"JUST THE MATHS"

UNIT NUMBER

18.2

STATISTICS 2
(Measures of central tendency)

by

A.J.Hobson

UNIT 18.2 - STATISTICS 2

MEASURES OF CENTRAL TENDENCY

18.2.1 INTRODUCTION

Having shown, in Unit 18.1, how statistical data may be presented in a clear and concise form, we shall now be concerned with the methods of analysing the data in order to obtain the maximum amount of information from it.

In the previous Unit, it was stated that statistical problems may be either "descriptive problems" (in which all the data is known and can be analysed) or "inference problems" (in which data collected from a "sample" population is used to infer properties of a larger population).

In both types of problem, it is useful to be able to measure some value around which all items in the data may be considered to cluster. This is called **"a measure of central tendency"**; and we find it by using several types of average value as follows:

18.2.2 THE ARITHMETIC MEAN (BY CODING)

To obtain the **"arithmetic mean"** of a finite collection of n numbers, we may simply add all the numbers together and then divide by n. This elementary rule applies even if some of the numbers occur more than once and even if some of the numbers are negative.

However, the purpose of this section is to introduce some short-cuts (called **"coding"**) in the calculation of the arithmetic mean of large collections of data. The methods will be illustrated by the following example, in which the number of items of data is not over-large:

EXAMPLE

The solid contents, x, of water (in parts per million) was measured in eleven samples and the following data was obtained:

4520	4490	4500	4500
4570	4540	4520	4590
4520	4570	4520	

Determine the arithmetic mean, $\overline{x}$, of the data.

Solution

(i) Direct Calculation

By adding together the eleven numbers, then dividing by 11, we obtain

$$\overline{x} = 49840 \div 11 \simeq 4530.91$$

(ii) Using Frequencies

We could first make a frequency table having a column of distinct values x_i, $(i = 1, 2, 3, ..., 11)$, a column of frequencies f_i, $(i = 1, 2, 3....., 11)$ and a column of corresponding values $f_i x_i$.

The arithmetic mean is then calculated from the formula

$$\overline{x} = \frac{1}{11} \sum_{i=1}^{11} f_i x_i.$$

In the present example, the table would be

x_i	f_i	$f_i x_i$
4490	1	4490
4500	2	9000
4520	4	18080
4540	1	4540
4570	2	9140
4590	1	4590
	Total	49840

The arithmetic mean is then $\overline{x} = 49840 \div 11 \simeq 4530.91$ as before.

(iii) Reduction by a constant

With such large data-values, as in the present example, it can be convenient to reduce all of the values by a constant, k, before calculating the arithmetic mean.

It is easy to show that, by adding the constant, k, to the arithmetic mean of the reduced data, we obtain the arithmetic mean of the original data.

Proof:

For n values, $x_1, x_2, x_3, ..., x_n$, the arithmetic mean is given by

$$\overline{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}.$$

If each value is reduced by a constant, k, the arithmetic mean of the reduced data is

$$\frac{(x_1 - k) + (x_2 - k) + (x_3 - k) + ... + (x_n - k)}{n} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} - \frac{nk}{n} = \overline{x} - k.$$

(iv) Division by a constant

In a similar way to the previous paragraph, each value in a collection of data could by divided by a constant, k, before calculating the arithmetic mean.

This time, we may show that the arithmetic mean of the original data is obtained on multiplying the arithmetic mean of the reduced data by k.

Proof:

$$\frac{\frac{x_1}{k} + \frac{x_2}{k} + \frac{x_3}{k} + ... + \frac{x_n}{k}}{n} = \frac{\overline{x}}{k}.$$

In order to summarise the shortcuts used in the present example, the following table shows a combination of the use of frequencies and of the two types of reduction made to the data:

x_i	$x_i - 4490$	$x_i' = (x_i - 4490) \div 10$	f_i	$f_i x_i'$
4490	0	0	1	0
4500	10	1	2	2
4520	30	3	4	12
4540	50	5	1	5
4570	80	8	2	16
4590	100	10	1	10
			Total	45

$$\text{The fictitious arithmetic mean, } \overline{x}' = \frac{45}{11} \simeq 4.0909$$

$$\text{The actual arithmetic mean, } \overline{x} \simeq (4.0909 \times 10) + 4490 \simeq 4530.91$$

(v) The approximate arithmetic mean for a grouped distribution

For a large number of items of data, we may (without losing too much accuracy) take all items within a class interval to be equal to the class mid-point.

A calculation similar to that in the previous paragraph may then be performed if we reduce each mid-point by the first mid-point and divide by the class width (or other convenient number).

EXAMPLE

Calculate, approximately, the arithmetic mean of the data in TABLE 4 on page 4 of Unit 18.1

Solution

Class Interval	Class Mid-point x_i	$x_i - 9$	$(x_i - 9) \div 3 = x_i'$	Frequency f_i	$f_i x_i'$
$7.5 - 10.5$	9	0	0	12	0
$10.5 - 13.5$	12	3	1	10	10
$13.5 - 16.5$	15	6	2	15	30
$16.5 - 19.5$	18	9	3	19	57
$19.5 - 22.5$	21	12	4	12	48
$22.5 - 25.5$	24	15	5	14	70
$25.5 - 28.5$	27	18	6	3	18
$28.5 - 31.5$	30	21	7	0	0
$31.5 - 34.5$	33	24	8	4	32
$34.5 - 37.5$	36	27	9	1	9
			Totals	90	274

$$\text{Fictitious arithmetic mean } \overline{x}' = \frac{274}{90} \simeq 3.0444$$

$$\text{Actual arithmetic mean} = 3.044 \times 3 + 9 \simeq 18.13.$$

Notes:

(i) By direct calculation from TABLE 1 in Unit 18.1, it may be shown that the arithmetic mean is 17.86 correct to two places of decimals; and this indicates an error of about 1.5%.

(ii) The arithmetic mean is widely used where samples are taken of a larger population. It

usually turns out that two samples of the same population have arithmetic means which are close in value.

18.2.3 THE MEDIAN

Collections of data often include one or more values which are widely out of character with the rest; and the arithmetic mean can be significantly affected by such extreme values.

For example, the values 8,12,13,15,21,23 have an arithmetic mean of $\frac{92}{6} \simeq 15.33$; but the values 5,12,13,15,21,36 have an arithmetic mean of $\frac{102}{6} \simeq 17.00$.

A second type of average, not so much affected, is defined as follows:

DEFINITION

The **"median"** of a collection of data is the middle value when the data is arranged in rank order. For an even number of values in the collection of data, the median is the arithmetic mean of the centre two values.

EXAMPLES

1. For both 8,12,13,15,21,23 and 5,12,13,15,21,36, the median is given by

$$\frac{13+15}{2} = 14.$$

2. For a grouped distribution, the problem is more complex since we no longer have access to the individual values from the data.

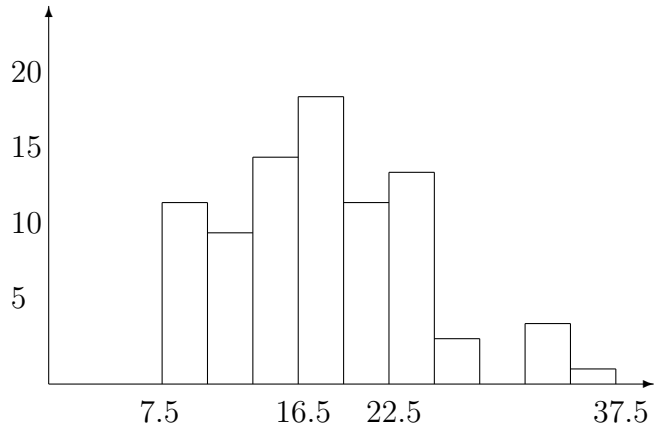
   However, the area of a histogram is directly proportional to the total number of values which it represents, since the base of all the rectangles are the same width and each height represents a frequency.

   We may thus take the median to be the value for which the vertical line through it divides the histogram into two equal areas.

   For non-symmetrical histograms, the median is often a better measure of central tendency than the arithmetic mean.

   Illustration

   Consider the histogram from Unit 18.1, representing rainfall figures over a 90 year period.



The total area of the histogram $= 90 \times 3 = 270$.

Half the area of the histogram $= 135$.

The area up as far as $16.5 = 3 \times 37 = 111$ while the area up as far as $19.5 = 3 \times 56 = 168$; hence the Median must lie between 16.5 and 19.5

The Median $= 16.5 + x$ where $18x = 135 - 111 = 24$ since 18 is the frequency of the class interval $16.5 - 19.5$

That is,

$$x = \frac{24}{18} = \frac{4}{3} \simeq 1.33,$$

giving a Median of 17.83

Notes:

(i) The median, in this case, is close to the arithmetic mean since the distribution is fairly symmetrical.

(ii) If a sequence of zero frequencies occurs, it may be necessary to take the arithmetic mean of two class mid-points, which are not consecutive to each other.

(iii) Another example of the advantage of median over arithmetic mean would be the average life of 100 electric lamps. To find the arithmetic mean, all 100 must be tested; but to find the median, the testing may stop after the 51st.

18.2.4 THE MODE

DEFINITIONS

1. For a collection of individual items of data, the **"mode"** is the value having the highest frequency.
2. In a grouped frequency distribution, the mid-point of the class interval with the highest frequency is called the **"crude mode"** and the class interval itself is called the **"modal class"**.

Note:
Like the median, the mode is not much affected by changes in the extreme values of the data. However, some distributions may have several different modes, which is a disadvantage of this measure of central tendency.

EXAMPLE

For the histogram discussed earlier, the mode is 18.0; but if the class interval, $22.5 - 25.5$, had 5 more members, then 24.0 would be a mode as well.

18.2.5 QUANTILES

To conclude this Unit, we shall define three more standard measurements which, in fact, extend the idea of a median; and we may recall that a median divides a collection of values in such a way that half of them fall on either side of it.

Collectively, these three new measurements are called **"quantiles"** but may be considered separately by their own names as follows:

(a) Quartiles

These are the three numbers dividing a ranked collection of values (or the area of a histogram) into 4 equal parts.

(b) Deciles

These are the nine numbers dividing a ranked collection of values (or the area of a histogram) into 10 equal parts.

(c) Percentiles

These are the ninety nine numbers dividing a ranked collection of values (or the area of a histogram) into 100 equal parts.

Note:
For collections of individual values, quartiles may need to be calculated as the arithmetic mean of two consecutive values.

EXAMPLES

1. (a) The 25th percentile = The 1st quartile.
   (b) The 5th Decile = The median.
   (c) The 85th Percentile = the point at which 85% of the values fall below it and 15% above it.
2. For the collection of values

$$5, 12, 13, 19, 25, 26, 30, 33,$$

   the quartiles are 12.5, 22 and 28.
3. For the collection of values

$$5, 12, 13, 19, 25, 26, 30,$$

   the quartiles are 12.5, 19 and 25.5

18.2.6 EXERCISES

1. The arithmetic mean of 75 observations is 52.6 and the arithmetic mean of 25 similar observations is 48.4; determine the Arithmetic Mean of all 100 observations.
2. Of 500 students, whose mean height is 67.8 inches, 150 are women. If the mean height of 150 women is 62.0 inches, what is the mean height of the men ?
3. By coding the following collection of data, determine the arithmetic mean correct to three places of decimals:

$$1.847, 1.843, 1.842, 1.847, 1.848, 1.841, 1.845$$

4. Using a histogram of the frequency distribution shown, determine
   (a) the arithmetic mean;
   (b) the median class;
   (c) the median;

(d) the modal class;

(e) the crude mode.

class interval	$15-25$	$25-35$	$35-45$	$45-55$	$55-65$	$65-75$
Frequency	4	11	19	14	0	2

5. The number of a certain component issued, per day, from stock over a 40 day period is given as follows:

83	80	91	81	88	82	87	97	83	99
75	85	72	92	84	90	87	78	93	98
86	80	93	86	88	83	82	101	89	82
85	95	80	89	84	92	76	81	103	94

Using class intervals $70-75$, $75-80$, $80-85$ etc., draw up a frequency distribution table and construct a histogram.

From the histogram, determine the median and the 7th Decile.

18.2.7 ANSWERS TO EXERCISES

1. $$51.55$$

2. $$70.3 \text{ inches.}$$

3. $$1.845$$

4. (a) $$40.20$$

   (b) $$35-45.$$

   (c) $$40.26$$

   (d) $$35-45.$$

   (e) $$40.$$

5. $$\text{Median} = 86.5, \quad 7\text{th Decile} = 91.25$$