Effects of Changing the Center and Scale

Example 1 Below are the 3 p.m. temperatures, in degrees Celsius, for 10 towns and cities in the North Island of New Zealand one day in February:

22 , 22 , 24 , 23 , 25 , 22 , 20 , 26 , 21 , 23 .

Summary statistics for this batch of numbers are:

$$\overline{x} = 22.8, \quad s_X = 1.81, \quad \text{Med} = 22.5, \quad \text{IQR}_X = 2.$$

If we now changed all the temperatures from degrees Celsius to degrees Fahrenheit, what would happen to these summary statistics? Let X be a temperature in $^\circ C$ and let Y be the same temperature in $^\circ F$. The scales are related in a linear way with $0^\circ C = 32^\circ F$ and $100^\circ C = 212^\circ F$. A formula for relating Y to X is then given by

$$Y = \left(\frac{212 - 32}{100 - 0}\right) X + 32, \quad \text{or} \quad Y = 1.8X + 32.$$

The above temperatures in $^\circ F$ are:

71.6, 71.6, 75.2, 73.4, 77.0, 71.6, 68.0, 78.8, 69.8, 73.4.

The summary statistics for these numbers are:

$$\overline{y} = 73.04, \quad s_Y = 3.26, \quad \text{Med}_Y = 72.5, \quad \text{IQR}_Y = 3.6.$$

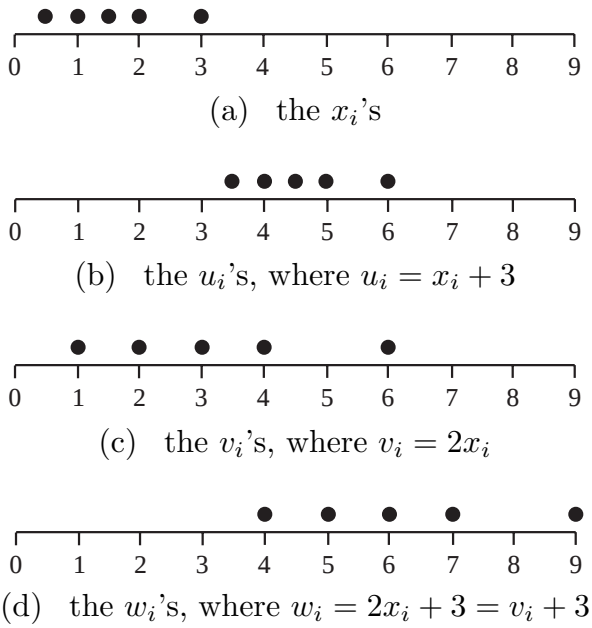

Figure 1 : Dot plot for values of $x_1 = 0.5$, $x_2 = 1$, $x_3 = 1.5$, $x_4 = 2$, $x_5 = 3$ and for three transformations.

We note that the relationship between X and Y in Example 1 is a special case of the ***linear transformation***: $Y = aX + b$.

To find out what happens to summary statistics and plots of the data under such transformations we look at some simple examples. Consider the dot plots in Fig. 1. Suppose our batch of numbers measured on the X-scale are $0.5, 1, 1.5, 2$ and 3 as shown in Fig. 1(a). In Fig. 1(b) we transform each X-value to $U = X + 3$, and the dot plot is now given for the corresponding values of U. Clearly the whole batch is simply displaced by 3 units to the right. Any sensible measure of the centre (or location) of the points should also be displaced 3 units to the right. We are therefore not surprised to find that $\overline{u} = \overline{x} + 3$ and $\text{Med}_U = \text{Med}_X + 3$. However, the distances between the points haven't changed so that sensible measures of spread based on these distances, such as standard deviation and interquartile range, will be left unchanged, i.e. $s_U = s_X$ and $\text{IQR}_U = \text{IQR}_X$.

If we now consider Fig. 1(c) we obtain the V-sample by doubling each element in the X-sample. It seems natural to expect that measures of location will correspondingly be doubled. For example, the mean will be doubled as it is the average of numbers that are doubled. Also, the median is either equal to one of the numbers or midway between two of the numbers and is therefore doubled along with the numbers. In addition, the distances between the points have doubled so that measures of spread like the standard deviation and the interquartile range should also be doubled. Hence $\overline{v} = 2\overline{x}$ and $s_V = 2s_X$, etc.

Finally, consider Fig. 1(d). We get from the X-sample to the W-sample by first doubling (which doubles measures of location and spread), and then shifting everything to the right by 3 units. This latter move only affects the location, which increases by 3 units. Thus $W = 2X + 3$ implies that $\overline{w} = 2\overline{x} + 3, \text{Med}_W = 2\text{Med}_X + 3, s_X = 2s_X$ and $\text{IQR}_W = 2\ \text{IQR}_X$.

If, instead of shifting our data 3 units to the right, we shifted it 3 units to the left so that all the X-values were reduced by 3, the only change in the above formulae is to replace 3 by -3. If we use a scale change of -2 so that $V = -2X$ we are doubling each number and changing its sign. The same will happen to a measure of location so that, for example, $\overline{v} = -2\overline{x}$. Although the values of V are now all negative, the distances between these values are still just double what they were before so that $s_V = 2s_X = |-2|s_X$, where $|-2|$ is the absolute value of -2, namely $+2$.

From the above arguments it is clear that, for the general case[1] :

[1]Changing from x_i to y_i is known as transforming the data (x_i). This particular transformation is called a *linear* transformation. Other transformations like $y_i = \log x_i$ or $y_i = \sqrt{x_i}$ are often useful in data analysis (see Chapter 10).

If each $y_i = ax_i + b$, where a and b can be positive or negative, we have

"**Centre**" : $\overline{y} = a\overline{x} + b$, $\quad \text{Med}_Y = a\text{Med}_X + b$.

"**Spread**" : $s_Y = |a|s_X$, $\quad \text{IQR}_Y = |a|\ \text{IQR}_X$.

Returning to our example of changing temperatures from $^\circ C$ to $^\circ F$, we have $a = 1.8$ and $b = 32$. Thus in $^\circ F$,

Sample mean :	$\overline{y} = 1.8\overline{x} + 32 = 1.8 \times 22.8 + 32 = 73.04.$
Sample median :	$\text{Med}_Y = 1.8\text{Med}_X + 32 = 1.8 \times 22.5 + 32 = 72.5.$
Sample standard dev. :	$s_Y = 1.8s_X = 1.8 \times 1.81 = 3.26.$
IQR :	$\text{IQR}_Y = 1.8\ \text{IQR}_X = 1.8 \times 2 = 3.6.$

Example 2 An electrician bases his charges for house calls on the time taken. The times (in hours) for eight jobs are

$$1.5, \quad 1.0, \quad 2.0, \quad 1.5, \quad 1.2, \quad 1.0, \quad 2.0, \quad 2.5\ .$$

The sample mean and standard deviation of the times taken are, respectively, $\overline{x} = 1.5875$ hours and $s_X = 0.5383507$ hours, where these numbers are given to calculator accuracy as we will be using them for further calculations.[2]

If the electrician charges \$40 per hour together with a fixed call-out charge of \$25 (travelling time) we can find the sample mean and standard deviation of the charges as follows. Let $Y =$ Charge and $X =$Time taken. Then $Y = 40X + 25$ so that the average charge is $\overline{y} = 40\overline{x} + 25 = 40 \times 1.5875 + 25 = \$88.50$ and the standard deviation of charges is $s_Y = 40\, s_X = 40 \times 0.5383507 = \$21.53$.

Exercises

1. If each member of a set of measurements

- **(a)** is multiplied by 5, what will happen to a measure of location? to a measure of spread?
- **(b)** has 10 added to it, what will happen to a measure of location? to a measure of spread?
- **(c)** is subject to $y = 7x + 9$, what will happen to a measure of location? to a measure of spread?

[2] In fact, we do not write them down and then use them for the calculations. They are already resident in the memories of the calculator, and that is where we get them to do the calculations which follow.

(d) is subject to $y = -7x$, what will happen to a measure of location? to a measure of spread?

(e) is subject to $y = -7x+9$, what will happen to a measure of location? to a measure of spread?

(f) is subject to $y = 12 - 2.3x$, what will happen to a measure of location? to a measure of spread?

2. Let $$y_i = \frac{x_i - \overline{x}}{s_X} = \frac{1}{s_X}x_i - \frac{1}{s_X}\overline{x}.$$

By identifying a and b in the previous discussion, show that $\overline{y} = 0$ and $s_Y = 1$. This transformation is sometimes known as *standardizing* the data. For the lengths of the 40 female coyotes (Table 2.3.2 in the text), $\overline{x} = 89.24$ and $s_X = 6.548196$. The first 3 observations are: 93.0, 97.0, 92.0, ... Standardize each of these 3 observations.