1

LOCATION AND DISPERSION

1. Location or central tendency. Fifteen 1968 British 10p coins, all completely new, were weighed separately, in grams. Their weights, in order, were:

12.48	12.63	12.64	12.65	12.66	12.67	12.68	12.71
12.73	12.77	12.78	12.80	12.82	12.90	12.93	

The *arithmetic mean*, or more simply the *mean*, of these fifteen weights is

$$\tfrac{1}{15}(12.48+12.63+12.64+\ldots+12.90+12.93) = \tfrac{1}{15}(190.85)\text{ g}$$
$$= 12.723\text{ g (3 decimals).}$$

Note that means are usually stated to one place (sometimes two places) of decimals more than the data.

The weight 12.723 g is commonly called the *average* of the fifteen weights, but in the study of Statistics it is usually called the mean. The mean of a group of observations is a measure of location of the group. It is a single number which enables us to assess the position in which the group is located with respect to other groups. This single number is also called the *central tendency* of the group.

Another useful measure of location is the *median*. This is the central observation of the group. In the above example it is 12.71 g. It has an equal number of observations above and below it and provides us with an actual specimen for the central tendency of the group. In a group of boys, the boy of median age can be interviewed, the physique of the boy of median weight can be examined, and the script of the examination candidate with the median mark can be subjected to further scrutiny. If a group contains an *even* number of observations the mean of the two central observations is taken as the median. The median is unaffected by abnormal individuals. It is also unaffected if the observations are *transformed* by some process such as *taking logarithms* or *squaring*. It is, however, unsuitable for work demanding mathematical manipulation.

Returning, therefore, to a study of the mean, we may regard the above example as a particular case of the general definition:

The mean of the n observations $x_1, x_2, \ldots, x_n$ *is*

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

$$= \frac{1}{n}\sum_{r=1}^{n} x_r$$

and we may write $$\bar{x} = \frac{1}{n}\Sigma x$$

if no confusion is likely to arise from the omission of the suffix and limits.

In 1968 the 10p cupro-nickel piece replaced the *florin*. Prior to 1947 the florin was minted from a *half silver* alloy. It is interesting, therefore, to compare the fifteen 10p cupro-nickel pieces with fifteen 'silver' florins minted between 1926 and 1946. When these were weighed, in grammes, their weights in order were:

12.05	12.05	12.08	12.15	12.60	12.70	12.71	12.72
12.83	12.86	12.89	12.90	12.91	12.92	12.93	

Thus for these fifteen florins, $\Sigma x = 189.30$ g and the mean weight $\bar{x} = 12.62$ g exactly. This indicates how the weights of the florins are located with respect to the weights of the new 10p pieces. It is important to decide, however, if the difference between the means 12.723 g and 12.62 g is so small as to be *probably negligible* or if it is large enough to support the view that, on the average, new 10p pieces are heavier than used florins. In other words, 'Is the difference between the means *significant*?' This question is introduced in this opening section but not answered. It is one of the many examples of significance which will be discussed later.

2. Change of origin and unit. The calculation of the mean weight of the fifteen florins can be simplified by working with an *arbitrary origin* such as 12.40 g and at the same time taking 0.01 g as *unit*. The fifteen observations then become

−35	−35	−32	−25	20	30	31	32
43	46	49	50	51	52	53	

and their mean is 22. From this the mean weight of the fifteen florins is deduced as

$$12.40 + 0.01 \times 22 = 12.62 \text{ g}.$$

In general, if the n observations $x_1, x_2, \ldots, x_n$ are converted to $X_1, X_2, \ldots, X_n$ by working with A as an arbitrary origin and B as unit then

$$x_1 = A + BX_1, \quad x_2 = A + BX_2, \ldots, x_n = A + BX_n.$$

Hence $$\Sigma x = nA + B\Sigma X \quad \text{and} \quad \bar{x} = A + B\bar{X}.$$

3. Variance. Standard deviation. If $\bar{x}$ is the mean of the n observations $x_1, x_2, \ldots, x_n$ then the n values

$$(x_1-\bar{x}), \quad (x_2-\bar{x}), \ldots, (x_n-\bar{x})$$

are called the *deviations from the mean*, and

$$(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2$$

is the *sum of the squares* of these deviations.

The *variance* of the n observations $x_1, x_2, \ldots, x_n$ is defined as

$$\frac{1}{n}\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2\} = \frac{1}{n}\Sigma(x-\bar{x})^2.$$

It may be described as the *mean-square deviation from the mean*. The *standard deviation* of the n observations $x_1, x_2, \ldots, x_n$ is

$$\mathrm{S} = \sqrt{\left\{\frac{\Sigma(x-\bar{x})^2}{n}\right\}}.$$

It is the positive square root of the variance and may be described as the *root-mean-square deviation from the mean*. In the examples under discussion the standard deviation will be denoted by the large capital English letter S. In certain circumstances, to be discussed later, it will be denoted by the Greek letter σ or by the small English letter s.

The variance of the weights of the fifteen florins is, therefore,

$$\begin{aligned}\mathrm{S}^2 &= \tfrac{1}{15}\{(12.05-12.62)^2+(12.05-12.62)^2+\ldots+(12.93-12.62)^2\}\ \mathrm{g}^2\\ &= 0.11403\ \mathrm{g}^2.\end{aligned}$$

Note that:

(i) the units of this variance are g^2,

(ii) the above method of calculation is awkward,

(iii) if, as in the case of the 10p pieces the mean is not exact, there is an error in each of the fifteen deviations from the mean.

The difficulties (ii) and (iii) above can be avoided by realizing that the sum of the squares of deviations from the mean

$$\begin{aligned}\Sigma(x-\bar{x})^2 &= \Sigma(x^2-2x\bar{x}+\bar{x}^2)\\ &= \Sigma x^2-2\bar{x}\Sigma x+n\bar{x}^2\\ &= \Sigma x^2\times 2\bar{x}(n\bar{x})+n\bar{x}^2\\ &= \Sigma x^2-n\bar{x}^2.\end{aligned}$$

Hence, the variance may be calculated by

$$\mathrm{S}^2 = \frac{1}{n}\Sigma x^2-\bar{x}^2,$$

or

$$\mathrm{S}^2 = \frac{1}{n}\Sigma x^2-\frac{1}{n^2}(\Sigma x)^2.$$

When using a desk calculating machine the second of the above formulae is best because it avoids squaring the rounding-off error in $\bar{x}$.

Thus, the variance of the weights of the fifteen florins is

$$\begin{aligned} S^2 &= \tfrac{1}{15}(12.05^2+12.05^2+\ldots+12.93^2)-\tfrac{1}{225}(12.05+12.05+\ldots+12.93)^2 \\ &= 159.37843-159.2644 \\ &= 0.11403\ \text{g}^2 \end{aligned}$$

and the standard deviation S = 0.338 g (3 decimals).

Note that on an electronic computer, where the number of significant figures (or decimal places) is fixed, it is best to compute

$$S^2 = \frac{1}{n}\Sigma(x-\bar{x})^2$$

using as many decimal places as possible for $\bar{x}$.

The alternative formulae may give S^2 negative if the variance is small.

4. Change of origin and unit. The calculation of the variance and standard deviation can be considerably simplified by converting $x_1, x_2, \ldots, x_n$ to $X_1, X_2, \ldots, X_n$ as in §2.

Thus, the sum of the squares of deviations from the mean

$$\begin{aligned} nS^2 &= \Sigma(x-\bar{x})^2 \\ &= \Sigma\{(A+BX)-(A+B\bar{X})\}^2 \\ &= \Sigma\{BX-B\bar{X}\}^2 \\ &= B^2\Sigma(X-\bar{X})^2. \end{aligned}$$

Hence the variance $$S^2 = B^2\left\{\frac{\Sigma(X-\bar{X})^2}{n}\right\}$$

and the standard deviation

$$\begin{aligned} S &= B\sqrt{\left\{\frac{\Sigma(X-\bar{X}^2)}{n}\right\}} \\ &= B\sqrt{\left\{\frac{\Sigma X^2}{n}-\bar{X}^2\right\}}. \end{aligned}$$

Suppose, in the case of the fifteen 10p pieces the weights $X_1, X_2, \ldots, X_n$ are taken as

$$-12 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 11 \quad 13 \quad 17 \quad 18 \quad 20 \quad 22 \quad 30 \quad 33,$$

then $A = 12.60$, $B = 0.01$, $\Sigma X = 185$, $\Sigma X^2 = 4119$,

$\bar{X} = 12.33$ (2 decimals), $\Sigma X^2/15 = 274.6$

and $$S = 0.01\sqrt{122.5}$$

$$= 0.111\ \text{g} \quad (3\ \text{decimals}).$$

Note that the value of A is not used in the calculation of S.

5. Dispersion or variability. Inspection of the separate weights of the fifteen 10p pieces shows that the *least* and *greatest* weights recorded are 12.48 g and 12.93 g respectively. That is to say, the *range* of the weights is 12.48 g to 12.93 g, a difference of 0.45 g. On the other hand the range of the weights of the florins is 12.05 g to 12.93 g, a difference of 0.88 g. This indicates that the *dispersion* or *variability* in the weights of the group of florins is approximately twice that of the group of 10p pieces. Alternatively, the weights of the florins may be said to be more widely *dispersed* or *spread* than the weights of the 10p pieces.

Some readers, no doubt, will be surprised that the weights of new 10p pieces vary so much. The weight of a new coin, like the diameter, is so often accepted as a *standard*. The mind often tends to accept the central tendency of a group as a rigid and unvarying standard. In Statistics we are particularly interested in the degree of departure from the central tendency. That is to say, we study and measure the dispersion or variability within the group. The universally accepted measures of dispersion are the variance and standard deviation defined in §3.

The range is unsatisfactory because it is based entirely on the two extreme members which may be abnormal. Moreover, the range generally depends on the size of the sample. If, in assessing the dispersion in a group, use is made of the range, it is generally done by first converting the range to the standard deviation by a table such as that given on page 112.

Measures of dispersion other than the range, variance and standard deviation were described fully in *A First Course in Statistics*. They will not be discussed further in this volume.

Let us note then that the standard deviation of the weights of the fifteen florins is 0.338 g, while the standard deviation of the weights of the fifteen 10p pieces is 0.111 g. This implies that the dispersion of the weights of the florins is approximately three times that of the 10p pieces. The full meaning of the standard deviation is not easily understood by the beginner. Only after studying a large number of examples such as those given in the first two chapters of this book will the student feel that some understanding of this measure of dispersion is being achieved.

The first four florins in the group of fifteen are obviously very worn. If these four observations are discarded, the mean and standard deviation of the other eleven are 12.815 g and 0.108 g. Thus the weights of the eleven florins in good condition have approximately the same location and dispersion as the weights of the fifteen new 10p pieces.

6. Exercises.

1. Working with 980 as origin and 0.1 as unit calculate the mean and the standard deviation of

980.8 981.1 980.7 980.3 981.8 982.5.

2. A sugar refiner uses machines which pack automatically 1 kilo cartons of sugar. To check that the machines are giving correct weight, cartons are selected at random and weighed accurately. The results of checking two machines are:

Accurate weights in kilos of eleven cartons

Machine *A*	1.017	1.051	1.078	0.996	1.033	1.059
Machine *B*	0.995	1.009	1.028	1.036	1.000	1.017
Machine *A*	1.082	1.014	1.040	1.072	0.998	
Machine *B*	1.027	1.045	1.006	1.018	1.039	

Calculate the means and the standard deviations and explain briefly what inference can be drawn from them.

3. A firm which manufactures lead-covered submarine cables checked the thickness of the cover on two of its cables by taking measurements in ten places:

Ten measurements, in cm of the thickness

Cable *A*	0.74	0.76	0.78	0.70	0.72
Cable *B*	0.70	0.72	0.73	0.74	0.72
Cable *A*	0.73	0.75	0.77	0.79	0.71
Cable *B*	0.72	0.74	0.71	0.72	0.73

Calculate for each cable the mean thickness of the cover and the standard deviation of the thickness. Explain briefly what deductions can be made about the covers of the two cables.

4. The times, in minutes, of a car journey made between 5.30 p.m. and 6.30 p.m. along the same route on five consecutive Mondays were as follows:

33 28 26 35 38.

Calculate the mean and the standard deviation of the times.

5. Find the mean and standard deviation of the set of numbers 8, 9, 10, 11, 12. From this set, ten samples each containing two numbers can be selected. Find the mean of each of these samples and calculate the standard deviation of these means. [London]

6. Two forms, one of 20 boys and the other of 30 boys, are given an examination. In the smaller form the average mark was 60 and the standard deviation was 7.0. In the other form the average mark was 50 and the standard deviation was 10.0. Find the standard deviation for the marks of the 50 boys taken as a single group. [London]

7. The numbers of members, means and standard deviations of three distributions are:

No. of members	280	350	630
Means	45	54	49
Standard deviations	6	4	8

Find the mean and standard deviation of the distribution formed by the three distributions taken together. [London]

8. The mean ages of n_1 boys is M_1, and the standard deviation from the mean of their age distribution is σ_1. The mean of the ages of n_2 girls is M_2, and the corresponding standard deviation from the mean is σ_2. Find the mean of the ages of the boys and girls combined.

If $M_1 = M_2$, obtain an expression for the standard deviation from the mean of the combined age distribution. [London]

9. The marks obtained by ten boys in an examination were 12, 17, 20, 23, 26, 29, 29, 35, 38, 41.

Find the standard deviation. The marks are now to be adjusted so that the mean is 60 and the standard deviation is 15. Calculate the highest and the lowest marks obtained on the new scale. What is the purpose of this adjustment? [London]

10. The table gives the number of minutes late or early for the arrival of a train on a number of runs:

Late	2	4	1	6	9	2	1	0
Early	3	1	.	.	.	.	.	.

Calculate the mean of these and the standard deviation.

After two more runs neither the mean nor the standard deviation is altered. Calculate, to the nearest half, the number of minutes late or early for each of these runs. [London]

11. Four boys sit for an examination. The average of their marks is M and the standard deviation is σ. The marks are converted to a new scale by the formula

$$y = 50 - 20(M - x)/\sigma,$$

where y is the new mark and x is the original mark. Find the mean and the standard deviation of the new marks.

If the original marks were 47, 57, 65, 71, find the new marks each to the nearest integer. [London]

12. The marks obtained by the n candidates who passed an examination but did not reach the credit standard ranged from 68 to 76. They were converted to a range of 50 to 60 by reading off the new mark, y, corresponding to an old mark, x, from the straight line graph joining the point (68, 50) to the point (76, 60). Find a formula for y in terms of x and deduce relationships between (i) the new mean, $\bar{y}$, and the old mean, $\bar{x}$; (ii) the new standard deviation, s', and the old standard deviation, s. [Northern]

13. Two dice, each of which has its faces numbered from 1 to 6, are thrown together, and the score found by squaring the difference between the numbers on the faces resting uppermost. Draw up a table showing the number of ways in which each possible score can be obtained. Determine the mean score and calculate the standard deviation from this mean. [London]

14. If the mean of two numbers x and y is 4, show that the mean of the four numbers x, $2x$, y and $2y$ is 6.

If the variance of the numbers x and y is 2, find the variance of the numbers x, $2x$, y and $2y$. [London]

15. Two numbers x and y have a mean 3 and variance 4. Three different numbers a, b and c have a mean 8 and a variance 36. Calculate the mean and variance of the five numbers x, y, a, b and c taken as one group. [London]

16. Given, as in §3, that $\bar{x}$ and S^2 are the mean and variance of the n observations

$$x_1, x_2, \ldots, x_n$$

and given also that the *mean-square deviation from an arbitrary value* a is

$$S_1^2 = \frac{1}{n}\Sigma(x-a)^2$$

show that $$S^2 = S_1^2 - (\bar{x}-a)^2.$$

7. From random sample to parent population. The fifteen 10p pieces and the fifteen florins are only of interest in that they enable us to make statements about 10p pieces and florins *in general*. They are *random samples* from which we are trying to estimate the properties of the *parent populations* from which they are drawn. To distinguish clearly between the mean and standard deviation of a random sample and the corresponding mean and standard deviation of its parent population English small italic letters are used for sample values and the corresponding Greek small letters for the population values. Thus m (rather than $\bar{x}$) and s are used to denote the mean and standard deviation of the sample while μ and σ are used to denote the mean and standard deviation of the parent population. The values μ and σ are examples of *parameters*. A parameter is a constant which takes different values for different populations.

Now m gives a good estimate of μ and we write

$$m = \frac{1}{n}\Sigma x = \text{Est}\,(\mu).$$

Thus 12.62 g may be accepted as an estimate of the mean weight of florins in general and 12.72 g as the mean weight of 10p pieces in general. At this stage the 12.723 g has been reduced to two decimal places because it seems reasonable to state the estimate with the same accuracy as the data.

For σ, however, a better estimate is obtained by using a divisor $(n-1)$ instead of n and denoting the value of the standard deviation obtained in this way by a small s. Thus

$$s = \sqrt{\left\{\frac{\Sigma(x-\bar{x})^2}{(n-1)}\right\}} = \text{Est}\,(\sigma),$$

or

$$s = \sqrt{\frac{n}{(n-1)}}\,\text{S} = \text{Est}\,(\sigma).$$

The reason for this is that σ should really be calculated by taking deviations from μ. However, we use m instead of μ and when the n deviations

$$(x_1-m),\quad (x_2-m),\ \ldots,\ (x_n-m)$$

are used their sum is zero. This implies that when $(n-1)$ of the deviations have been written down the nth deviation is predetermined and we say that, for a random sample of n observations, only $(n-1)$ *degrees of freedom* are available for the calculation of s. For example, in the case of the weights of the fifteen florins the deviations, in g, from the mean are: -0.57, -0.57, -0.54, -0.47, -0.02, $+0.08$, $+0.09$, $+0.10$, $+0.21$, $+0.24$, $+0.27$, $+0.28$, $+0.29$, $+0.30$, $+0.31$.

The sum of the first fourteen deviations is -0.31 and, therefore, the fifteenth deviation is predetermined as $+0.31$. The above explanation is not a proof but it is hoped that it will be sufficient to convince students at this stage of their course.

When calculating s, the formula

$$s = \sqrt{\left\{\frac{\Sigma(x-\bar{x})^2}{(n-1)}\right\}} = \text{Est}\,(\sigma)$$

is generally used in its equivalent form

$$s = \sqrt{\left\{\frac{\Sigma x^2}{(n-1)} - \frac{n}{(n-1)}\bar{x}^2\right\}} = \text{Est}\,(\sigma),$$

or

$$s = \sqrt{\left\{\frac{\Sigma x^2}{(n-1)} - \frac{(\Sigma x)^2}{n(n-1)}\right\}} = \text{Est}\,(\sigma).$$

The last form is convenient when using a desk calculating machine. To estimate the standard deviation of the weights of 10p pieces, therefore, the divisor 14 is used instead of 15 giving

$$\sigma = 0.115\ \text{g}\quad (3\ \text{decimals}).$$

Similarly, the estimate of the standard deviation of the weights of florins

$$\sigma = 0.350\ \text{g}\quad (3\ \text{decimals}).$$

8. Large samples. When n is large the estimate of σ obtained by using the divisor $(n-1)$ will differ very little from the estimate obtained by the divisor n.

Thus, when n is large,

$$\sqrt{\frac{n}{(n-1)}} = \left(1-\frac{1}{n}\right)^{-\frac{1}{2}}$$

$$= 1+\frac{1}{2n} \quad \text{approximately}$$

and the relation $\quad s = \sqrt{\frac{n}{(n-1)}}\,\mathrm{S} = \mathrm{Est}\,(\sigma) \quad$ of §7

becomes $\quad s = \left(1+\frac{1}{2n}\right)\mathrm{S} = \mathrm{Est}\,(\sigma).$

Hence, when $n = 50$ the difference between s and S is 1 % and when $n = 100$ the difference is 0.5 %. Moreover, s is merely an estimate and so for *large samples* (say $n = 50$ or more) it is customary to use the divisor n.

9. Exercises.

1. The values given in §6, Ex. 1, are the results of six determinations of g by a particular piece of apparatus. Estimate from them the standard deviation, σ, to be expected if a large number of determinations of g were made using the same apparatus.

2. Use the times given in §6, Ex. 4, to estimate the standard deviation of the parent population. Explain briefly the meaning, in this example, of the phrase 'parent population'.

3. Estimate the mean and standard deviation of the whole output for each of the machines A and B of §6, Ex. 2.

4. In §6, Ex. 3, if it were possible to measure the thickness of the cover in a very large number of places throughout the whole length of each cable, what would you expect the mean and standard deviation, in each case, to be.

10. Frequency distribution with unequal group intervals. The results of a count of craters on the surface of the Moon are shown in table 1 A. In it the total number of 1596 craters have been divided into *groups* whose sizes are given in the left-hand column. Note that in the first of these groups, since the values are given to one decimal place, the range 5.0–9.9 implies that the diameters are greater than 4.95 and less than 9.95. Thus the range of the first group is 5. Note also that the *group intervals* in the left-hand column are unequal, the first being 5, the next seven being 10, the ninth being 20 and the tenth being 50. The numbers in the right-hand column indicate the *frequency* with which each particular size occurs. In table 1 A then, the 1596 separate observations have been grouped into a *frequency distribution.*