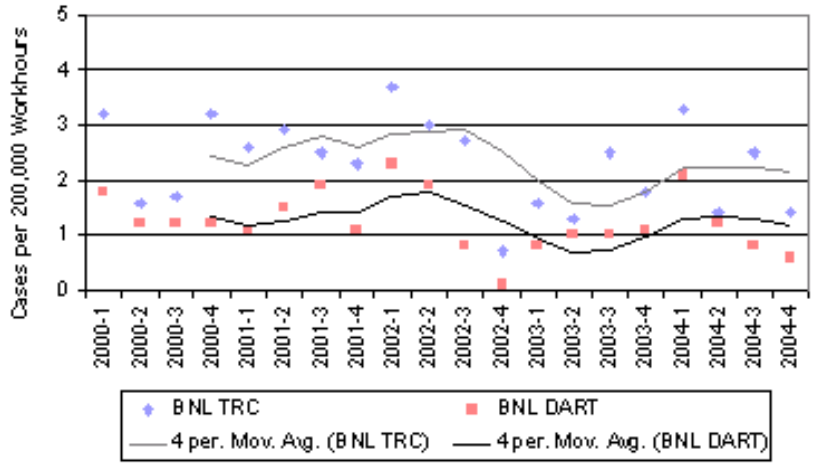
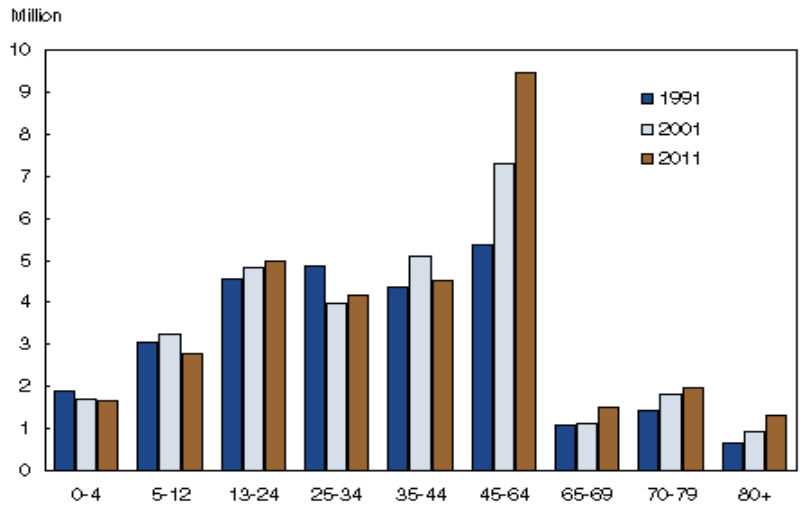
Introduction to Statistics

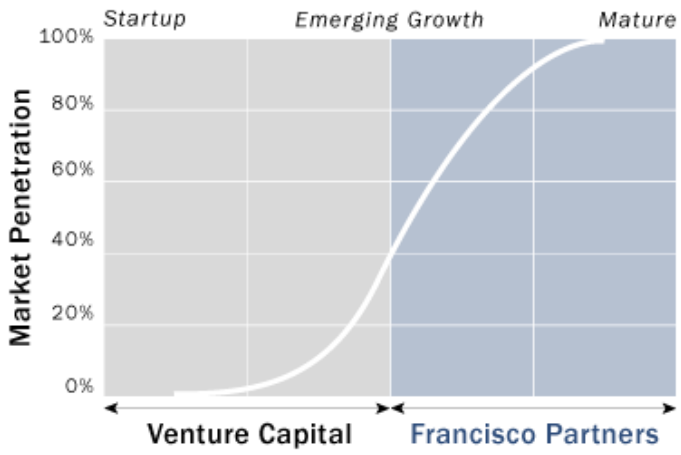


Q: Why do scientists get paid to do science?
A: To take the hundreds of repeat readings needed to prove a theory statistically.



Canadian age profiles. (Statistics Canada. Ref: http://www12.statcan.ca/english/census01/Products/Analytic/companion/age/canada.cfm)

Q: On what do governments base their decisions?
A: To a large extent on Statistics. Population stats, wage stats, welfare stats



Q: On what do commercial companies base their decisions?
A: Viewing stats, advertisment stats, buying stats

Statistics looks at ways of representing data to make groups of numbers more readable to those you want to communicate with.

- ordering and grouping data
- Use of charts such as Histograms
- And graphs such as Ogives
- Leading up to the idea of representing a group of numbers by just two or three numbers
- One of those numbers would be some kind of 'average' or 'middle' number depending on the type and distribution of the data, such as Mean, Median or Mode
- One or two other numbers would then be used to describe the distribution or spread of numbers such as Standard Deviation or Inter-quartile range
- We look at the concept of sampling and Standard Error

Definitions

Variate	the variable that the given data refers to ie. colour, size, velocity, time etc.
Raw data	the values of the variate as obtained ie. not organised in any way
Organised data	data that has been group or sorted in some way
Frequency	the number of times that a particular value for the variate appears in the raw data
Range	the spread of values for the data. This data is determined by subtracting the lowest value of the variate from the highest value within the raw data
Class	the spread of values for a section of the raw data once it has been organised
Population	the complete set of of all the possible values that could be measured or considered
Sample	the subset of values that are measured in a population from which statistical values can be obtained

The *variate* may be continuous or discrete.
A continuous *variate* may have any value within the range.
A discrete *variate* may only have whole number values within the range.
In general, if the *variate* has to be measured in order to determine its value it is continuous, if it has to be counted, it is discrete.

In most cases, the raw data is of little use. It is almost impossible to identify patterns, trends or significant points from blocks of figures. The raw data must be organised in some way to make it useful. There are many ways in which this can be done, one of the most useful being the frequency distribution table (FDT) which we will discuss later. Now we look at working out the first number that can represent a set of data values, namely the average.

Mean, Median and Mode

The *mean*, *median* and *mode* are all forms of average. Each is used depending on the nature of the data presented and the type of average required.

Mean

The *mean* is the mathematical average of a given set of data. To determine the value of the *mean*, the total of all data values is determined, this total is then divided by the number of individual data values in the data set.

E.g. Determine the *mean* of 1, 5, 7, 8, 9, 12

total value = 42
no of values = 6
mean $= \frac{42}{6} = 7$

The *mean* is used because it can be calculated accurately, and gives a useful figure for further statistical calculations. However, it is only suitable if there are no extreme values within the data set or if the number of data values is large. The *mean* does not always give an acceptable average figure, as it is sensitive to any extreme values.

E.g. Determine the *mean* of 1, 5, 7, 8, 9, 12, 98

total value = 140
no of values = 7
mean $= \frac{140}{7} = 20$

But, 20 is a much higher value than six out of the seven individual data values in the data set, so is not representative of the set as a whole. When the set has an extreme value (98) a more acceptable average figure would be given by the *median*.

Median

When the raw data is given, it must first be reorganised, in order (up or down), this is then called an array. The *median* is the middle figure of an array.

E.g. Determine the *median* of 8, 7, 1, 5, 12, 98, 9
First reorganise into an array 1, 5, 7, **8**, 9, 12, 98
8 is the middle figure therefore is the *median*.

This is a better estimate of the 'average' for **most** of the numbers in the array. If there is an even number of data values in the data set, then the *median* is the *mean* of the middle two.

E.g. Determine the *median* of 3, 3, 4, **5, 7**, 9, 11, 12, $\frac{5+7}{2} = 6$

So 6 is the median for this array, even though the figure six is not one of the actual data values in the original array.

Mode

Even the median does not always give a value that best represents a particular situation. For example, a survey was carried out into the number of people living in each house on a certain estate, the results were as follows:

Number in house	Number of houses	Total people
1	5	5
2	19	38
3	28	84
4	40	160
5	8	40
	Total 100	Total 327

This gives a total of 100 houses and 327 people. The *mean* of people per house would be 3.27, the *median* of people per house would be 3, but there were far more houses with 4 people than with any other number, so the best 'average' in this case is the most common number, 4. The *mode* is defined as the most common number.

E.g. Determine the *mode* for the following data set 3, 5, 7, 5, 4, 3, 6, 8, 3, 6
The figure 3 occurs more often than any other number, so 3 is the *mode*.

It is possible for a set data to have one *mode* (uni-modal), to have more than one *mode* (bio-modal or tri-modal) or to have no *mode* at all.

It is common practice to give data sets in the form of a table

E.g

Value (£)	£100	£120	£140	£160	£180	£200
Number	5	7	10	15	12	9

The value of the *mode* is very easy to determine, simply look along the values to see which has the highest number - £160 has 15, so *modal* value is £160.

Median value is also fairly easy – count up how many values there are in all
5 + 7 +10 +15 +12 +9 = 58 the *median* value will be at the halfway point i.e. the 29th value. The 29th value occurs in the 15 block, so median value is also £160.

The *mean* value will require more working out. From the *median* we already know that there are 58 different values, the total amount of money must now be calculated.

E.g.

100 x 5 = 500 120 x 7 = 840 140 x 10 = 1400
160 x 15 = 2400 180 x 12 = 2160 200 x 9 = 1800

$$\text{Total} = \frac{9100}{58}$$

Mean = £156.90

Frequency distribution tables

To construct a frequency distribution table (FDT), the raw data is sorted into small groups or classes, each group covering the same spread of values. The class can be of any size but in practice 2, 5 or their multiples are nearly always used. The method is best explained using an example:

The following raw data shows the marks achieved in an examination by fifty students

66	22	11	28	31	35	58	42	65	34
22	46	51	47	58	45	51	67	38	47
31	24	48	56	93	17	49	53	58	68
69	51	62	76	36	55	88	55	43	74
43	54	35	58	42	68	37	76	26	52

The first step to identify the range Highest value = 93, lowest value = 11

Range = 93 – 11 = 82

Classes must begin at a convenient value at or below the lowest value in the *range*, continue to at or above the highest value in the range and be sufficient in number to make a reasonable FDT (never less than 5 nor more than 20). To determine the class size, divide the *range* by the proposed spread of values, for a reasonable FDT the result should be somewhere between 6 and 12. In the example given above, suppose the classes of 5 were considered, this would give a

$$\frac{\textit{range}}{\text{proposed class size}} = \frac{82}{5} = 16.4$$

This result does not lie between 6 and 12 so would give a table with too many classes. If classes of 10 were considered, this would give $\frac{82}{10} = 8.2$

This result does lie between 6 and 12 so classes with a spread of 10 will give a FDT with a suitable number of classes. The next step is to decide on the starting point for the classes. The lowest value is 11 and if each class is to have a spread of 10 values it seems reasonable to start from ten. The classes will then be 10-19, 20-29, 30-39, 40-49 etc. The classes must continue until the last class covers the highest value in the raw data.

The raw data must then be scanned and each figure placed in the appropriate class so that the frequencies can be established. The most convenient method is the tally system:

Class	Tally	Frequency
10-19	II	2
20-29	IIII	4
30-39	IIIII IIII	9
40-49	IIIII IIIII	10
50-59	IIIII IIIII III	13
60-69	IIIII II	7
70-79	III	3
80-89	I	1
90-99	I	1

The frequency distribution table can be used in several ways once constructed.

In this example the raw data consisted of whole numbers only, so there is no doubt which class each value in the raw data belongs in. If the raw data includes figures with decimal points, the limit for each class must be established. In the same way, each class has recognised boundaries and limits, for example, for class 3 in the FDT shown above

Upper boundary	$B_u = 29$	lower boundary	$B_l = 20$
Upper limit	$L_u = 29.4$	lower limit	$L_l = 19.5$

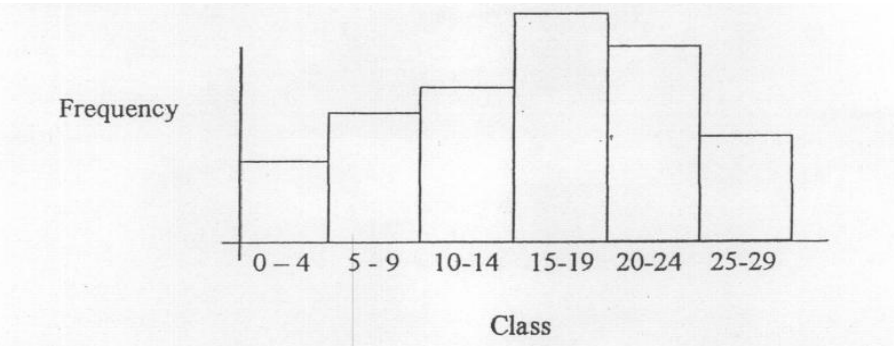
The numerical difference between the upper and lower boundaries is called the class width or class interval.

Histograms

A frequency distribution table is produced from a set of raw data. The FDT can then be used to produce a histogram – a bar graph of classes plotted against frequency.

Eg.

Class	0-4	5-9	10-14	15-19	20-24	25-29
Frequency	3	5	6	10	8	4

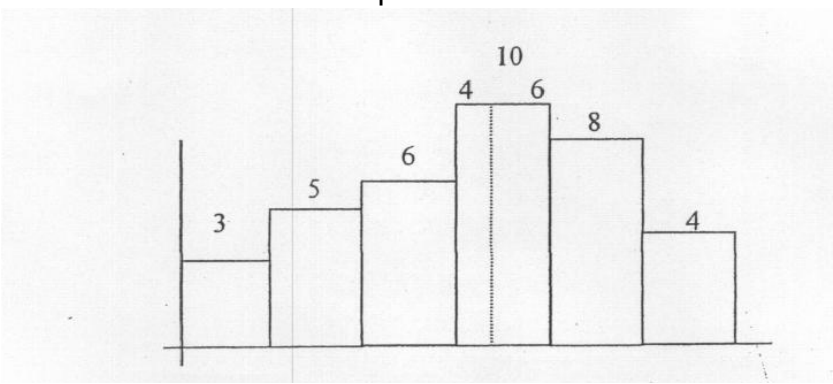


Once the histogram has been constructed, it can be used to determine the value of the *median* and *mode*.

To determine the *median* value

The median is the middle value, or the value that has as many frequencies above as below. In the example above, the total frequencies = 3 + 5 + 6 +10 + 8 + 4 = 36

The *median* will therefore occur where 18 frequencies lie to the left and 18 lie to the right



The median frequency will occur in the 15-19 class

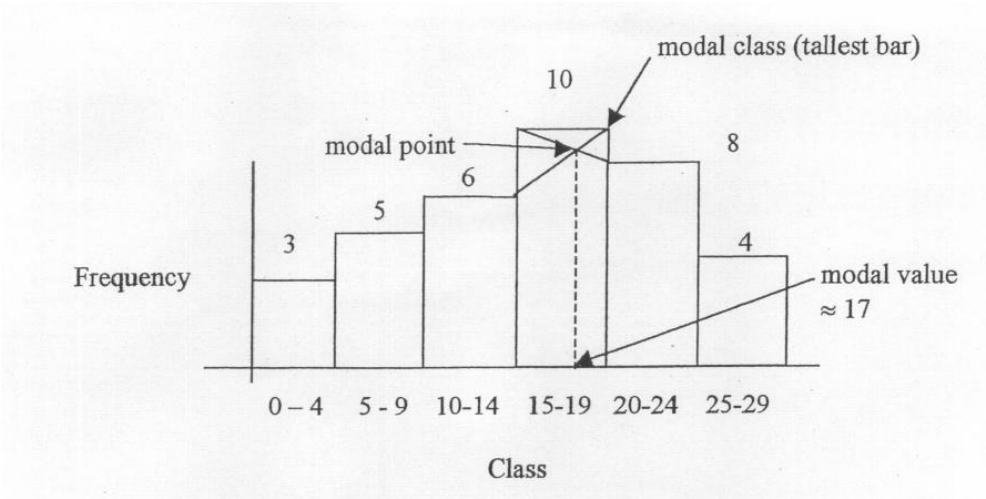
3 + 5 + 6 + $\boxed{4 \mid 6}$ + 8 + 4

To determine the *median* value, the 10 frequencies in the 15-19 class must be divided in the ratio of 4:6, so the class width or interval must be divided in the same ratio.

Class interval = 19 – 15 = 4 and as there are 10 frequencies, each one represents 0.4 of the interval. The interval is divided in the ratio of 4:6

Median value therefore = either 15 + (4 x 0.4) = 16.6
or 19 – (6 x 0.4) = 16.6

To determine *modal* value



The *modal* class is easily identified as it will always be the tallest bar. If an approximate value is acceptable it can be estimated from the graph baseline by drawing straight lines from each of the adjoining class barriers to the opposite corner of the *modal* class. Where these two lines intersect identifies the *modal* point, its corresponding value can be read from the base line.

If a more accurate value is required, the following formula can be used

$$\textit{modal}\ \text{value} = L + \frac{D_1}{D_1 + D_2} \times W$$

Where
- L = lower boundary value of *modal* class
- D_1 = difference in frequency between *modal* class and one class below
- D_2 = difference in frequency between *modal* class and one class above
- W = class width (or interval)

In the example above L = 15 D_1 = 10 – 6 = 4 D_2 = 10 – 8 = 2 W = 19 – 15 = 14

$$\textit{modal}\ \text{value} = 15 + \frac{4}{4+2} \times 4 = 15 + 2.67 = 17.67$$

Introduction to Statistics - Worksheet 1

1. Determine the median value for the following sets of data

   (a) £24, £36, £19, £43, £28, £29, £31
   (b) 84, 76, 39, 47, 81, 56, 73, 62
   (c) 61.8, 63.2, 64.5. 61.2, 64.1, 85.9, 65.9, 62.3, 63.6

2. Calculate the mean, median and modal wage for the following casual workers:
   8 workers earn £76.50 per week, 7 earn £82.40 per week and 5 earn £83.60 per week.

Produce a frequency distribution table for the following set of raw data and from that construct a histogram. For each histogram determine, as accurately as possible the *Median* and *Modal* values.

3. The following table shows the number of industrial accidents reported each week over a period of 80 weeks.

52	43	61	29	50	28	56	36	53	70	95	59	64
60	51	81	35	38	22	23	86	57	81	92	69	48
43	67	72	60	70	65	91	44	67	33	66	43	29
34	61	75	43	45	74	64	33	73	59	60	66	89
41	46	75	48	57	89	55	72	87	39	84	76	44
87	91	39	90	49	73	55	50	58	97	24	57	50
87	73											

Introduction to Statistics - Worksheet 2

4. The following table shows the number of rejects produced by a factory,for 15 weeks

week no	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
rejects	7	55	96	38	17	55	5	49	28	83	72	66	23	41	14

   Calculate the mean weekly rejects. If the acceptable mean rejects is 45, how many more rejects could have been made over the 15 weeks and the target still met?

5. Calculate the mean, median and modal values for each of the following sets of data

(a)	diameter	14.96	14.97	14.98	14.99	15.00	15.01	15.02
	number	3	5	13	21	26	24	8
(b)	length	29.5	29.6	29.7	29.8	29.9	30.0	30.1	30.2
	number	3	7	22	28	18	12	7	3

Produce a frequency distribution table for the following set of raw data and from that construct a histogram. For each histogram determine, as accurately as possible the *Median* and *Modal* values.

6. The cutting life, in minutes, of 50 identical drill bits was as follows

150	517	464	444	466	289	358	532	204	418
436	196	408	262	240	403	332	323	491	383
342	568	218	213	510	397	453	581	459	396
173	520	254	241	164	555	369	437	428	350
292	284	372	289	341	388	365	303	322	310