Problem 1: What would you interpret the different fluctuations in the picture to mean (this is part of the problem of dealing with data, at the end of the day you need to describe what you think the data tells you)?

The small fluctuations represent the day/night cycle temperature swings. The fluctuations on a longer timescale represent temperature trends that the average person might call "a cool spell."

Problem 2: For each of the data sets identify the portions of the month that are not well represented by the mean. What fraction of the month does this account for (an estimate as opposed to a calculation is what we are looking for here). How did the standard deviations come into your answer?

In the first graph, the data from days 7 through 12 or so are not well represented by the mean. Similarly in the second graph it is the data from days 17 through 24 or so which are not well represented by the mean. In both cases this is about 1/4 of the month. We see that the data which is not well represented by the mean also falls outside one standard deviation from the mean.

Problem 3: Compare the two figures. Explain how the sample average is different from the average shown in Figure 1. What is the formula for the sample average at some moment in time?

The sample average is different because it averages multiple data sets at the same time value. In Figure 1 the average is for one sample, but over the entire time interval. If f, g, h, r are the different data sets, then in this case the formula for the sample average s at time t is

$$s(t) = \frac{1}{4}[f_t + g_t + h_t + r_t],$$

and this formula holds for every t. Compare this with the formula for the mean of a function over the entire time interval

$$\langle f \rangle = \frac{1}{N}[f_1 + \cdots + f_N],$$

where here N is the number of data points in the data set.

Problem 4: Use all the figures to choose a time period for a beach camping trip. Justify your answer.

The beginning of August is the best choice. From Figure 3 we see that the variance of the different records is lowest, and from Figure 2 we see that the temperature is high. If the data is representative we would expect early August to be the best choice. The end of August is also a good choice, but the variance is slightly higher, so not quite as good as the beginning of August.

We use instruments to take measurements of what we're interested in like temperature. We record the values of those measurements in a list called a data set. We'll write

$$f = \{f_1, f_2, \ldots, f_N\}$$

where each f_i is a measurement, and there are N measurements in the list. In the plots in the the presentation the green lines showed the average temperature. The average, or mean, of the data finds the centre of the data.

Ex 1. $f = \{1, 2\}$. The number half way between 1 and 2 is $\frac{1+2}{2} = 1.5$

Ex 2. $f = \{1, 2, 3\}$. The number in the middle of 1,2 and 3 is $\frac{1+2+3}{3} = 2$

Using these examples as inspiration we define *the mean* of the data set $f = \{f_1, f_2, \ldots, f_N\}$ to be

$$\langle f \rangle = \frac{1}{N}(f_1 + f_2 + \cdots + f_N) = \frac{1}{N}\sum_{i=1}^{N} f_i$$

Problem 1: Calculate the mean for the following data sets:

a) $f = \{1, 2, 3, 4, 5\}$

$$\langle f \rangle = \frac{1}{5}(1 + 2 + 3 + 4 + 5) = 3$$

b) $f = \{1, 0, 1, 4, 10\}$

$$\langle f \rangle = \frac{1}{5}(1 + 0 + 1 + 4 + 10) = \frac{16}{5} = 3.2$$

Problem 2: Take two data sets $f = \{f_1, f_2, \ldots, f_N\}, g = \{g_1, g_2, \ldots, g_N\}$. We say that $f + g = \{f_1 + g_1, f_2 + g_2, \ldots, f_N + g_N\}$. Use the definition to show that

a) $\langle f + g \rangle = \langle f \rangle + \langle g \rangle$

$$\begin{aligned}\langle f + g \rangle &= \frac{1}{N}[(f_1 + g_1) + (f_2 + g_2) + \cdots + (f_N + g_N)] \\ &= \frac{1}{N}[f_1 + f_2 + \cdots + f_N] + \frac{1}{N}[g_1 + g_2 + \cdots + g_N] \\ &= \langle f \rangle + \langle g \rangle\end{aligned}$$

b) $\langle af \rangle = a\langle f \rangle$, where a is any number and $af = \{af_1, af_2, \ldots, af_N\}$

$$\begin{aligned}\langle af \rangle &= \frac{1}{N}[af_1 + af_2 + \cdots + af_N] \\ &= a\left(\frac{1}{N}[f_1 + f_2 + \cdots + f_N]\right) \\ &= a\langle f \rangle\end{aligned}$$

When an operation satisfies these two properties we call it a *linear operator*.

Ex 3. Let $f = \{2, 2, 2, 2, 2\}$, $g = \{0, 0, 1, 3, 6\}$. Then

$$\langle f \rangle = \frac{1}{5}(2 + 2 + 2 + 2 + 2) = \frac{10}{5} = 2$$

$$\langle g \rangle = \frac{1}{5}(0 + 0 + 1 + 3 + 6) = \frac{10}{5} = 2$$

But these two data sets look very different (Make a graph for yourself). From the graphs we see that g has different values but f is all the same value. We want a mathematical way to describe this difference. Since the mean finds the middle of the data we can use it to measure how much the data varies:

$$f_i - \langle f \rangle = \text{ distance of } f_i \text{ from the mean}$$

but this number could be positive or negative. We want only positive numbers because we care about how far from the mean f_i is, not whether it's above or below. Remember that for any number a, even if it's negative, we have

$$a^2 \geq 0$$

We define *variance* as

$$\text{var}(f) = \frac{1}{N}\sum_{i=1}^{N}(f_i - \langle f \rangle)^2 = \frac{1}{N}[(f_1 - \langle f \rangle)^2 + (f_2 - \langle f \rangle)^2 + \cdots + (f_N - \langle f \rangle)^2]$$

Problem 3:

a) Using f from Example 3, and the fact that $\langle f \rangle = 2$, find $\text{var}(f)$

$$\text{var}(f) = \frac{1}{N}\sum_{i=1}^{N}(f_i - \langle f \rangle)^2 = \frac{1}{5}\sum_{i=1}^{N}(2 - 2)^2 = 0$$

b) Using g from Example 3, and the fact that $\langle g \rangle = 2$, find $\text{var}(g)$

$$\begin{aligned}\text{var}(g) &= \frac{1}{5}[(g_1 - \langle g \rangle)^2 + (g_2 - \langle g \rangle)^2 + \cdots + (g_5 - \langle g \rangle)^2] \\ &= \frac{1}{5}[(0 - 2)^2 + (0 - 2)^2 + (1 - 2)^2 + (3 - 2)^2 + (6 - 2)^2] \\ &= \frac{1}{5}[4 + 4 + 1 + 1 + 16] = \frac{1}{5}[26] = 5.2\end{aligned}$$

c) Our idea was that variance should capture the idea of the spread of the data. Do the variances of from a) and b) show which of f and g is more spread out? Yes. We have $\text{var}(f) = 0$, corresponding to no spread in the data, and $\text{var}(g) = 5.2 > 0$, which tells us that the data in g is more spread out. In general, higher variance indicates that the data is more spread out from the mean.