1. A biased die produces numbers with the following probability distribution:

1	2	3	4	5	6
0.15	0.1	0.2	0.25	0.15	0.15

Calculate

(a) The expected value of a single roll of the die.

(b) The expected value if the die is rolled twice and added together.

(c) The variance of a single roll of the die.

(a) The expected value for a single roll is:

$$0.15 \times 1 + 0.1 \times 2 + 0.2 \times 3 + 0.25 \times 4 + 0.15 \times 5 + 0.15 \times 6 = 3.6$$

(b) To get the expected value for two rolls, you need to calculate the probability of getting each score, remembering that there are several different ways of obtaining each score: for example 6 can be 1 + 5, 2 + 4, 3 + 3, 4 + 2 or 5 + 1 so the probability of it is

$$0.15 \times 0.15 + 0.1 \times 0.25 + 0.2 \times 0.2 + 0.25 \times 0.1 + 0.15 \times 0.15 = 0.135$$

The answer is 7.2.

(c) To calculate the variance we also need the expectation of the square:

$$0.15 \times 1 + 0.1 \times 4 + 0.2 \times 9 + 0.25 \times 16 + 0.15 \times 25 + 0.15 \times 36 = 15.5$$

Then the variance is $15.5 - (3.6)^2 = 2.54$.

2. The mean of a set of n data points is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(a) Calculate an expression for the change in the mean if a new data point x_{n+1} is measured.

(b) Calculate (i) the mean and (ii) the median of the data set { 1.9, 2.3, 3.1, 3.6, 3.7, 4.0, 4.1, 4.5 }.

(c) Calculate the change in (i) the mean and (ii) the median if the data points { 198.2, 206.6 } are added to the data set.

(d) Comment on these changes. Do you think either average is still a useful measure of the location? How would you use summary statistics to understand the changes in the data set?

(a) If the previous mean is $\bar{x}_0$ and the new mean is $\bar{x}_1$, the expression is

$$\bar{x}_1 = \frac{n}{n+1}\bar{x}_0 + \frac{x_{n+1}}{n+1}$$

(b) (i) 3.4 (ii) 3.65 (remember the median is the mean of the two central points if there is an even number of data points).

(c) The new mean is 54 and the new median is 3.85, so the changes are (i) 50.6 and (ii) 0.2.

(d) The mean has changed very substantially because of the effect of the outliers, while the median has not changed very much. The median is therefore a more useful measure of the location of most of the dataset while the mean is no longer very useful. It would be sensible to add a measure of spread such as the variance to show that the dataset is much more spread out with the new data points (for example the variance goes from 0.7175 to 6340.262 when the new points are added).

3. The following data represents the daily energy output for a small solar power generator.

Date	**Energy generated/kJ**
2023-06-01	101.29
2023-06-02	112.41
2023-06-03	-1
2023-06-04	122.54
2023-06-05	1.12
2023-06-06	75.56
2023-06-07	87.10
2023-06-08	69.49
2023-06-09	96.43
2023-06-10	121.11
2023-06-11	0.00
2023-06-12	59.35

(a) What type of data is this?

(b) Which of the data points are unusual? What is unusual about each one, and can you make a suggestion for what might have caused the anomaly?

(c) What external data source do you think would be important to help you analyse this data series?

(a) A time series.

(b) The third (-1), fifth (1.12) and eleventh (0.00) are unusual; the first is a negative value which shouldn't be permitted, the second is very low and the third is zero. The first looks like an anomaly record as it's negative. The second could be caused by a measurement error or an equipment failure in the generator, and the second could also be an equipment outage somewhere in the generator.

(c) Weather data will be very important, as you need to know how much the sun has been shining.

4. A car racing team are testing their new car on a straight track. They have a camera set up alongside the track, radars to measure the velocity and an accelerometer inside the car which records the acceleration in the directions along (x) and across (y) the track. They would like to use a Kalman filter to model the position and velocity of the car.

   (a) What assumptions are they making about the noise that affects the car's movement and the accelerometer measurements?

   (b) At the start of a step, the car's position (in metres) and velocity (in metres per second) are estimated to be

$$\hat{\mathbf{x}}_{t-1|t-1} = \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix} = \begin{bmatrix} 123.0 \\ 10.0 \\ 90.0 \\ 1.2 \end{bmatrix}$$

   The accelerations were measured as $(a_x, a_y) = (1.4, -0.2)$. Use the dynamical equations

$$s = vt + \frac{1}{2}at^2$$
$$v = at$$

   to calculate the dynamical estimate of the car's position and velocity after a step of 1 second.

   (c) At the new step, the camera measures the car's position in the x direction with an error of 0.12 m, and the radars measure the x and y velocities with an error of 0.04 m/s; the velocity errors are correlated with correlation 0.5. What is the covariance matrix used for the measurement errors?

   (d) How would you expect the covariance of the state to change when a measurement is made? Does this suggest a problem with the testing set-up?

(a) They are assuming that the noise can be modelled by a normal/Gaussian distribution.

(b) Using the given dynamical equations, the update equations for position and velocity after a time interval Δt are

$$\Delta s = v(\Delta t) + \frac{1}{2}a(\Delta t)^2$$

$$\Delta v = a(\Delta t)$$

Therefore the matrix update equation for the state vector (x, y, v_x, v_y) is

$$\hat{\mathbf{x}}_{t|t-1} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix} + \begin{bmatrix} \frac{1}{2}a_x(\Delta t)^2 \\ \frac{1}{2}a_y(\Delta t)^2 \\ a_x \Delta t \\ a_y \Delta t \end{bmatrix}$$

Putting in the provided numbers, for a step of $\Delta t = 1$ second, the updated estimate is

$$\hat{\mathbf{x}}_{t|t-1} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 123.0 \\ 10.0 \\ 90.0 \\ 1.2 \end{bmatrix} + \begin{bmatrix} 0.7 \\ -0.1 \\ 1.4 \\ -0.2 \end{bmatrix} = \begin{bmatrix} 213.7 \\ 12.1 \\ 91.4 \\ 1.0 \end{bmatrix}$$

(c) The three variables measured are x, v_x and v_y. The covariance matrix is

$$\mathbf{\Sigma} = \begin{bmatrix} 0.12 & 0 & 0 \\ 0 & 0.04 & 0.02 \\ 0 & 0.02 & 0.04 \end{bmatrix}$$

(d) The covariance in the measured directions is expected to be reduced when a measurement is made. This means the variance in the x, v_x and v_y variables but not in y as this isn't measured. This is potentially a problem with the test set-up, and another camera or an overhead drone to measure the y position would be an improvement.