2 Summarizing Data

A common way to summarize a data set is with a single value considered to be the centre of the data, along with a measure of how spread out the data is from that centre. Two ways to define the centre of a data set are *mean* and *median*, and the most widely used measure of spread is the *standard deviation*.

Decimal place convention: mean, median and standard deviation are usually rounded off to one more decimal place than was present in the original data.

Units: mean, median and standard deviation all have the same unit as the original data.

2.1 Mean

The mean is the number most people simply call the "average". It is sometimes referred to as the *arithmetic mean*.

Definition: The mean of a data set $x_1, x_2, \cdots, x_n$ is the average value

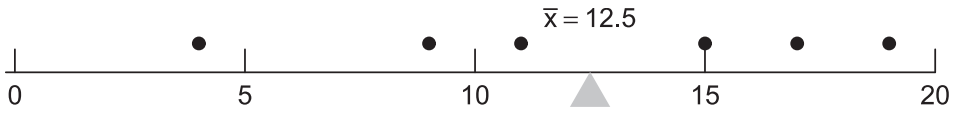
$$\text{mean} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}.$$

Notation: μ is used to denote a population mean and $\bar{x}$ is used for a sample mean. If a data set is not labelled as a sample, we will assume that it is a population.

The SHARP EL-531X has a function to calculate the mean of a data set. Instructions are in the Appendix.

Example 2.1. Find the mean of the sample $11, 9, 17, 19, 4, 15$.

Physical Interpretation of the Mean: If each value in a data set is represented along a weightless horizontal axis by a ball of equal weight, then the mean corresponds to the centre of inertia or balance point of the data. For the previous example, this looks like [2, p. 30]:



Example 2.2. A student has test marks 58, 63, 71. What mark on his 4th test gives him an average of 70?

Example 2.3. If the following two samples are combined into one sample, find the mean.

	sample size	$\bar{x}$
Sample 1	43	71
Sample 2	26	68

If a data set is given using a frequency distribution table instead of a list of numbers,

value	frequency
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_n	f_n

then

$$\boxed{\text{mean} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum xf}{f}}.$$

Example 2.4. Find the mean for the following sample:

Temperature (°C)	Frequency
22	11
23	6
25	3

If a data set is given using relative frequencies,

value	relative frequency
x_1	r_1
x_2	r_2
$\vdots$	$\vdots$
x_n	r_n

then

$$\boxed{\text{mean} = x_1r_1 + x_2r_2 + \cdots + x_nr_n = \sum xr}.$$

Example 2.5. Find the mean for the following sample:

mass (g)	relative frequency
82	0.1
85	0.35
86	0.5
88	0.05

2.2 Median

The second way to define the centre of a data set is to order the values from smallest to largest and then simply select the middle value. This way the data set contains the same number of values that are larger than the median and that are smaller than the median. One convenient property of the median is that it eliminates the effect of outliers (values that are much larger or smaller than the rest of the data).

Definition: If the data set $x_1, x_2, \cdots, x_n$ is listed in order from smallest to largest, then the median is defined as

$$\text{median} = \begin{cases} \text{middle value} & \text{if } n \text{ is odd} \\ \text{average of the 2 middle values} & \text{if } n \text{ is even} \end{cases}.$$

Specifically, if n is odd then the median is the value in position $\dfrac{n+1}{2}$ in the list. And if n is even, then the median is the average of the values in position $\dfrac{n}{2}$ and $\dfrac{n}{2}+1$.
There is no standard notation for the median.

Example 2.6. Find the median of the following data sets:

(a) 2, 9, 11, 5, 6.

(b) 2, 9, 11, 5, 6, 10.

Example 2.7. Find the median for the sample in Example 2.4:

Temperature (°C)	Frequency
22	11
23	6
25	3

If a data set is given using relative frequencies,

value	relative frequency
x_1	r_1
x_2	r_2
$\vdots$	$\vdots$
x_n	r_n

where the values $x_1, x_2, \cdots, x_n$ are listed in order from smallest to largest, then the median is x_i where i is the smallest index for which $r_1 + r_2 + \cdots + r_i \geq 0.5$. That is, we add the relative frequencies in order until 0.5 is first reached or exceeded, and the median is the corresponding value.

Example 2.8. Find the median for the sample in Example 2.5:

mass (g)	relative frequency
82	0.1
85	0.35
86	0.5
88	0.05

2.3 Standard Deviation

Now we turn our attention to measure how spread out a data set is. The spread is small if the values are all bunched close to the mean, and it is large if the values are scattered widely. An intuitive way to do this would be to take the average of the difference of each value from the mean, but this average is always 0.

Definition: The population $x_1, x_2, \cdots, x_n$ has a population variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

and a standard deviation (SD)

$$\sigma = \sqrt{\sigma^2}.$$

The SHARP EL-531X has a function to calculate population SD. Instructions are in the Appendix.

Example 2.9. Find the population SD of 2, 5, 8, 9.

Definition: The sample $x_1, x_2, \cdots, x_n$ has a sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

and a standard deviation (SD)

$$s = \sqrt{s^2}.$$

For the purpose of calculating SD, if a given data set is not labelled as a sample, we assume that it is a population.

Example 2.10. Find the sample SD of 12, 13, 17.

Example 2.11. Which sample is more spread out?

(a) 1, 4, 10

(b) 31, 36, 38

A data set is considered to be accurate if the mean is close to the target value, and it is precise if the variance or SD is small.

Example 2.12. Two machines are filling 355 mL cans of pop. A sample of volumes has the following means and variances (in mL).

	Machine 1	Machine 2
$\overline{x}$	355.8	355.2
s^2	0.3	1.4

(a) Which machine is more accurate?

(b) Which machine is more precise?

Example 2.13. Let a population consist of the salaries at a small engineering firm, where the second highest salary is \$50,000 less than the highest salary. What happens to the mean, median and SD in each situation:

(a) Each employee get a \$2,000 raise.

(b) Each employee's salary is doubled.

(c) The highest salary is decreased by \$10,000.

Example 2.14. Compare the mean, median and SD for the results from the demonstration *An experiment that looks like a survey* in Section 1.

Additional Notes

Additional Notes