Section 8 (Cont'd)
Measures of Central Tendency

In this section we continue our analysis of the basics of statistics with an examination of the measures used to determine the most probable value, the most common value and the value in the middle of our data. From these we define the measures associated with sub-groups of the data. We also introduce the coding method; the method you will adopt in your examination questions.

Averages and measures of central tendency

An *average* is usually a value which is typical of the set of data in question. As such values tend to the center of data sets they are referred to as *measures of central tendencies*. The several averages we'll deal with here are *arithmetic mean, median,* and *mode*.

Arithmetic Mean

The arithmetic mean or *mean* of a data set containing N numbers $X_1, X_2, \ldots, X_N$ is denoted by $\bar{X}$ defined as

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

For numbers which have frequencies associated with them in a data set denoted f_i then

$$\bar{X} = \frac{f_1 X_1 + f_2 X_2 + \ldots + f_K X_K}{N} = \frac{\sum_{i=1}^{K} f_i X_i}{N}$$

where $N = \sum_{i=1}^{K} f_i$. Note that K obviously has to be less than N.

Weighted Arithmetic Mean

When dealing with some data sets we sometimes attribute higher importance to some numbers over others. This may be due to

higher frequencies or some other factor. Such attributes are called *weighting factors*, w_i, and give rise to a mean very closely related to the frequency sensitive mean above.

$$\bar{X} = \frac{w_1 X_1 + w_2 X_2 + \ldots + w_K X_K}{w_1 + w_2 + \ldots + w_K} = \frac{\sum_{i=1}^{K} w_i X_i}{\sum_{i=1}^{K} w_i}$$

Note that the frequency sensitive mean is contained within the definition of the weighted mean defined here.

Properties of the Arithmetic Mean

(i) The algebraic sum of the deviations of a data set of numbers about their mean is zero.

(ii) The sum of the squares of the deviations of a data set $\{X_i\}$ about a number *a* is a minimum if $a = \bar{X}$.

(iii) If f_1 numbers have mean m_1, f_2 numbers have mean m_2, … , f_K numbers have mean m_K then the mean of all the numbers, $\bar{X}$, is

$$\bar{X} = \frac{f_1 m_1 + f_2 m_2 + \ldots + f_K m_K}{f_1 + f_2 + \ldots + f_K} = \frac{\sum_{i=1}^{K} f_i X_i}{\sum_{i=1}^{K} f_i}$$

(iv) If *A* is any *guessed* or *assumed arithmetic mean* and if $d_i = X_i - A$ are the deviations of the X_i from *A* then the mean can be expressed

$$\bar{X} = A + \frac{\sum_{i=1}^{N} d_i}{N} \quad \text{or} \quad \bar{X} = A + \frac{\sum_{i=1}^{K} f_i d_i}{N}$$

where $N = \sum_{i=1}^{K} f_i$. The latter formula is used for data X_i that have associated frequencies f_i.

Arithmetic Mean from Group Data

For data in a frequency distribution the values in a class interval are considered coincident with the midpoint or class mark of the interval. If we interpret X_i as the class mark (in the previous weighted equations for $\bar{X}$), f_i its corresponding frequency, A any guessed or assumed class mark and $d_i = X_i - A$ the deviations of X_i from A then the previous equations of $\bar{X}$ are valid:

$$\bar{X} = \frac{f_1 X_1 + f_2 X_2 + \ldots + f_K X_K}{f_1 + f_2 + \ldots + f_K} = \frac{\sum_{i=1}^{K} f_i X_i}{N}$$

$$\bar{X} = A + \frac{\sum_{i=1}^{K} f_i d_i}{\sum_{i=1}^{K} f_i}$$

If the class intervals are the same size c then the deviation can be expressed as $c\,u_i$ where u_i can be positive or negative integers or zero; i.e. 0, $\pm 1, \pm 2, \cdots$ and the latter formula above becomes

$$\bar{X} = A + \left(\frac{\sum_{i=1}^{K} f_i u_i}{N} \right) c$$

which is equivalent to the equation $\bar{X} = A + c\,\bar{u}$. This is called the *coding method* for computing means. It should always be used for grouped data when class intervals are equal.

Median

The median of a set of numbers arranged in order of magnitude (in an array) is defined to be the arithmetic mean of the two middle values. For grouped data the median is given by

$$\text{median} = L_1 + \left(\frac{\frac{N}{2} - \left(\sum f \right)_1}{f_{median}} \right) c$$

where

L_1 = lower class boundary of the class containing the median
(i.e. the median class).

N = number of data elements.

$\left(\sum f\right)_1$ = sum of all frequencies in classes lower than the median class.

f_{median} = frequency of the median class

c = size of the median class interval.

The median can be considered geometrically as that vertical line which divides the histogram into two equal parts. The median is often denoted by $\tilde{X}$.

Mode

The mode of a set of numbers is that number whose frequency is the largest (the most common value). Unique modes are not absolute (i.e. if a mode exists then it may not be unique). Modes don't have to exist. A distribution with one mode is *unimodal*, that with two modes is *bimodal* while that of many modes is called *multi-modal.* In the case for grouped data where a frequency curve has been fitted to the data then the mode is the maximum point on the curve. The point is denoted as $\hat{X}$. For a frequency distribution or histogram then the mode can be calculated from

$$\text{Mode} = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)c$$

where

L_1 = lower class boundary of the class containing the mode
(i.e. the modal class).

Δ_1 = excess of modal frequency over that of next lower class.

Δ_2 = excess of modal frequency over that of next higher class.

c = size of the modal class interval.

Root Mean Square

The root mean square (r.m.s) or *quadratic mean* of a set of numbers is sometimes denoted by $\sqrt{\overline{X}^2}$ and is defined by

$$\text{r.m.s.} = \sqrt{\overline{X}^2} = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N}}$$

This is normally used in physical applications (i.e. mains voltage, current etc.)

Quartiles, Deciles & Percentiles

We saw how the median was defined as the value which split the set into two equal parts (the set is arranged in increasing or decreasing order of magnitude i.e. an array). If we decide to split the set into more than two equal parts we could split it into fourths (*quartiles*), tenths (*deciles*) or even into hundredths (*percentiles*).

The values of the set corresponding to the splits are denoted

- $Q_1, Q_2,$ and Q_3 with Q_2 equal to the median for the quartiles
- $D_1, D_2, \ldots, D_9$ for deciles
- $P_1, P_2, \ldots, P_{99}$ for percentiles.

P_{25}, P_{50} and P_{75} correspond to $Q_1, Q_2,$ and Q_3 respectively. D_5 corresponds to the median as does P_{50}. These are analogous to partitioning the array as you would an interval in real analysis.