Section 8 (Cont'd)
Measures of Deviation

The extent to which numerical data tend to spread about an average value is called the *dispersion* or *variation* of the data. The various measures we will deal with in this section are the range, mean deviation, semi-interquartile range, 10-90 percentile range and the standard deviation.

Range

The range of a set of numbers is the difference between the smallest and largest number in the data set.

Mean Deviation (or Average Deviation)

The *mean deviation (or average deviation)* of a data set containing N numbers $X_1, X_2, \ldots, X_N$ is defined as

$$\text{M.D.} = \frac{|X_1-\bar{X}|+|X_2-\bar{X}|+\ldots+|X_N-\bar{X}|}{N}$$

$$= \frac{\sum_{i=1}^{N}|X_i-\bar{X}|}{N} = \frac{\sum|X_i-\bar{X}|}{N} = \overline{|X-\bar{X}|}$$

where $\bar{X}$ is the arithmetic mean of the numbers and $|X_i-\bar{X}|$ is the absolute value between the data element X_i and the mean $\bar{X}$. If the numbers $X_1, X_2, \ldots, X_K$ occur with frequencies $f_1, f_2, \ldots, f_K$ respectively then the mean deviation can be written as

$$\text{M.D.} = \frac{\sum_{i=1}^{N} f_i|X_i-\bar{X}|}{N} = \frac{\sum f_i|X_i-\bar{X}|}{N} = \overline{|X-\bar{X}|}$$

where $N=\sum_{i=1}^{K} f_i$. Note that K obviously has to be less than N. This form is useful when the data is grouped

with the X_i's representing class marks and the f_i's their associated frequencies.

Sometimes the mean deviation is defined in terms of the absolute deviation about the median (or some other average) and not the mean. The interesting property about the summation $\sum_{i=1}^{N} |X_i - a|$ is that it is a minimum when *a* corresponds to the median. Thus the mean deviation about the median is a minimum.

Semi-Interquartile Range or Quartile Deviation

The semi-interquartile range of a set of data is defined as

$$\text{Semi-Interquartile Range SIQR} = Q = \frac{Q_3 - Q_1}{2}$$

Where Q_1 and Q_3 are the first and third quartiles for the data.

10-90 Percentile Range

This range for a set of data is defined to be

$$\text{10-90 Percentile Range} = P_{90} - P_{10}$$

where P_{10} and P_{90} are the tenth and ninetieth percentiles for the data.

Standard Deviation

The standard deviation for a set of numbers consisting of *N* elements $X_1, X_2, \ldots, X_N$ is denoted by *s* and defined to be

$$s = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}{N}} = \sqrt{\frac{\sum \left(X_i - \bar{X}\right)^2}{N}} = \sqrt{\frac{\sum x^2}{N}}$$

where x is defined to be the deviations of the X_i about the mean $\bar{X}$. Thus s represents the root mean square of the deviations from the mean or, as is often called, the *root mean square deviation*.

If the numbers $X_1, X_2, \ldots, X_K$ occur with frequencies $f_1, f_2, \ldots, f_K$ respectively then the standard deviation can be re-written as

$$s = \sqrt{\frac{\sum_{i=1}^{K} f_i \left(X_i - \bar{X}\right)^2}{N}} = \sqrt{\frac{\sum f_i \left(X_i - \bar{X}\right)^2}{N}} = \sqrt{\frac{\sum f x^2}{N}}$$

where $N = \sum_{i=1}^{K} f_i$. This latter form is useful for grouped data.

In some situations the N above in the standard deviation definition is replaced by $(N-1)$ because the resultant value is a better estimate of the standard deviation of the population from which the sample is taken. For N greater than 30 the difference is negligible between the two definitions. In cases where the better estimate is needed we need just multiply the above definition by $\sqrt{N/(N-1)}$. So we'll stick with the above for now.

Variance

The variance of a set of data composed of either N numbers $X_1, X_2, \ldots, X_N$ or the frequency sensitive data $X_1, X_2, \ldots, X_K$ with frequencies $f_1, f_2, \ldots, f_K$ is defined as the squares of the standard deviations defined for the respective data sets above.

When it is necessary to distinguish between the standard deviation of a population from that of a sample drawn

from the population, we use the symbol s for the sample standard deviation and σ for the population standard deviation. Thus s^2 refers to the *sample variance* and σ^2 the *population variance*.

Methods for computing the Standard Deviation

The last two equations can be written in the equivalent forms

$$s=\sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N}-\left(\frac{\sum_{i=1}^{N} X_i}{N}\right)^2}=\sqrt{\frac{\sum X^2}{N}-\left(\frac{\sum X}{N}\right)^2}=\sqrt{\overline{X^2}-\overline{X}^2}$$

$$s=\sqrt{\frac{\sum_{i=1}^{K} f_i X_i^2}{N}-\left(\frac{\sum_{i=1}^{K} f_i X_i}{N}\right)^2}=\sqrt{\frac{\sum f X^2}{N}-\left(\frac{\sum f X}{N}\right)^2}=\sqrt{\overline{X^2}-\overline{X}^2}$$

where $\overline{X^2}$ denotes the mean of the squares of the various X, while $\bar{X}^2$ denotes the square of the mean of the various values of X. If $d_i=X_i-A$ are the deviations of X_i from some arbitrary constant A, the results above become, respectively

$$s=\sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}-\left(\frac{\sum_{i=1}^{N} d_i}{N}\right)^2}=\sqrt{\frac{\sum d^2}{N}-\left(\frac{\sum d}{N}\right)^2}=\sqrt{\overline{d^2}-\overline{d}^2}$$

$$s=\sqrt{\frac{\sum_{i=1}^{K} f_i d_i^2}{N}-\left(\frac{\sum_{i=1}^{K} f_i d_i}{N}\right)^2}=\sqrt{\frac{\sum f d^2}{N}-\left(\frac{\sum f d}{N}\right)^2}=\sqrt{\overline{d^2}-\overline{d}^2}$$

When data are grouped into a frequency distribution whose class intervals have equal size c, we have $d_i=c\,u_i$ or $X_i=A+c\,u_i$ and the last equation becomes

$$s = c\sqrt{\frac{\sum_{i=1}^{K} f_i u_i^2}{N} - \left(\frac{\sum_{i=1}^{K} f_i u_i}{N}\right)^2} = c\sqrt{\frac{\sum f u^2}{N} - \left(\frac{\sum f u}{N}\right)^2} = c\sqrt{\overline{u^2} - \overline{u}^2}$$

This last equation provides a very short method for calculating the standard deviation and should always be used for grouped data with equal class intervals. It is exactly analogous to the *coding method* used for calculating the arithmetic mean for grouped data.

Properties of the Standard Deviation

1. The standard deviation can be defined by

$$s = \sqrt{\frac{\sum_{i=1}^{N} (X_i - a)^2}{N}}$$

where a is an average besides the arithmetic mean. The minimum of the above is when $a = \bar{X}$

2. For normal distributions it turns out that
   68.27% of all cases fall between $\bar{X} - s$ and $\bar{X} + s$
   95.45% falls within $\bar{X} - 2s$ and $\bar{X} + 2s$
   99.73% falls between $\bar{X} - 3s$ and $\bar{X} + 3s$

3. If you have two sets containing N_1 and N_2 numbers (or two frequency distributions of total frequencies N_1 and N_2) with variances s_1^2 and s_2^2 respectively and the same mean $\bar{X}$ then the *combined* or *pooled variance* of both sets is a weighed arithmetic mean of the variances:

$$s = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2}$$

Empirical relations between measures of dispersion

For moderately skewed distributions we have the empirical formulae

Mean deviation = 4/5 (standard deviation)

Semi-interquartile Range = 2/3 (standard deviation)

These result from the fact that for normal distributions the mean deviation is 0.7979 times the standard deviation and the semi-interquartile range is 0.6745 times the standard deviation.

Standardised Variable, Standard Scores.

The variable

$$z = \frac{X - \bar{X}}{s}$$

which measures the deviation from the mean in units of standard deviation is called a *standardised variable* and is dimensionless. If deviations are given in terms of standard deviatons then they are in a form which allows inter-comparison of distributions.