

Chapter 4

The Standard Deviation and other Measures of Dispersion

DISPERSION or VARIATION

The degree to which numerical data tend to spread about an average value is called the *variation* or *dispersion* of the data. Various measures of dispersion or variation are available, the most common being the range, mean deviation, semi-interquartile range, 10-90 percentile range, and the standard deviation.

THE RANGE

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

Example: The range of the set 2, 3, 3, 5, 5, 5, 8, 10, 12 is $12 - 2 = 10$. Sometimes the range is given by simply quoting the smallest and largest numbers. In the above example, for instance, the range could be indicated as 2 to 12 or 2-12.

THE MEAN DEVIATION, or AVERAGE DEVIATION, of a set of N numbers X_1, X_2, \dots, X_N is defined by

$$\text{Mean Deviation} = \text{M.D.} = \frac{\sum_{j=1}^N |X_j - \bar{X}|}{N} = \frac{\sum |X - \bar{X}|}{N} = \overline{|X - \bar{X}|} \quad (1)$$

where \bar{X} is the arithmetic mean of the numbers and $|X_j - \bar{X}|$ is the absolute value of the deviation of X_j from \bar{X} . (The *absolute value* of a number is the number without the associated sign and is indicated by two vertical lines placed around the number. Thus $|-4| = 4$, $|+3| = 3$, $|6| = 6$, $|-0.84| = 0.84$.)

Example: Find the mean deviation of the set of numbers 2, 3, 6, 8, 11.

$$\begin{aligned} \text{Arithmetic Mean} &= \bar{X} = \frac{2+3+6+8+11}{5} = 6 \\ \text{Mean Deviation} &= \text{M.D.} = \frac{|2-6| + |3-6| + |6-6| + |8-6| + |11-6|}{5} \\ &= \frac{|-4| + |-3| + |0| + |2| + |5|}{5} = \frac{4+3+0+2+5}{5} = 2.8 \end{aligned}$$

If X_1, X_2, \dots, X_K occur with frequencies f_1, f_2, \dots, f_K respectively, the mean deviation can be written as

$$\text{Mean Deviation} = \text{M.D.} = \frac{\sum_{j=1}^K f_j |X_j - \bar{X}|}{N} = \frac{\sum f |X - \bar{X}|}{N} = \overline{|X - \bar{X}|} \quad (2)$$

where $N = \sum_{j=1}^K f_j = \sum f$. This form is useful for grouped data where the X_j 's represent class marks and the f_j 's are the corresponding class frequencies.

Occasionally the mean deviation is defined in terms of absolute deviations from the median or other average instead of the mean. An interesting property of the sum $\sum_{j=1}^N |X_j - a|$ is that it is a minimum when a is the median, i.e. the mean deviation about the median is a minimum.

Note that it would be more appropriate to use the terminology, *mean absolute deviation* rather than mean deviation.

THE SEMI-INTERQUARTILE RANGE or **QUARTILE DEVIATION** of a set of data is defined by

$$\text{Semi-interquartile Range} = Q = \frac{Q_3 - Q_1}{2} \quad (3)$$

where Q_1 and Q_3 are the first and third quartiles for the data. See Problems 6 and 7. The interquartile range $Q_3 - Q_1$ is sometimes used but the semi-interquartile range is more common as a measure of dispersion.

THE 10-90 PERCENTILE RANGE of a set of data is defined by

$$10-90 \text{ Percentile Range} = P_{90} - P_{10} \quad (4)$$

where P_{10} and P_{90} are the 10th and 90th percentiles for the data (see Prob. 8). The semi-10-90 percentile range, $\frac{1}{2}(P_{90} - P_{10})$, can also be used but is not commonly employed.

THE STANDARD DEVIATION of a set of N numbers X_1, X_2, \dots, X_N is denoted by s and is defined by

$$s = \sqrt{\frac{\sum_{j=1}^N (X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} = \sqrt{\frac{\sum x^2}{N}} = \sqrt{(X - \bar{X})^2} \quad (5)$$

where x represents the deviations of each of the numbers X_j from the mean \bar{X} .

Thus s is the root mean square of the deviations from the mean or, as it is sometimes called, the *root mean square deviation* (see Page 49).

If X_1, X_2, \dots, X_K occur with frequencies f_1, f_2, \dots, f_K respectively, the standard deviation can be written as

$$s = \sqrt{\frac{\sum_{j=1}^K f_j (X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum f (X - \bar{X})^2}{N}} = \sqrt{\frac{\sum f x^2}{N}} = \sqrt{(X - \bar{X})^2} \quad (6)$$

where $N = \sum_{j=1}^K f_j = \sum f$. In this form it is useful for grouped data.

Sometimes the standard deviation for the data of a sample is defined with $(N - 1)$ replacing N in the denominators of the expressions in (5) and (6) because the resulting value represents a better estimate of the standard deviation of a population from which the sample is taken. For large values of N (certainly $N > 30$) there is practically no difference between the two definitions. Also, when the better estimate is needed we can always obtain it by multiplying the standard deviation computed according to the first definition by $\sqrt{N/(N - 1)}$. Hence we shall adhere to the definition given above.

THE VARIANCE

The variance of a set of data is defined as the square of the standard deviation and is thus given by s^2 in (5) and (6).

CHAP. 4] STANDARD DEVIATION and other MEASURES OF DISPERSION

When it is necessary to distinguish the standard deviation of a population from the standard deviation of a sample drawn from this population, we often use the symbol s for the latter and σ for the former. Thus s^2 and σ^2 would represent the *sample variance* and *population variance* respectively.

SHORT METHODS for COMPUTING the STANDARD DEVIATION

The equations (5) and (6) can be written respectively in the equivalent forms

$$s = \sqrt{\frac{\sum_{j=1}^N X_j^2}{N} - \left(\frac{\sum_{j=1}^N X_j}{N}\right)^2} = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2} \quad (7)$$

$$s = \sqrt{\frac{\sum_{j=1}^K f_j X_j^2}{N} - \left(\frac{\sum_{j=1}^K f_j X_j}{N}\right)^2} = \sqrt{\frac{\sum f X^2}{N} - \left(\frac{\sum f X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2} \quad (8)$$

where $\overline{X^2}$ denotes the mean of the squares of the various values of X , while \bar{X}^2 denotes the square of the mean of the various values of X . See Problems 12-14.

If $d_j = X_j - A$ are the deviations of X_j from some arbitrary constant A , the results (7) and (8) become respectively

$$s = \sqrt{\frac{\sum_{j=1}^N d_j^2}{N} - \left(\frac{\sum_{j=1}^N d_j}{N}\right)^2} = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\overline{d^2} - \bar{d}^2} \quad (9)$$

$$s = \sqrt{\frac{\sum_{j=1}^K f_j d_j^2}{N} - \left(\frac{\sum_{j=1}^K f_j d_j}{N}\right)^2} = \sqrt{\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N}\right)^2} = \sqrt{\overline{d^2} - \bar{d}^2} \quad (10)$$

See Problems 15 and 17.

When data are grouped into a frequency distribution whose class intervals have equal size c , we have $d_j = cu_j$ or $X_j = A + cu_j$ and (10) becomes

$$s = c \sqrt{\frac{\sum_{j=1}^K f_j u_j^2}{N} - \left(\frac{\sum_{j=1}^K f_j u_j}{N}\right)^2} = c \sqrt{\frac{\sum f u^2}{N} - \left(\frac{\sum f u}{N}\right)^2} = c \sqrt{\overline{u^2} - \bar{u}^2} \quad (11)$$

This last formula provides a very short method for computing the standard deviation and should always be used for grouped data when class interval sizes are equal. It is called the *coding method* and is exactly analogous to that used in computing the arithmetic mean of grouped data in Chapter 3. See Problems 16-19.

PROPERTIES of the STANDARD DEVIATION

The standard deviation can be defined as $s = \sqrt{\frac{\sum_{j=1}^N (X_j - a)^2}{N}}$

where a is an average besides the arithmetic mean. Of all such standard deviations, the minimum is that for which $a = \bar{X}$, because of Property (b), Chap. 3, Page 46. This property provides an important reason for defining the standard deviation as above. For a proof of this property see Prob. 27.

For normal distributions (see Chapter 7) it turns out that:

- (a) 68.27% of the cases are included between $\bar{X} - s$ and $\bar{X} + s$
(i.e. one standard deviation on either side of the mean)

- (b) 95.45% of the cases are included between $\bar{X} - 2s$ and $\bar{X} + 2s$
 (i.e. two standard deviations on either side of the mean)
- (c) 99.73% of the cases are included between $\bar{X} - 3s$ and $\bar{X} + 3s$
 (i.e. three standard deviations on either side of the mean)

as indicated in Fig. 4-1.

For moderately skewed distributions the above percentages may hold approximately (see Prob. 24).

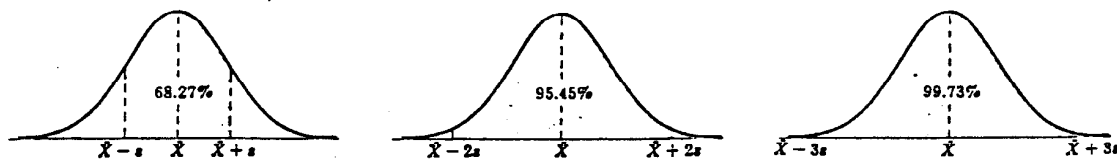


Fig. 4-1

3. Suppose that two sets consisting of N_1 and N_2 numbers (or two frequency distributions with total frequencies N_1 and N_2) have variances given by s_1^2 and s_2^2 respectively and the same mean \bar{X} . Then the *combined* or *pooled variance* of both sets (or both frequency distributions) is given by

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2} \quad (12)$$

Note that this is a weighted arithmetic mean of the variances. This result can be generalized to 3 or more sets.

Small Sampling Theory,

"STUDENT'S" t DISTRIBUTION and THE CHI-SQUARE DISTRIBUTION

SMALL SAMPLES

In previous chapters we often made use of the fact that for samples of size $N > 30$ called *large samples*, the sampling distributions of many statistics were approximately normal, the approximation becoming better with increasing N . For samples of size $N < 30$ called *small samples*, this approximation is not good and becomes worse with decreasing N , so that appropriate modifications must be made.

A study of sampling distributions of statistics for small samples is called *small sampling theory*. However, a more suitable name would be *exact sampling theory*, since the results obtained hold for large as well as small samples. In this chapter we study two important distributions, called "*Student's*" t distribution and the *chi-square distribution*.

"STUDENT'S" t DISTRIBUTION

Let us define the statistic

$$t = \frac{\bar{X} - \mu}{s} \sqrt{N-1} = \frac{\bar{X} - \mu}{\hat{s}/\sqrt{N}} \quad (1)$$

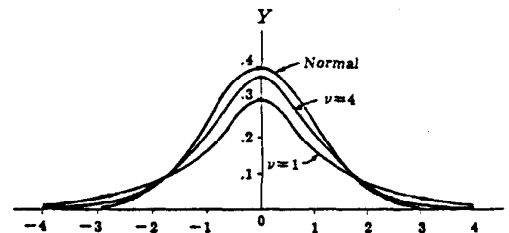
If we consider samples of size N drawn from a normal (or approximately normal) population with mean μ and if for each sample we compute t , using the sample mean \bar{X} and sample standard deviation s or \hat{s} , the sampling distribution for t can be obtained. This distribution (see Fig. 11-1) is given by

$$Y = \frac{Y_0}{\left(1 + \frac{t^2}{N-1}\right)^{N/2}} = \frac{Y_0}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}} \quad (2)$$

where Y_0 is a constant depending on N such that the total area under the curve is one, and where the constant $\nu = (N-1)$ is called the *number of degrees of freedom* (ν is the Greek letter *nu*). For a definition of degrees of freedom, see Page 191.

The distribution (2) is called "*Student's*" t distribution after its discoverer *Gosset*, who published his works under the pseudonym of "*Student*" during the early part of the twentieth century

For large values of ν or N (certainly $N \geq 30$) the curves (2) closely approximate the standardized normal curve $Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$ as indicated in Fig. 11-1.



Student's t distributions for various values of ν

Fig. 11-1

CONFIDENCE INTERVALS

As done with normal distributions in Chapter 9, we can define 95%, 99% or other confidence intervals by using the table of the t distribution in the Appendix, Page 344. In this manner we can estimate within specified limits of confidence the population mean μ .

For example, if $-t_{.975}$ and $t_{.975}$ are the values of t for which 2.5% of the area lies in each "tail" of the t distribution, then a 95% confidence interval for t is

$$-t_{.975} < \frac{\bar{X} - \mu}{s} \sqrt{N-1} < t_{.975} \quad (3)$$

from which we see that μ is estimated to lie in the interval

$$\bar{X} - t_{.975} \frac{s}{\sqrt{N-1}} < \mu < \bar{X} + t_{.975} \frac{s}{\sqrt{N-1}} \quad (4)$$

with 95% confidence (i.e. probability .95). Note that $t_{.975}$ represents the 97.5 percentile value, while $t_{.025} = -t_{.975}$ represents the 2.5 percentile value.

In general, we can represent confidence limits for population means by

$$\bar{X} \pm t_c \frac{s}{\sqrt{N-1}} \quad (5)$$

where the values $\pm t_c$, called *critical values* or *confidence coefficients*, depend on the level of confidence desired and the sample size. They can be read from the table (Appendix III)

TESTS of HYPOTHESES and SIGNIFICANCE

Tests of hypotheses and significance as discussed in Chapter 10 are easily extended to problems involving small samples, the only difference being that the z score or z statistic is replaced by a suitable t score or t statistic.

1. Means.

To test the hypothesis H_0 that a normal population has mean μ , we use the t score or t statistic

$$t = \frac{\bar{X} - \mu}{s} \sqrt{N-1} = \frac{\bar{X} - \mu}{\bar{s}} \sqrt{N} \quad (6)$$

where \bar{X} is the mean of a sample of size N .

2. Differences of Means.

Suppose that two random samples of sizes N_1 and N_2 are drawn from normal populations whose standard deviations are equal ($\sigma_1 = \sigma_2$). Suppose further that

SMALL SAMPLING THEORY

[CHAP. 11

these two samples have means and standard deviations given by \bar{X}_1, \bar{X}_2 and s_1, s_2 respectively. To test the hypothesis H_0 that the samples come from the same population (i.e. $\mu_1 = \mu_2$ as well as $\sigma_1 = \sigma_2$), we use the t score given by

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{1/N_1 + 1/N_2}} \quad \text{where} \quad \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \quad (7)$$

The distribution of t is Student's distribution with $\nu = N_1 + N_2 - 2$ degrees of freedom.

Use of (7) is made plausible on placing $\sigma_1 = \sigma_2 = \sigma$ in the z score of Equation (2), Page 170 and then using as an estimate of σ^2 the weighted mean $\frac{(N_1 - 1)\hat{s}_1^2 + (N_2 - 1)\hat{s}_2^2}{(N_1 - 1) + (N_2 - 1)}$
 $= \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}$ where \hat{s}_1^2 and \hat{s}_2^2 are the unbiased estimates of σ_1^2 and σ_2^2 (see Property 3, Page 72).

See your lab manual (Appendix III) for an example!