Unit 17

Measures of Dispersion and Variability

Contents

Learning Objectives

It is expected that after reading Unit 17 you would be able to

- Obtain a measure of dispersion of data
- Explain the meaning of the term 'range' and work out how to measure the range of one's data
- Discuss the element of variance in one's data and find out the standard variation in it
- Work out the coefficient of variation in the data.

17.1 Introduction

In addition to a measure of central tendency, it is generally desirable to have a measure of dispersion of data. A measure of dispersion (or a measure of variability◎, as it is sometimes called) is an indication of the clustering of measurements around the center of the distribution, or, conversely, it is an indication of how variable the measurements are. Sanders (1955) held that you need to measure dispersion to evaluate the extent to which the average value depicts the data. Another reason for measuring dispersion is to find out the spread in order to improve or control the existing variations.

17.2 The Range

The difference between the highest and the lowest measurements in a group of data is termed range◎. If sample measurements are arranged in an increasing order of magnitude, as if the median were about to be determined, then

Sample range = $X_n - X_1$ 1

Where, X_1 and X_n are the lowest and the highest value of the series respectively.

See Box 17.1 for Example 1.

Box 17.1 Example 1

The number of cattle owned by members of a community is recorded as: 12, 11, 13, 20, 15, 18, 19, 17, 22 and 23. Calculate the range.

X_1= 11; X_n = 23

Sample range = 23 - 11 = 12

The range is a relatively crude measure of dispersion, inasmuch as it does not take into account any measurement except the highest and the lowest. Furthermore since it is unlikely that a sample will contain both the highest and the lowest values in the population, the sample range usually underestimates the population range; therefore, it is a biased and inefficient estimator. Nonetheless, it is useful in some circumstances to present the sample range as an estimate (although a poor one) of the population range. Whenever the range is specified in reporting data, it is usually a good practice to report another measure of dispersion as well.

The Mean Deviation

It is clear that no information is provided by the range about the distribution of the measurements in the middle. Since the mean is so useful a measure of central tendency, one might express dispersion in terms of deviations from the mean.

The sum of all deviations from the mean ((Σ(X_i - M)) will always be zero, therefore such a summation would be useless as a measure of dispersion. On the other hand, the sum of the absolute values of the deviation from the mean expresses dispersion about the mean. Dividing this sum by the total number yields a measure that is known as *mean deviation*, or *mean absolute deviation* of the sample, is obtained.

Sample mean deviation = $(\Sigma| X_i - M |) / n$ 2

Where, M is the mean (sample), X_i are the scores, n is the total number of scores and ?is 'the sum of' and the vertical lines indicate that the values are absolute (irrespective of sign). See Box 17.2 for example 2.

Box 17.2 Example 2

The number of cattle owned by members of a community is recorded as: 12, 11, 13, 20, 15, 18, 19, 17, 22 and 23. Calculate the mean deviation.

ΣX_i = 12 + 11 + 13 + 20 + 15 + 18 + 19 + 17 + 22 + 23 = 170

N = 10

M =ΣX_i / N; M = 170 / 10=17

$(\Sigma|X_i - M|)$ = (12 - 17) + (11 - 17) + (13 - 17) + (20 - 17) + (15 -17) + (18 - 17) + (19 - 17) + (17 - 17) + (22 - 17) + (23 - 17) = 5 + 6 + 4 + 3 + 2 + 1 + 2 + 0 + 5 + 6 = 34

Sample mean deviation = 34 / 10 = 3.4

It is possible that the two samples may have the same range, but not the mean deviation. Mean deviation can also be defined by using the sum of the absolute deviations from the median rather than from the mean.

17.3 The Variance

Another method of eliminating the signs of deviations from the mean is to square the deviations. The sum of the square of deviation from the mean is called the *sum of squares*, abbreviated SS, and is defined as follows:

Sample SS = $\Sigma\ (X_i - M)^{2}$ 3

Where, M is the mean (sample), X_i are the scores, and Σis 'the sum of'.

From the sample SS, population SS can be estimated.

Population SS = $\Sigma\ (X_i - \mu)^{2}$ 4

Where M is the mean (sample), X_i are the scores, and ?is 'the sum of'.

The mean sum of square is called *variance* (or *mean square*, the latter being short for *mean squared deviation*), and for a population is denoted by ó ^{2} ("sigma squared", using the lowercase Greek letter).

Calculating variance from ungrouped data

Population Variance = ó ^{2} = $\Sigma(X_i - \mu)^{2} / N$ 5

The best estimate of the population variance, ó ^{2}, is the sample variance, s^2:

Sample Variance = $s^2 = \Sigma(X_i - M)^{2} / (n - 1)$ 6

Where M is the mean (sample), X_i are the scores, n is the total number of scores (sample) and Σ is 'the sum of'.

The replacement of *μ* by M and N by n in the above equation results in a quantity which is a biased estimate of ó ^{2}. Dividing the sample's sum of squares by n-1 (called the degree of freedom, abbreviated DF) rather than by n, yields an unbiased estimate and the above equation should be used to calculate the sample variance. If all observations are equal, then there is no variability and $s^2 = 0$; and s^2 becomes increasingly large as the amount of variability, or dispersion, increases. Since s^2 is a mean sum of squares, it can never be a negative quantity.

The variance expresses the same type of information as does the mean deviation, but it has certain important properties relative to probability and hypothesis testing that makes it distinctly superior. Thus, the mean deviation is very seldom encountered in social or bio-statistical analysis.

The variance has square units. If measurements are in grams, their variance will be in grams squared, or if the measurements are in cubic centimeters, their variance will be in terms of cubic centimeters squared, even though such squared units have no physical interpretation.

The sample variance◎ can be calculated using the following formula

Sample variance = $s^2 = ((\Sigma\ X_i^{2}) - (\Sigma\ X_i)^{2} / n)) / (n - 1)$ 7

The above formula is often called the machine formula, because of its computational advantages. There are, in fact, two major advantages in calculating SS by Equation 7 rather than by Equation 6. First, here fewer computational steps are involved, a fact that decreases the chance of error. On a good desk calculator, the summed quantities, ΣX_i and ΣX_i^2 can both be obtained with only one pass through the data, whereas Equation 6 requires one pass through the data to calculate M, and at least one more pass to calculate and sum the squares of the deviations, $X_i - M$. Second, there may be a good deal of rounding error in calculating each $X_i - M$, a situation which leads to decreased accuracy in computation, but which is avoided by the use of Equation 7. See Box 17.3 for example 3.

Box 17.3 Example 3

The number of cattle owned by members of a community is recorded as: 12, 11, 13, 20, 15, 18, 19, 17, 22 and 23. Calculate the sample variance.

$\Sigma X_i = 12 + 11 + 13 + 20 + 15 + 18 + 19 + 17 + 22 + 23 = 170$

$n = 10$

$M = \Sigma X_i / n$; $M = 170 / 10 = 17$

X_i	12	11	13	20	15	18	19	17	22	23	$\Sigma X_i = 170$
$X_i - M$	-5	-6	-4	+3	-2	+1	+2	0	+5	+6	
$(X_i - M)^2$	25	36	17	9	4	1	4	0	25	36	$\Sigma(X_i - M)^2 = 156$

Sample variance = $s^2 = \Sigma (X_i - M)^2 / (n - 1) = 156 / 9 = 17.33$

Alternate formula (often called machine formula)

Sample variance = $s^2 = ((\Sigma X_i^2) - (\Sigma X_i)^2 / n)) / (n - 1)$

X_i	12	11	13	20	15	18	19	17	22	23	$\Sigma X_i = 170$
X_i^2	144	121	179	400	225	324	361	289	484	529	$\Sigma X_i^2 = 3046$

Sample variance = $s^2 = (3046 - ((170)^2)/ 10) / 9 = 156 / 9 = 17.33$

Calculating the variance from grouped data

The sample variance in the grouped data can be calculated using the following formula.

$$\text{Sample Variance} = s^2 = \Sigma f_i (X_i - M)^2 / (n - 1) \qquad \text{..........8}$$

Where, M is the mean (sample), f is the frequency of observations with magnitude $X_{i.}$, n is the total number of scores (sample) and $\Sigma$is 'the sum of'.

The manual calculation becomes complex, if the mean value is having several places after decimal. A commonly used method is from assumed mean. The formula is listed below.

$$\text{Sample Variance} = s^2 = \{(\Sigma f_i * d_i^2)/ n - (\Sigma f_i * d_i / n)^2\} * i \qquad \text{.........9}$$

$$d_i = (X_i - A)/ i$$

Where, i is the size of the class interval, f_i is the frequency of observations with magnitude X_i, n is the total number of scores (sample) and Σis 'the sum of'. See Box 17.4 for example 4.

Box 17.4 Example 4

Agricultural land (in acres) owned by various households is grouped under the following seven groups. The frequency of households in each category is listed below. Find the variance in the land owning.

Land Owned (in Acres)	Frequency (F_i)	Mid-Point of the Interval (X_i)	$d_i = (X_i - A) / i$	$f_i * d_i$	d_i^2	$f_i * d_i^2$
20 - 30	18	25	-3	-54	9	172
30 - 40	19	35	-2	-38	4	76
40 - 50	12	45	-1	-12	1	2
50 - 60	19	Assumed Mean (A) = 55	0	0	0	0
60 - 70	17	65	1	17	1	17
70 - 80	15	75	2	30	4	60
80 - 90	10	85	3	30	9	90
	110			-27		417

$\Sigma f_i * d_i^2 = 417$ $\Sigma f_i * d_i = - 27$ $n = 110$

Sample Variance = $\{(\Sigma f_i * d_i^2)/ n - (\Sigma f_i * d_i / n)^2\} * i$

Sample Variance = $\{(417/ 110) - (- 27/ 110)^2\} * 10 = (3.79 - .06) / 10 = 37.3$

The variance in the grouped data can also be calculated using the following equation (often called machine formula).

Sample variance $(s^2) = ((\Sigma f_i * X_i^2) - (\Sigma f_i * X_i)^2 / n)) / (n - 1)$ 10

Where f_i is the frequency of observations with magnitude X_i.

But with a desk calculator it is often faster to use Equation 7 for each individual observation, disregarding the class groupings. See Box 17.5 for example 5.

Box 17.5 Example 5

An investigation in a community on the bride price yielded the following data. Find the variance in bride price.

Bride Price (in Thousand Rs)	Frequency (F_i)	Mid-Point of the Interval (X_i)	$f_i * X_i$	X_i^2	$f_i * X_i^2$
10 - 20	8	15	120	225	1800
20 - 30	9	25	225	625	5625
30 - 40	12	35	420	1225	14700
40 - 50	9	45	405	2025	18225
50 - 60	7	55	385	3025	21175
60 - 70	5	65	325	4225	21125
	50		1880		82650

$\Sigma f_i * X_i^2 = 82650$ $\quad \Sigma f_i * X_i = 1880$ $(\Sigma f_i * X_i)^2 = (1880)^2 = 3534400$
$n = 50$

Sample variance (s^2) = $((\Sigma f_i * X_i^2) - (\Sigma f_i * X_i)^2 / n)) / (n - 1)$

Sample variance (s^2) = (82650 - (3534400 / 50) / 49 = (82650 - 70688) / 49 = 11962 / 49 = 244.12

Reflection and Action 17.1

Following the examples in the text, provide your own examples for calculating variance from ungrouped and grouped data.

17.4 The Standard Deviation

The standard deviation◎ is the positive square root of the variance; therefore, it has the same units as the original measurements. It can be calculated using the following formula.

Standard deviation (s) = v (Sample Variance)

In Example 5, you found the sample variance to be = 244.12, and therefore you can work out the standard deviation to be (s) = v 244.12 = 15.62

Thus various examples given above for the calculation of variance explain the procedure of calculating standard deviation.

17.5 Coefficient of Variation

Ratio scales are useful in social science research when an investigator is interested in the variability of a sample on one characteristic as compared to another.

The coefficient of variation is the percentage ratio of standard deviation to mean and it is calculated using the following formula.

Coefficient of variation = standard deviation *100 / Mean

It is a useful measure of dispersion, when comparison of variability is being made between the variables of unequal magnitude and/ or have different units of measurements, for example, height and weight.

In example 4, you would find that

Mean (M) = AM + $(\Sigma f_i * d_{i)} / n) * i$ = 55 + (-27 / 110) *10 = 55 - 2.45 = 52.55 and

Standard deviation (s) = v (Sample Variance) = v37.3 = 6.107

Coefficient of variation = s *100 / M = 6.107 / 52.55 = 11.62

Reflection and Action 17.2

Work out standard deviation and coefficient of variation of the examples you selected in Reflection and action 17.1.

17.6 Conclusion

After working out in Unit 16 how to measure the central tendency in one's data, in Unit 17 you acquired the skill of measuring dispersion of data, which indicates the clustering of measurements around the center of the distribution, or, you may say that it is an indication of how variable the measurements are.

You may agree with Sanders (1955: 90-91) who said that the range is an easy measure to work out and understand because it requires only one subtraction and it places stress on the extreme values. The mean absolute deviation, on the other hand, places equal weight to the deviation in every observation and it is equally easy to work out and understand. The squaring of deviations in calculating standard deviation emphasies the extreme value. The standard deviation is a more common measure of dispersion. The value of every observation in a series affects the value of this measure. A change in the value of any observation will generate a change in the standard deviation value. Relatively few extreme values can distort its value. The standard deviation is not possible to compute from an open ended distribution. Finally, the co-efficient of deviation is similar to the range as it is based on only two values, which identify the range of the middle fifty percent of the values.It is mostly used in the sets of skewed data and it is possible to compute it in an open-ended distribution.

Further Reading

Sanders, Donald 1955, *Statistics*. McGraw-Hill: New York