UNIVERSITY OF
WATERLOO

Faculty of Mathematics
Waterloo, Ontario N2L 3G1

Centre for Education in
Mathematics and Computing

Intermediate Math Circles

March 22-23, 2016

Data - Mean and Variance

The Environmental and Geophysical Fluid Dynamics Group Crew:
Marek Stastna, Kris Rowe, Justin Shaw, David Deepwell, Aaron Coutino, Jared Penney.

Introduction

Now that we know how to compute the mean, or in every day words, the average, let's think back to the temperature data and how the average over a month might look. In figure 1 we show the temperature records for two hypothetical Augusts (we made them up).

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)?

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?

The mean provides a decent idea of the overall data (we picked it to work out to be just under 17.5 degrees Centigrade for both data sets), but it is pretty inconvenient in terms of daily predictions. To see something better, the upper panel of Figure 2 shows 4 sample Augusts (we made these up too). In the bottom panel I did a different average. Instead of averaging over time, we averaged over the four samples at EACH point in time. This is called a *sample average* or sometimes an *ensemble average*.

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?

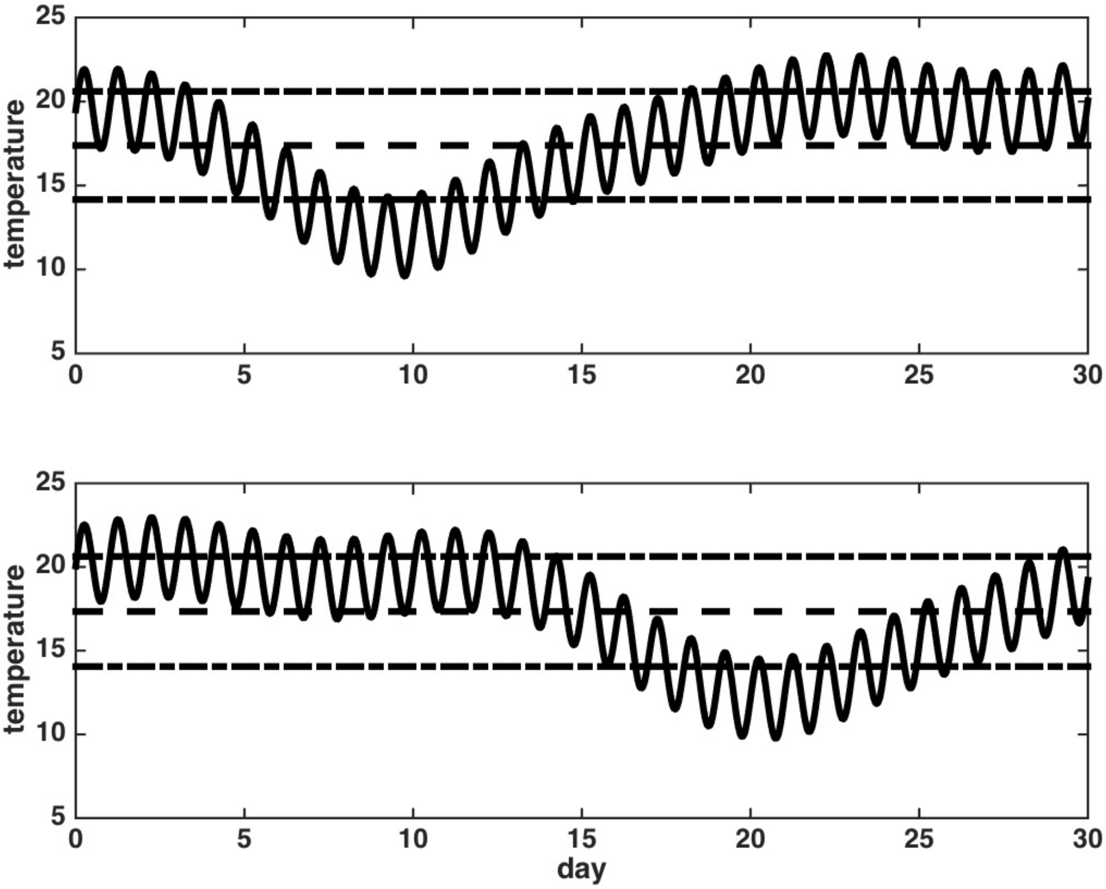

Figure 1: The time records for two different Augusts. The mean is shown by a dashed line. It is essentially the same for both samples. The mean plus one standard deviation and the mean minus one standard deviation are shown as dot-dashed lines.

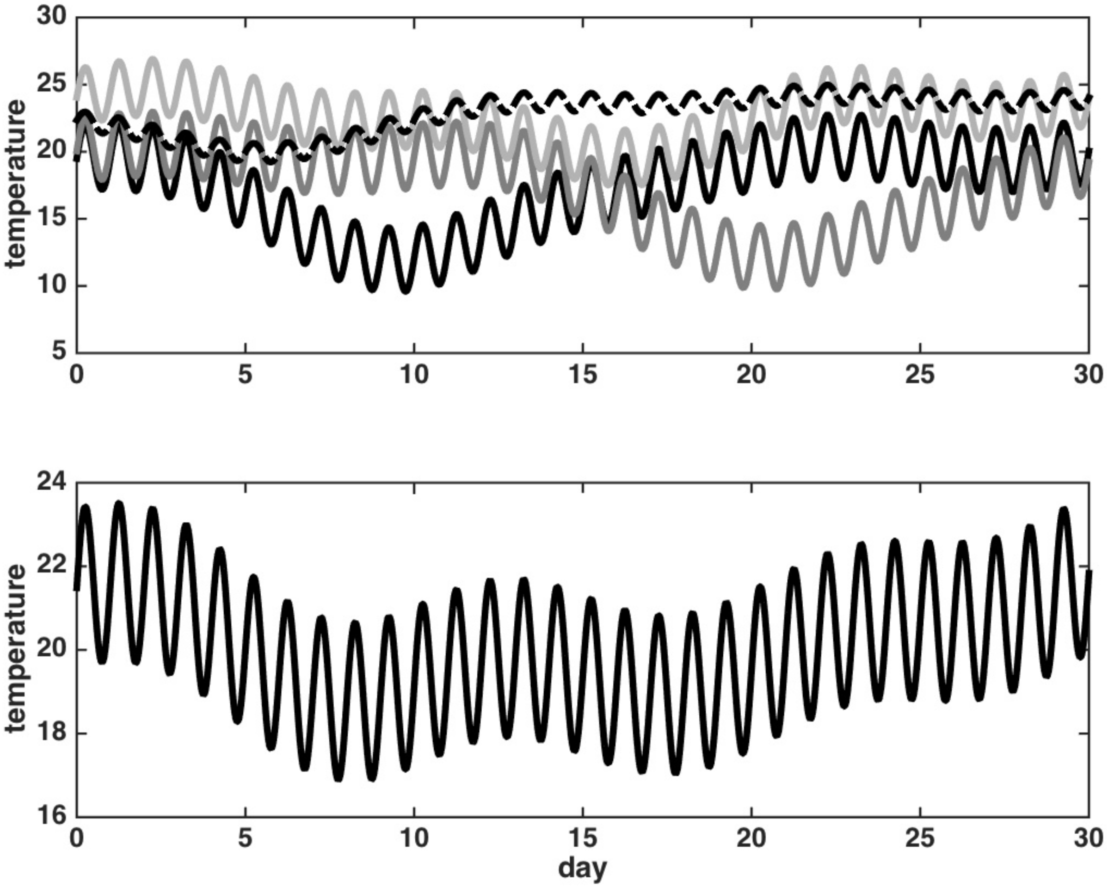

Figure 2: The time records for four different Augusts (upper panel). In the lower panel I compute the average of the four sample Augusts at EACH moment in time.

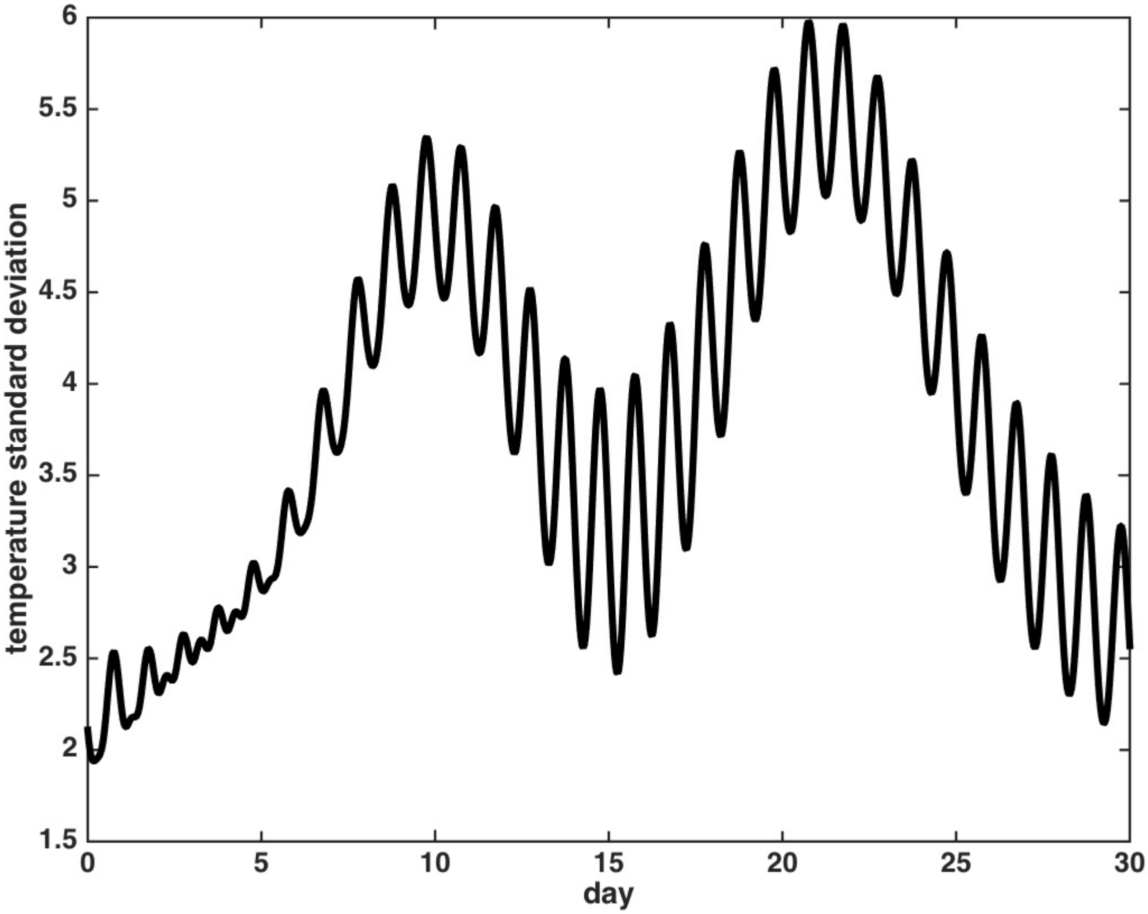

Figure 3: The standard deviation of the four sample Augusts.

OK, so the two types of mean, or averaging make sense. How would we think about the variance or the standard deviation? Well if you recall how we calculated this quantity you will recall that this is a computationally heavy task (when doing it by hand). But we can do it on a computer very easily. In Figure 3 we show the standard deviation (this is the square root of variance) as a function of the day. It says, that some days have temperature variations that are much bigger than this on other days.

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

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

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

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

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

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 $\text{var}(f)$

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

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?