20	-1,6	2,56
22	0,4	0,16
19	-2,6	6,76
21	-0,6	0,36
18	-3,6	12,96
25	3,4	11,56
16	-5,6	31,36
20	-1,6	2,56
23	1,4	1,96
19	-2,6	6,76
45	23,4	547,56
18	-3,6	12,96
17	-4,6	21,16
18	-3,6	12,96
21	-0,6	0,36
		$\sum (x - \bar{x})^2 = 677,76$

Substitute these values into the formula:

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

$$s^2 = \frac{677,76}{15} = 45,184$$



Calculate the square root of the variance to find the standard deviation:

$$s = \sqrt{45,184} = 6,72.$$

Since the mean is 21,6 months, roughly 90% of the survival times lies between  $(21,6 - 3 \times 6,72)$  months and  $(21,6 + 3 \times 6,72)$  months. That means that 90% of the data lie between 1,44 months and 41,76 months.

If the data are in the form of a frequency distribution, the standard deviation can be found using the following formula:

#### Formula for the standard deviation

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

where  $n$  = the sample size

$x$  = score

$f$  = frequency of score

#### EXAMPLE 4.28

The numbers of defective items produced per day were recorded as follows (where the frequency  $f$  for each  $x$  equals the number of days on which  $x$  defective items per day were produced).

Numbers of defective items $x$	Frequency $f$
0	17
1	12
2	19
3	28
4	21
5	19
6	9
7	2

Calculate the standard deviation.

**SOLUTION:**

Number of defective items $x$	Frequency $f$	$fx$	$x^2$	$fx^2$
0	17	0	0	0
1	12	12	1	12
2	19	38	4	76



3	28	84	9	252
4	21	84	16	336
5	19	95	25	475
6	9	54	36	324
7	2	14	49	98
	$\sum f = 127$	$\sum (fx)^2 = 145161$		$\sum fx^2 = 1573$

First calculate the variance:

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

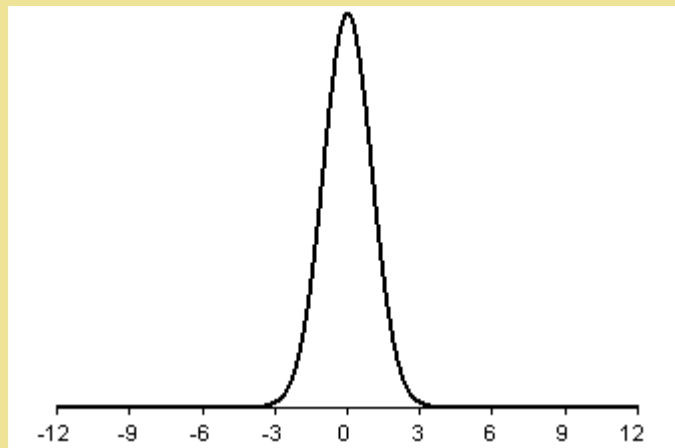
$$s^2 = \frac{1573 - \frac{1}{127}(145161)}{126}$$

$$s^2 \approx 3,413$$

The square root of the variance gives the standard deviation:  $s \approx 1,85$

Maybe the following example will give you a better idea of standard deviation.

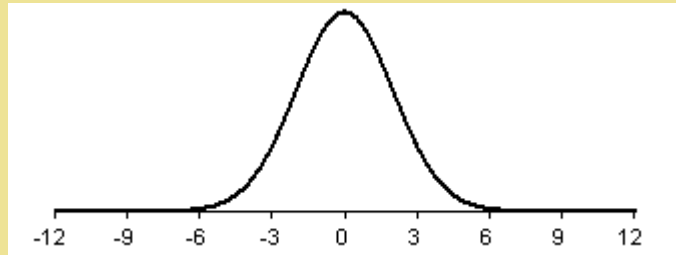
*One conceptual way to think about the standard deviation is that it is a measure of how far the bell-shaped distribution is spread out. Shown below is a bell shaped curve with a standard deviation of 1. Note how tightly concentrated around the mean the distribution is.*



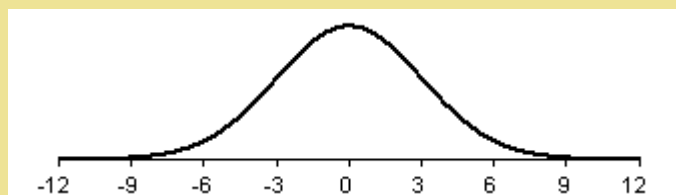
*Shown below is a different bell shaped curve, one with a standard deviation of 2. Notice that the curve is wider, which implies that the data are less concentrated around the mean, and more spread out.*







Finally, a bell shaped curve with a standard deviation of 3 appears below. This curve shows that the data is spread out even more than the data above.



<http://www.childrensmemory.org/stats/definitions/stdev.htm>

(03-03-2009)



#### ASSESSMENT ACTIVITY 4.9

1. Given the data set {5, 6, 8, 9}, calculate the standard deviation. Calculate their deviations from the mean and show that the sum of the deviations is equal to zero.
2. The college netball team played eight games. The number of baskets scored in each game is as follows:
  - 2.1. 8, 6, 6, 7, 9, 4, 8, 5
  - 2.2. Calculate the mean.
  - 2.3. Calculate the median.
  - 2.4. Calculate the mode.
  - 2.5. Calculate the range.
  - 2.6. Calculate the standard deviation for this data set. Explain your findings.
3. Calculate the mean and standard deviation for each of the following two sets of data. Explain your findings.
 

**Data Set 1:** 98, 99, 101, 102

**Data Set 2:** 1, 51, 149, 199
4. Peter and Johnny play for the college cricket team. Here are their scores of their last 10 innings:



**Peter:** 13 127 64 37 83 27 0 103 12 58

**Johnny:** 61 53 40 48 62 32 40 69 43 53

Which one of the two would you pick for the next important game?

5. Two bakeries are selling bread that should weigh 500 gram each. If the weight of a bread is found to be less than 500 gram, the bakery can be taken to court and get a fine. A quality control official weighs a sample of 5 breads from each bakery.

**Bakery A:** 501 500 499 501 499

**Bakery B:** 502 498 503 497 505

Work out the mean and standard deviation for each bakery.

Which bakery has the best chance of avoiding the court?

## Measures of dispersion for grouped data

### VARIANCE

#### Formula for the variance of grouped data

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

where  $n$  = the sample size

$m$  = midpoint of the interval

$f$  = frequency of score

### STANDARD DEVIATION

The standard deviation is obtained by taking the square root of the variance.

#### Formula for the standard deviation of grouped data

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

where  $n$  = the sample size

$m$  = midpoint of the interval

$f$  = frequency of score

#### EXAMPLE 4.29

Calculate the:

1. mean



2. variance
3. and standard deviation of the following grouped data set.

Class interval $x$	Frequency $f$
0 – 2	1
3 – 5	4
6 – 8	9
9 – 11	10
12 – 14	16
15 – 17	11
18 – 20	8
21 – 23	3
24 – 26	1

**SOLUTION:**

First complete the table:

Class interval $x$	Frequency $f$	Midpoint $m$	$fm$	$fm^2$
0 – 2	1	1	1	1
3 – 5	4	4	16	64
6 – 8	9	7	63	441
9 – 11	10	10	100	1000
12 – 14	16	13	208	2704
15 – 17	11	16	176	2816
18 – 20	8	19	152	2888
21 – 23	3	22	66	1452
24 – 26	1	25	25	625
	$\sum f = 63$		$\sum fm = 807$ $(\sum fm)^2 = 651249$	$\sum fm^2 = 11991$

$$1. \quad \bar{x} = \frac{\sum f_i m_i}{n}$$

$$\bar{x} = \frac{807}{63} \approx 12,8$$

2. Substitute into the formula:

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

$$s^2 = \frac{11991 - \frac{1}{63}(651249)}{62}$$

$$s^2 \approx 26,67$$

3. Calculate the square root of the variance to find the standard deviation:  
 $s \approx 5,16$



## Characteristics of standard deviation

The standard deviation is the most frequently used measure of dispersion. It can never be negative. When all the scores in the sample are exactly the same, the standard deviation is equal to zero. In such a case, the range is also zero.

The value of the standard deviation is affected by the value of every score in the sample. If the data contain a number of extreme values, the value of  $s$  may be affected in such a way that it is not a good 'representative' measure of dispersion any more. A large standard deviation indicates that most of the data points are far from the mean and a small standard deviation indicates that almost all data points are clustered closely around the mean.

Why is the characteristic of dispersion so important to the business world? Companies have gone out of business because of their inability to control variation. Goods and services must not only have a good average level of quality, but they must also have small variation in quality. If you buy a battery that says it has a lifetime of 540 days, you expect its lifetime to be very close to 540 days. You do not want a battery for which the life time differs a lot from that value, say with a lifetime of only 300 days.



### ASSESSMENT ACTIVITY 4.10

- The data table below gives the heights in centimeters of a sample of 100 15-year-old children.

165	161	170	182	176	185	180	155	154	166
165	152	174	167	165	171	172	150	181	165
166	161	174	158	166	168	164	150	155	170
168	144	164	154	177	173	178	158	165	175
180	174	152	167	148	175	153	162	180	175
157	172	155	140	147	160	152	166	168	158
153	165	160	143	166	167	167	163	158	160
150	157	172	167	184	172	165	159	158	177
179	174	156	178	165	179	174	148	175	166
157	159	163	165	162	153	145	170	176	180

[http://www19.statcan.ca/02/02\\_017-eng.htm](http://www19.statcan.ca/02/02_017-eng.htm)

- Complete a grouped frequency and cumulative frequency table for the data.
- What percentage of the boys have a height less than 164 centimetres?
- What percentage of the boys have a height of at least 182 centimetres?
- Which interval contains the median?
- Which interval contains the third quartile?
- Find the percentile rank for a height of 170 centimetres.
- Calculate  $P_{65}$ .



- 1.8. Calculate  $Q_1$ .
- 1.9. Calculate  $D_8$ .
- 1.10. Calculate the median height for the boys.
- 1.11. Calculate the average height for the boys.
- 1.12. Indicate whether or not there is evidence of skewness in the distribution, and, if so, its direction.
- 1.13. Which measure of central tendency is the best one to use?
- 1.14. Calculate the range.
- 1.15. Calculate the Standard Deviation. Explain your findings.



#### GROUP ACTIVITY 4.11

Work in groups of three.

Tutorial for using statistics in Excel:

<http://www.bioss.ac.uk/smart/unix/mbasexclides/frames.htm>

(03-03-2009)

[http://en.wikipedia.org/wiki/Image:Standard\\_deviation\\_diagram.svg](http://en.wikipedia.org/wiki/Image:Standard_deviation_diagram.svg)

<http://phoenix.phys.clemson.edu/tutorials/excell/stats.html>

(03-03-2009)

Check all the statistical functions:

<http://phoenix.phys.clemson.edu/tutorials/excell/mathstats.html>

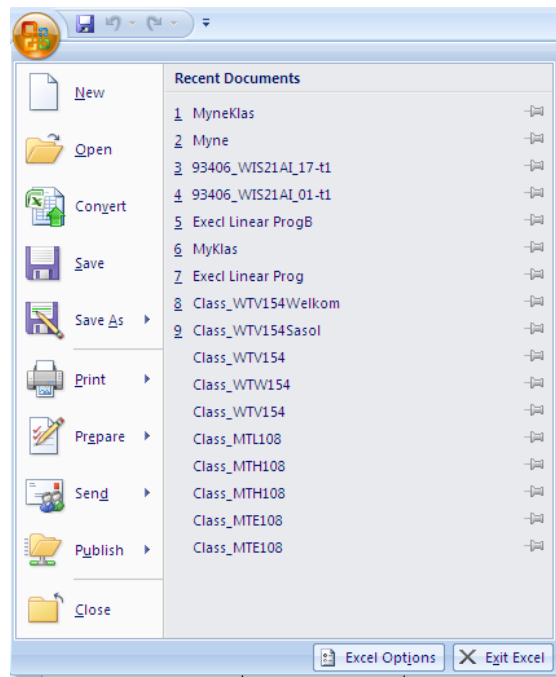
(03-03-2009)

We now solve Statistical Problems by making use of Excel.

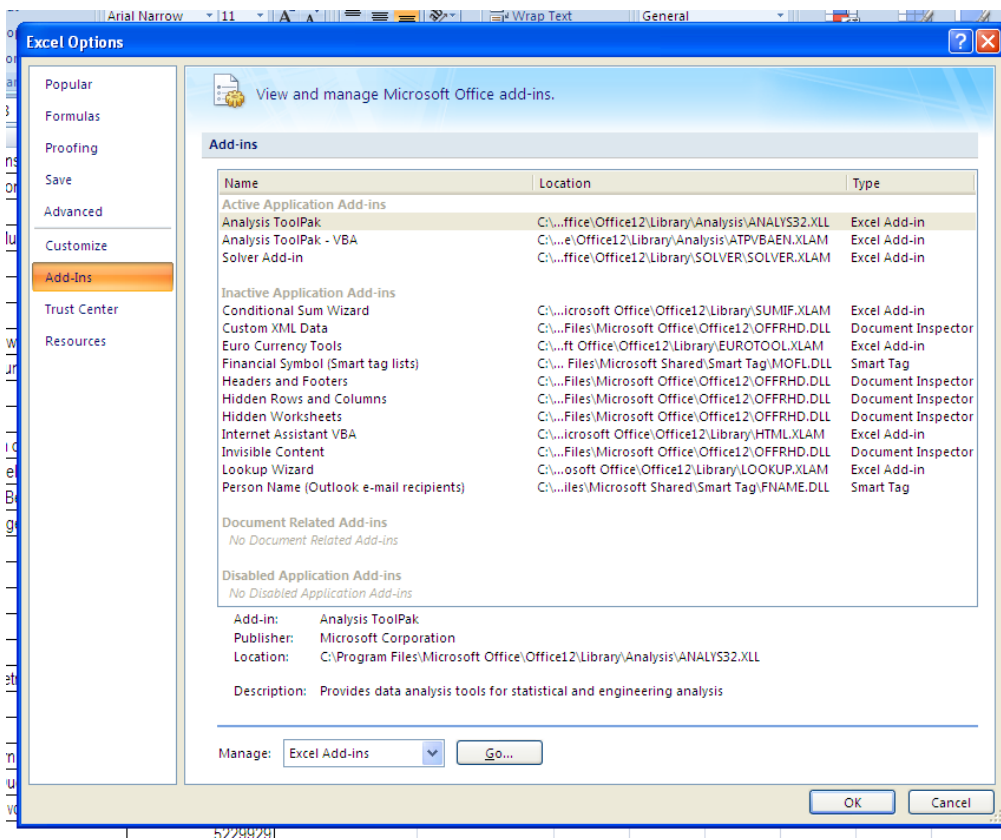
**Note:** Before starting this section, you need to install **Data Analysis** into Excel.

Click **Office button (at the top left)** and choose Excel Options.





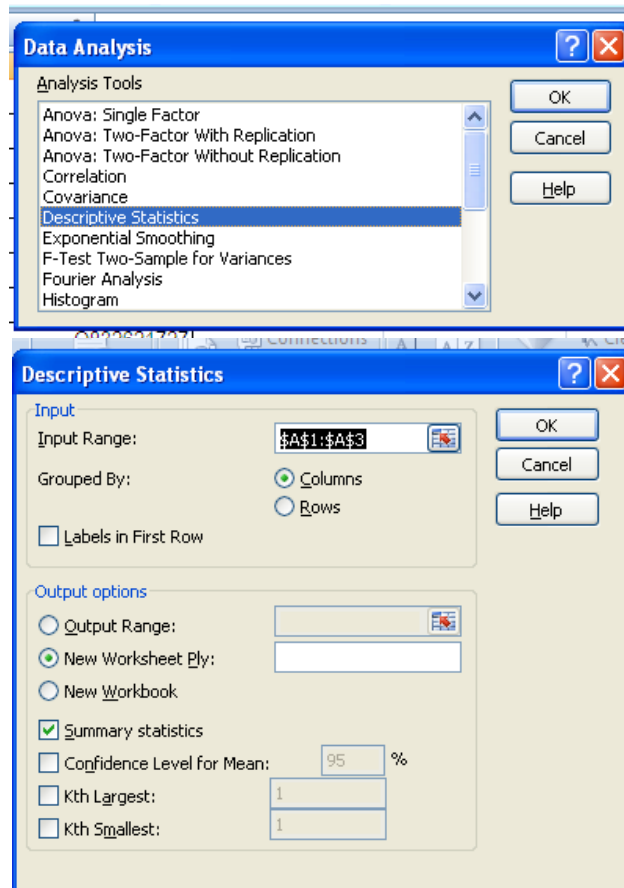
**Add-ins, and Analysis Toolpak: Click on the Go button.**



The Data Analysis package is now installed.



Click on **Data** and you will see **Data Analysis**. In the Data Analysis we will choose the Descriptive Statistics, then choose Summary Statistics as well.



This will give us all the information that we need on mean, median, mode, standard deviation, range and a lot more.

**NOTE:** You cannot use Data Analysis for grouped data.

Let us work together through the following example.

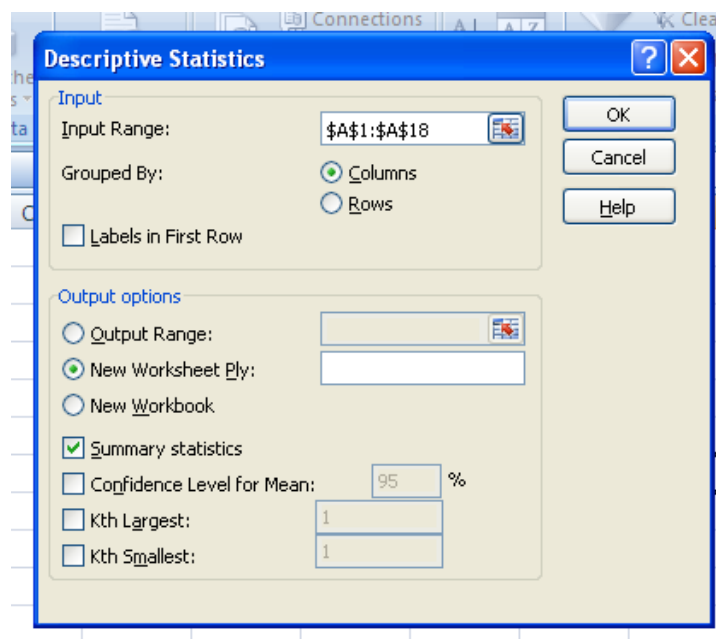
1. Type the following data into an Excel spread sheet:

56	89	69
36	65	53
69	79	72
25	34	69
16	96	82
46	58	75

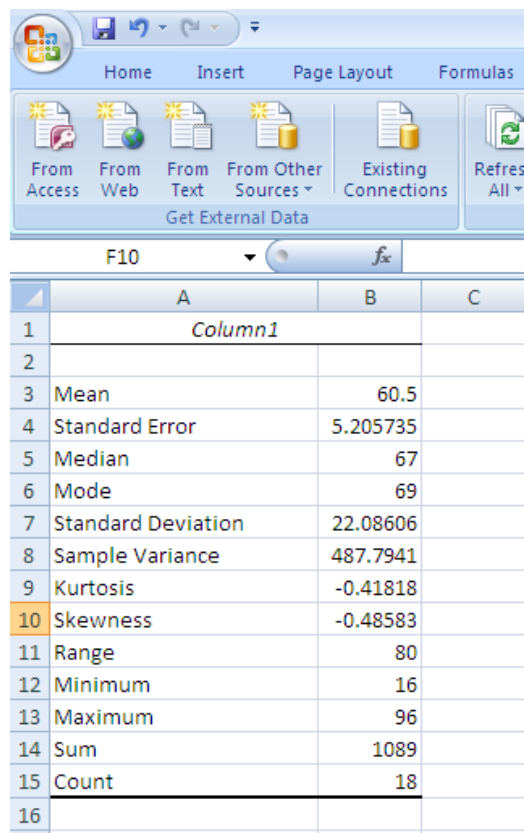


	A	B	C	D
1	56			
2	36			
3	69			
4	25			
5	16			
6	46			
7	89			
8	65			
9	79			
10	34			
11	96			
12	58			
13	69			
14	53			
15	72			
16	69			
17	82			
18	75			
19				

- 1.1. Use Data Analysis to give all the Descriptive Statistics. Remember to choose “Summary statistics”.



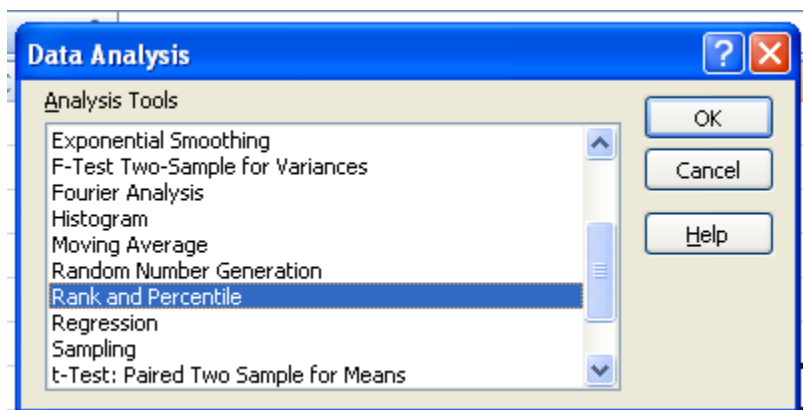


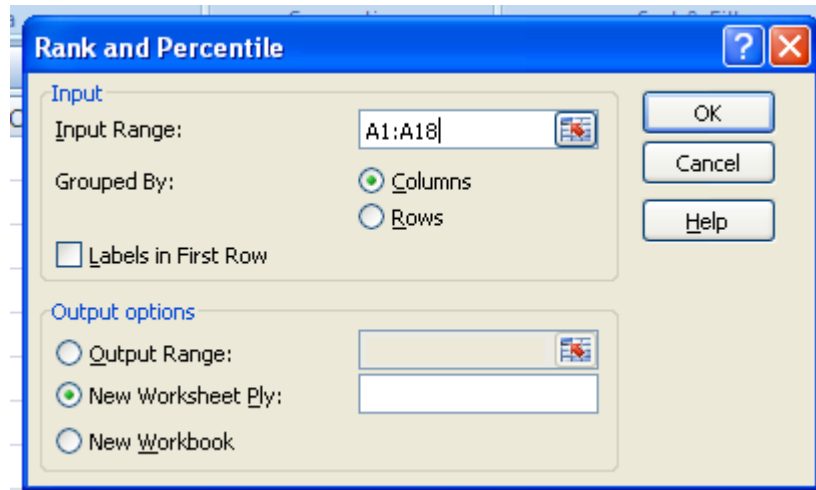
**Descriptive Statistics:**


The screenshot shows the 'Formulas' ribbon in Microsoft Excel, specifically the 'Get External Data' group. Below the ribbon, a summary statistics table is displayed in a spreadsheet format. The table has three columns: A, B, and C. Column A contains statistical measures, and column B contains their corresponding numerical values. The row for 'Skewness' is highlighted in orange.

	A	B	C
1	Column1		
2			
3	Mean	60.5	
4	Standard Error	5.205735	
5	Median	67	
6	Mode	69	
7	Standard Deviation	22.08606	
8	Sample Variance	487.7941	
9	Kurtosis	-0.41818	
10	Skewness	-0.48583	
11	Range	80	
12	Minimum	16	
13	Maximum	96	
14	Sum	1089	
15	Count	18	
16			

- 1.2. Use Data Analysis and give the rank and percentile for each data score.

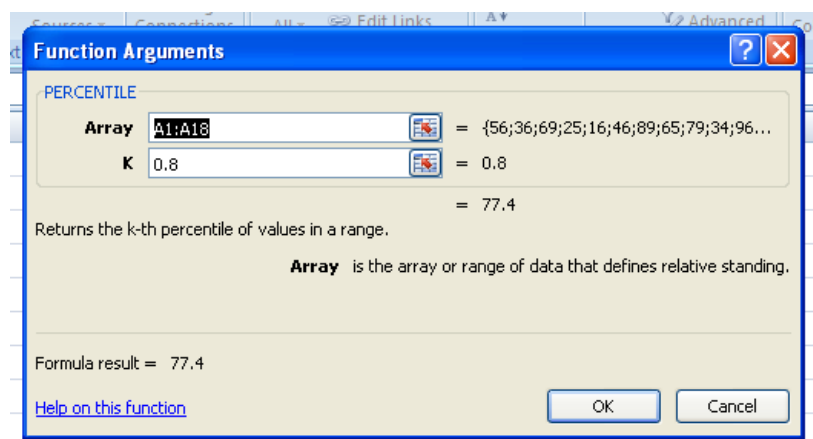
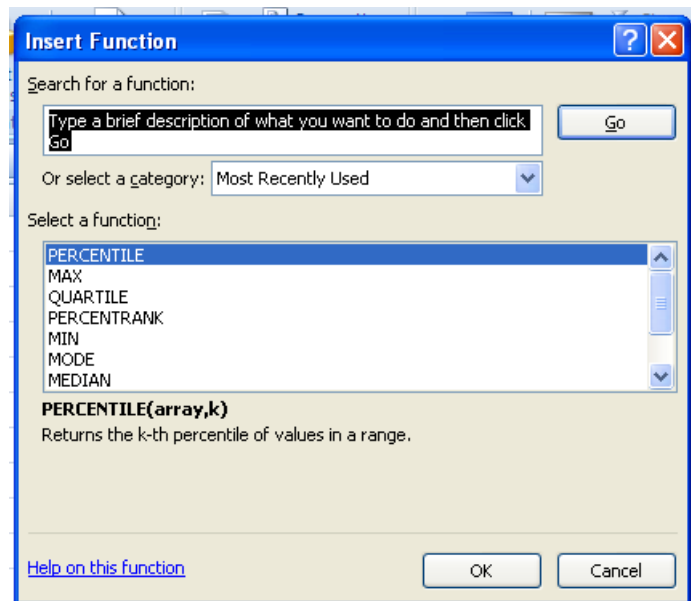




	A	B	C	D	E
1	Point	Column1	Rank	Percent	
2	11	96	1	100.00%	
3	7	89	2	94.10%	
4	17	82	3	88.20%	
5	9	79	4	82.30%	
6	18	75	5	76.40%	
7	15	72	6	70.50%	
8	3	69	7	52.90%	
9	13	69	7	52.90%	
10	16	69	7	52.90%	
11	8	65	10	47.00%	
12	12	58	11	41.10%	
13	1	56	12	35.20%	
14	14	53	13	29.40%	
15	6	46	14	23.50%	
16	2	36	15	17.60%	
17	10	34	16	11.70%	
18	4	25	17	5.80%	
19	5	16	18	0.00%	
20					

- 1.3. Now use the formulas in Excel to give the score for the 80-th and the 30-th percentile.



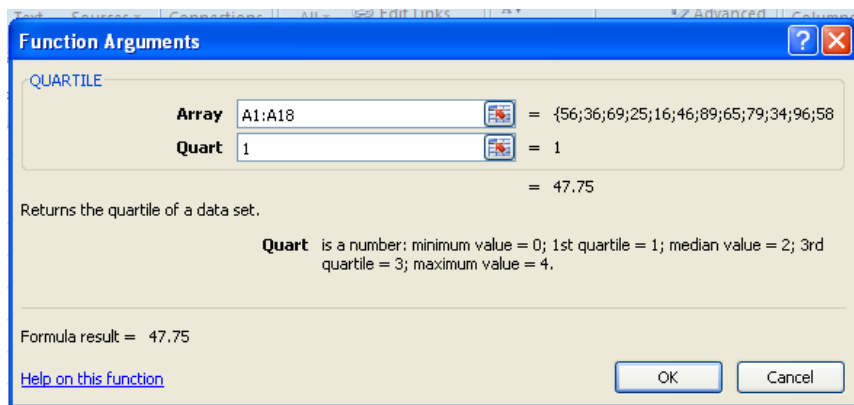
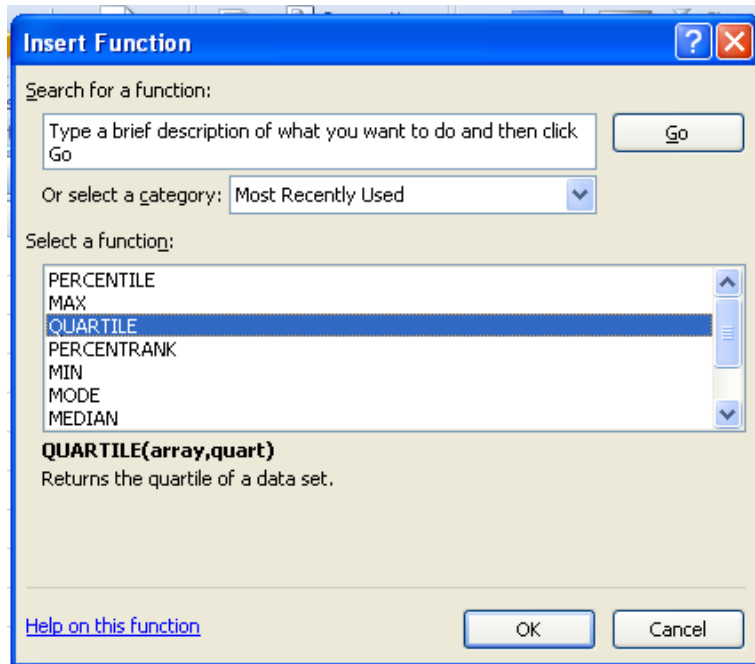


**80-th percentile** = 77.4

**30-th percentile** = 53.3

- 1.4. Now use the formulas in Excel to give the score for the first quartile.

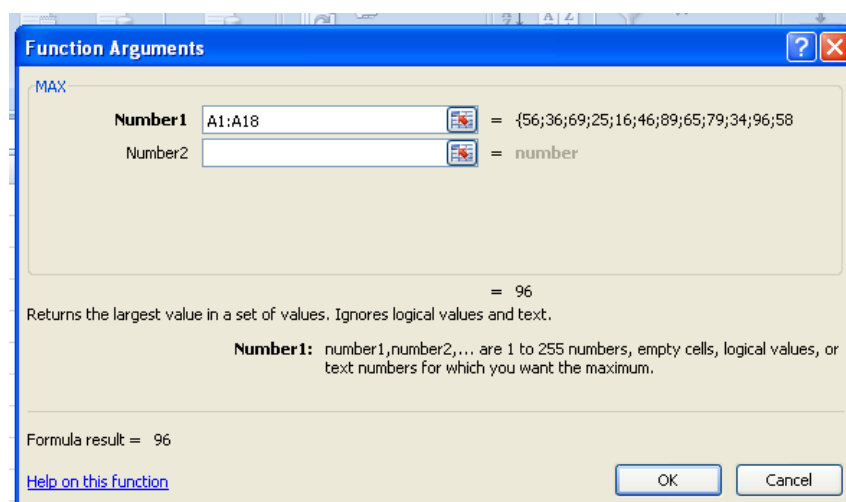
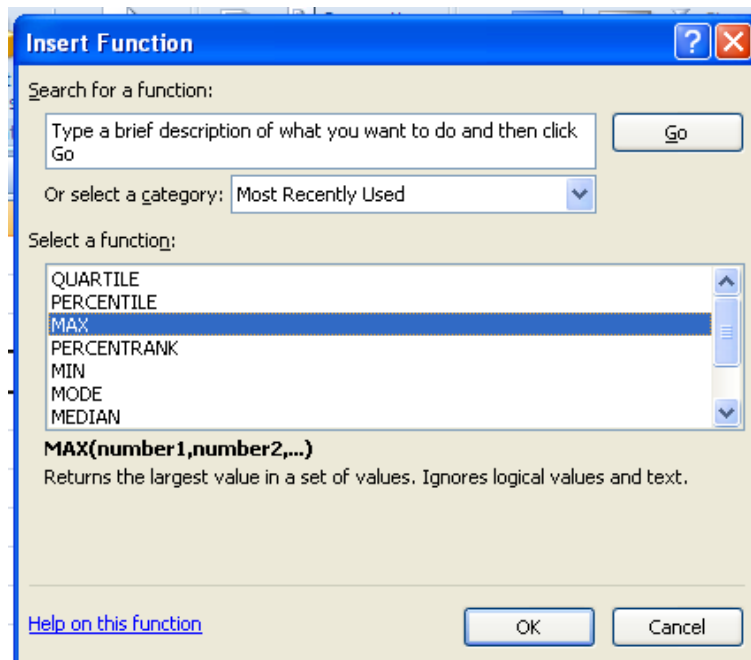




**First quartile = 47.75**

- 1.5. Now use the formulas in Excel to find the maximum score in the data set.





**maximum score = 96**

1.6. Now use the formulas in Excel to find the minimum score in the data set.

**minimum score = 16**

Now proceed by choosing some of the examples that we have discussed in this text and solve them using Data Analysis. Compare the answers that we have obtained with the answers given by Excel.

MAKE A PRINT OUT OF YOUR EXCEL SPREADSHEET FOR ALL THE QUESTIONS BELOW. Make sure your name and student number is also on the paper.



2. You are facilitating at your local school. The principle asks you to write a report on the progress of the students after their first test.

The results for the first test were as follows:

56	89	56	68	42	71	65	52	54	75
25	69	67	63	68	63	63	67	61	65
78	53	23	76	62	67	79	54	73	87
36	46	69	55	75	56	56	53	86	52
45	23	79	76	78	46	62	46	48	64
96	15	62	64	65	52	54	71	67	76
78	49	58	54	62	64	85	64	73	84
67	93	63	42	54	35	74	78	74	34
53	68	71	89	82	95	69	53	65	51
48	34	70	74	76	47	78	36	38	50

Write a report to your principle based on the following questions:

- 2.1. Explain what the mean, mode, median, maximum, minimum, range and standard deviation means, and give the value of each.
- 2.2. Explain what rank and percentile mean. Give a summary of the rank and percentile of each student.
- 2.3. Explain what the first quartile means and also give the value of it.



## End of section comments

In this section we tried to describe a set of numbers by means of a single representative number. In its broadest sense, an average is simply a single figure which represents as best as it can a set of data. Statisticians simply use the term 'measure of central tendency' to describe this idea. However, a measure of central tendency in itself is not sufficient to describe a set of data adequately. A measure of dispersion (or spread) of the data is usually also required.

## Feedback

**ANSWERS TO:** Check how statistical literate you are:

1. According to the axis on the right hand side: it was around 27\$ a barrel
2. According to the axis on the left hand side: it was around 5\$ cwt per crate
3.  $\pm 270\%$
4.  $\pm 1100\%$
5. onions

### ANSWERS TO START UP ACTIVITY 4.1

1. A representative measure of a list of prices that is calculated by taking the sum of the values and dividing it by the number of prices being examined.
2. Median indicates that for 50% of the time the price was higher than \$19,04 a barrel and 50% of the time the price was lower than \$19,04 a barrel.
3. It is \$103,33 and not \$100.

### ANSWERS TO LEARNING ACTIVITY 4.2

1.
  - 1.1. See table below
  - 1.2. 63

Score	Frequency $f$	Cumulative Frequency $F$	Score	Frequency $f$	Cumulative Frequency $F$
34	1	1	68	1	21
35	1	2	74	1	22
39	2	4	75	1	23
45	1	5	76	2	25
46	1	6	77	3	28
48	2	8	79	1	29
54	2	10	81	2	31
56	2	12	83	2	33
57	2	14	84	1	34
58	1	15	87	1	35



61	2	17	88	1	36
62	1	18	92	2	38
63	2	20	94	1	39
			98	1	40
				$\sum f = 40$	

2.

2.1.

Category	Frequency $f$	Cumulative Frequency $F$
X	6	6
A	7	13
P	7	20
	$\sum f = 20$	

3.

3.1. See table below

3.2. 3

Scale	Frequency $f$	Cumulative Frequency $F$
1	8	8
2	9	17
3	13	30
4	8	38
5	12	50

## ANSWERS TO ASSESSMENT ACTIVITY 4.3

1.

Class interval $x$	Frequency $f$	Cumulative frequency $F$
[34 – 44)	4	4
[44 – 54)	4	8
[54 – 64)	12	20
[64 – 74)	1	21
[74 – 84)	12	33
[84 – 94)	5	38
[94 – 104)	2	40
	$\sum f = 40$	

The “halfway” score is in the interval [54 – 64).





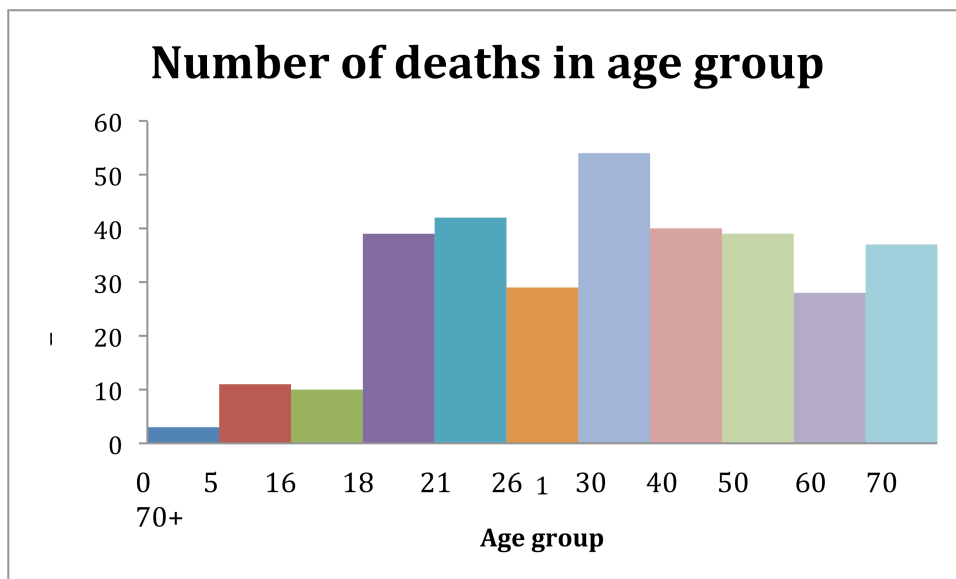
2.  
2.1.

Class interval $x$	Frequency $f$	Cumulative frequency $F$
[10 -12)	5	5
[12 -14)	6	11
[14 - 16)	13	24
[16 -18)	8	32
[18 -20)	3	35
[20 -22)	0	35
[22 -24)	0	35
	$\sum f = 35$	

2.2. The "halfway" score is between 14 and 16 kilogram.

ANSWERS TO ASSESSMENT ACTIVITY 4.4

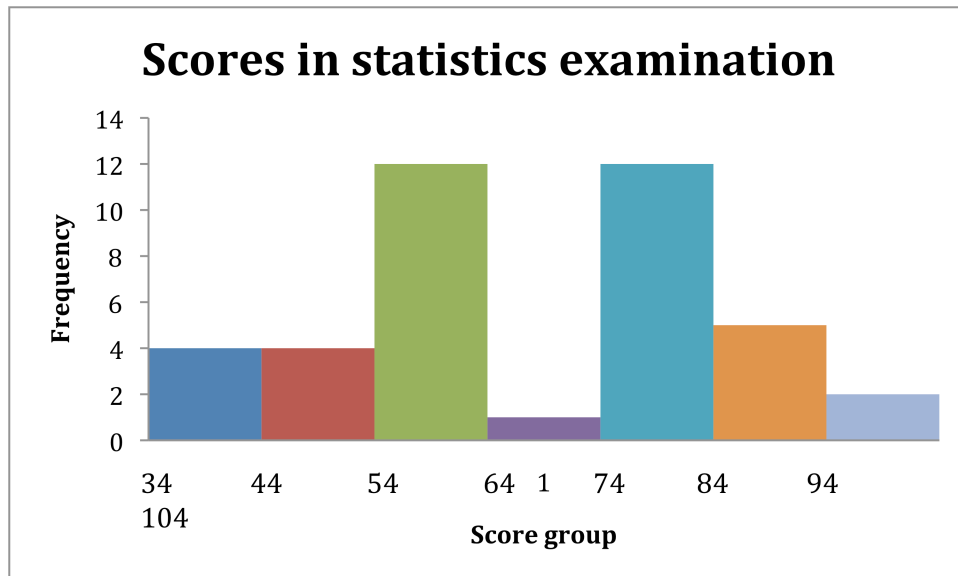
1.



The frequency distribution is skew to the left.



2.



#### ANSWERS TO ASSESSMENT ACTIVITY 4.5

1.

- 1.1. 27,2
- 1.2. 26
- 1.3. 23

2.

- 2.1. 66,95
- 2.2. 65,5
- 2.3. 77

3.

- 3.1. Mean = 5; median = 5; mode = 7
- 3.2. Mean = 22,17; median = 4,5; mode = 4
- 3.3. For the data in 3.2 there is an extreme value

4.

- 4.1. 849785,6
- 4.2. 463554
- 4.3. The best measure of central tendency is the median, because of the extreme values.
- 4.4. 6,9% increase
- 5. The mode is A (adult) and P (parental).
- 6. The mode rate is 3.

#### ANSWERS TO ASSESSMENT ACTIVITY 4.6

1.

- 1.1. 68
- 1.2. 62

2.

- 2.1. 14,89



- 2.2. 15  
2.3. 15,17

#### ANSWERS TO ASSESSMENT ACTIVITY 4.7

1. Skew to the left
2. Skew to the right
3. Symmetric
4.
  - 4.1. Skew to the left
  - 4.2. Median

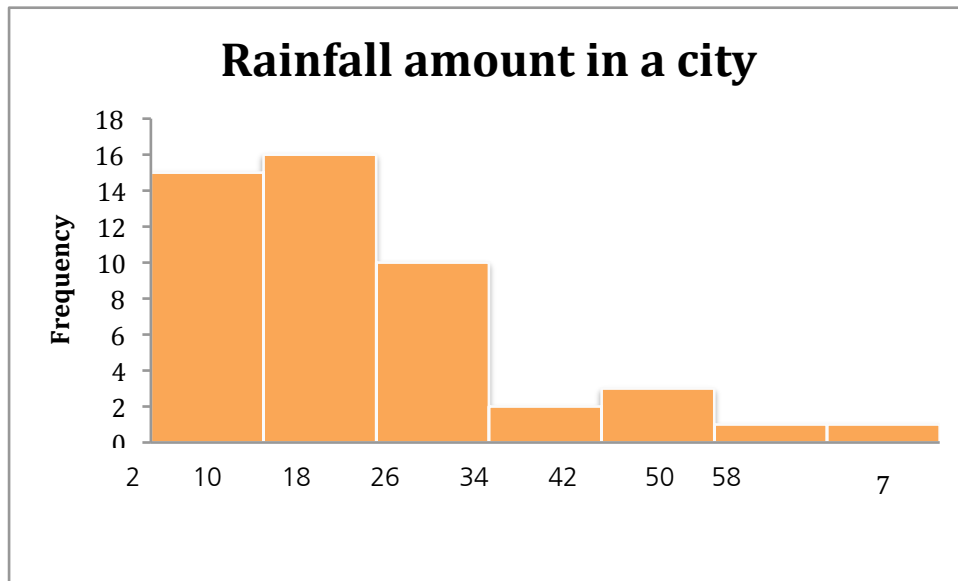
#### ANSWERS TO ASSESSMENT ACTIVITY 4.8

1. Yes I should be. I scored higher (or at least as high) than 90% of the people taking the test.
2. 25,6%
3. 22,7%
4.
  - 4.1. 2,9%
  - 4.2. 16,3
  - 4.3. 13,05
  - 4.4. 15,6
  - 4.5. 15,2 (it is the same)
5.
  - 5.1.

Class interval $x$	Frequency $f$	Cumulative frequency $F$
[2 -10)	15	15
[10 -18)	16	31
[18 - 26)	10	41
[26 - 34)	2	43
[34 - 42)	3	46
[42 - 50)	1	47
[50 - 58)	1	48
	$\sum f = 48$	

- 5.2.





Skew to the right

5.3. 35%

5.4. 85%

5.5. [10,18)

5.6. [2,10)

5.7. 95,1

5.8. 34,5

5.9. 22

5.10. 14,5

5.11. 14,5

5.12. 16,83

5.13. From the information we can also see that it is skew to the right

5.14. Median

6.

6.1.

Class interval $x$	Frequency $f$	Cumulative frequency $F$	Cumulative frequency percentage
[2 -10)	15	15	31,25
[10 -18)	16	31	64,5
[18 - 26)	10	41	85,4
[26 - 34)	2	43	89,6
[34 - 42)	3	46	95,8
[42 - 50)	1	47	97,9
[50 - 58)	1	48	100
	$\sum f = 48$		

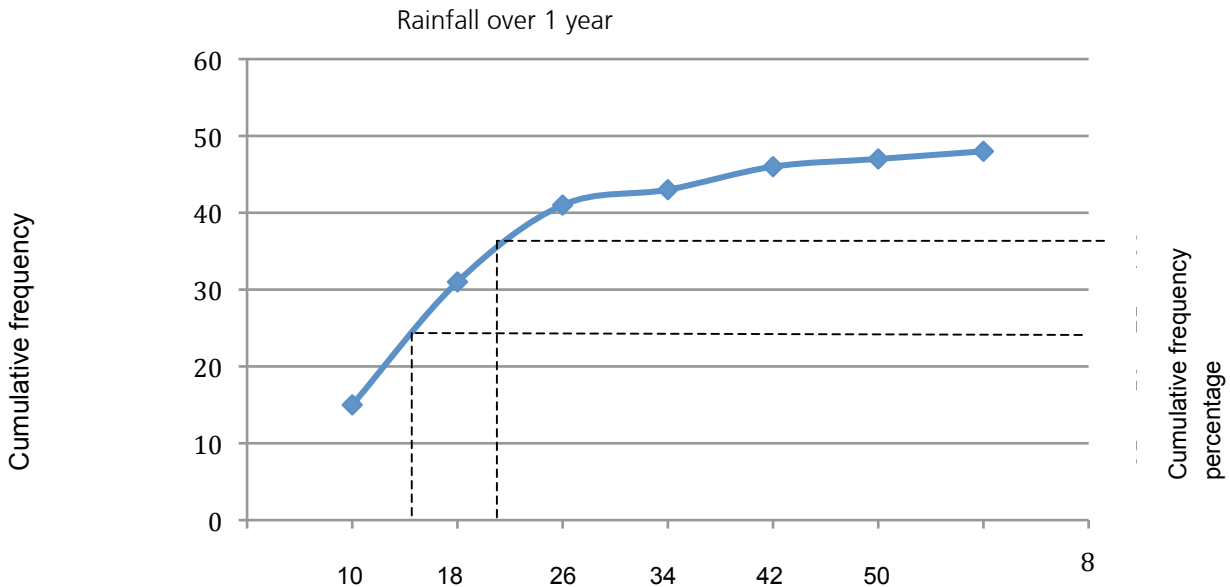
6.2. 12

6.3. 24



6.4. 36

6.5.



6.6. 75% of the measurements are below 22 mm. The same answer as for 5.9.

6.7. 50% of the measurements are below 15 mm. Almost the same answer as for 5.10.

**ANSWERS TO ASSESSMENT ACTIVITY 4.9**

1. Standard deviation = 1,826
2.
  - 2.1. 6,625
  - 2.2. 6,5
  - 2.3. 8
  - 2.4. 5
  - 2.5. 1,685 Since the mean is 6,625, roughly 90% of the scores are between  $6,625 \pm 3 \times 1,685$ .
3. Data set 1: mean = 100; standard deviation = 1,826  
Data set 2: mean = 100; standard deviation = 90,192
4. Peter: standard deviation = 42,217  
Johnny: standard deviation = 11,647  
I would choose Johnny because he has the smaller standard deviation
5. Bakery A: mean = 500; standard deviation = 1  
Bakery B: mean = 501; standard deviation = 3,391  
Bakery A has the best chance of avoiding the court



### ANSWERS TO ASSESSMENT ACTIVITY 4.10

1.

1.1.

Class interval $x$	Frequency $f$	Cumulative frequency $F$
[140 - 146)	4	4
[146 - 152)	6	10
[152 - 158)	15	25
[158 - 164)	16	41
[164 - 170)	25	66
[170 - 176)	18	84
[176 - 182)	13	97
[182 - 188)	3	100
	$\sum f = 100$	

1.2. 41%

1.3. 3%

1.4. [164 – 170)

1.5. [170 – 176)

1.6. 66

1.7. 167,8

1.8. 152

1.9. 174,7

1.10. 166,2

1.11. 165,4

1.12. Symmetric

1.13. Mean

1.14. 45

1.15. 10,12 Since the mean is 165,4 it means that roughly 90% of the boys' heights are in the interval  $165,4 \pm 3 \times 10,12$ .

### ANSWERS TO GROUP ACTIVITY 4.11

1.

1.1.

Column1	
Mean	61.96
Standard Error	1.638607
Median	64
Mode	78
Standard Deviation	16.38607
Sample Variance	268.5034



<b>Kurtosis</b>	0.178237
<b>Skewness</b>	-0.412842
<b>Range</b>	81
<b>Minimum</b>	15
<b>Maximum</b>	96
<b>Sum</b>	6196
<b>Count</b>	100

## 1.2.

Point	Column1	Rank	Percent
6	96	1	100.00%
59	95	2	98.90%
18	93	3	97.90%
11	89	4	95.90%
39	89	4	95.90%
93	87	6	94.90%
84	86	7	93.90%
67	85	8	92.90%
97	84	9	91.90%
49	82	10	90.90%
25	79	11	88.80%
63	79	11	88.80%
3	78	13	83.80%
7	78	13	83.80%
45	78	13	83.80%
70	78	13	83.80%
78	78	13	83.80%
33	76	18	79.70%
35	76	18	79.70%
50	76	18	79.70%
96	76	18	79.70%
44	75	22	77.70%
91	75	22	77.70%
40	74	24	74.70%
68	74	24	74.70%
88	74	24	74.70%
83	73	27	72.70%
87	73	27	72.70%
29	71	29	69.60%
51	71	29	69.60%
76	71	29	69.60%



30	70	32	68.60%
12	69	33	65.60%
24	69	33	65.60%
69	69	33	65.60%
19	68	36	62.60%
31	68	36	62.60%
42	68	36	62.60%
8	67	39	57.50%
22	67	39	57.50%
53	67	39	57.50%
72	67	39	57.50%
86	67	39	57.50%
46	65	44	53.50%
61	65	44	53.50%
89	65	44	53.50%
92	65	44	53.50%
36	64	48	49.40%
57	64	48	49.40%
77	64	48	49.40%
95	64	48	49.40%
28	63	52	45.40%
32	63	52	45.40%
52	63	52	45.40%
62	63	52	45.40%
26	62	56	41.40%
43	62	56	41.40%
47	62	56	41.40%
65	62	56	41.40%
82	61	60	40.40%
27	58	61	39.30%
1	56	62	35.30%
21	56	62	35.30%
54	56	62	35.30%
64	56	62	35.30%
34	55	66	34.30%
37	54	67	29.20%
48	54	67	29.20%
66	54	67	29.20%
73	54	67	29.20%
81	54	67	29.20%
9	53	72	25.20%
13	53	72	25.20%





74	53	72	25.20%
79	53	72	25.20%
56	52	76	22.20%
71	52	76	22.20%
94	52	76	22.20%
99	51	79	21.20%
100	50	80	20.20%
17	49	81	19.10%
10	48	82	17.10%
85	48	82	17.10%
60	47	84	16.10%
14	46	85	13.10%
55	46	85	13.10%
75	46	85	13.10%
5	45	88	12.10%
38	42	89	10.10%
41	42	89	10.10%
90	38	91	9.00%
4	36	92	7.00%
80	36	92	7.00%
58	35	94	6.00%
20	34	95	4.00%
98	34	95	4.00%
2	25	97	3.00%
15	23	98	1.00%
23	23	98	1.00%
16	15	100	0.00%

1.3. 52,75



## Tracking my progress

You have reached the end of this section. Check whether you have achieved the learning outcomes for this section.

LEARNING OUTCOMES	✓ I FEEL CONFIDENT	✓ I DON'T FEEL CONFIDENT
Set up a frequency and cumulative frequency distribution table for grouped and ungrouped data		
Draw a histogram and describe the distribution		
Calculate the mean for grouped and ungrouped data		
Calculate the median for grouped and ungrouped data		
Calculate the mode for grouped and ungrouped data		
Choose the best measure of central tendency		
Calculate the percentile rank for grouped and ungrouped data		
Calculate the score for a given percentile rank for grouped data		
Calculate the score for given quartile or deciles for grouped data		
Draw and interpret an ogive		
Calculate the range		
Calculate the variance and standard deviation for grouped data		
Calculate the variance and standard deviation for ungrouped data		
Use Excel to determine descriptive statistics		



Now answer the following questions honestly:

**1** What did you like best about this section?

---

---

---

**2** What did you find most difficult in this section?

---

---

---

**3** What do you need to improve on?

---

---

---

**4** How will you do this?

---

---

---

