32

Recursion and Averages

(This section optional)

The Arithmetic Average

You should remember the arithmetic average. Given n data points, their arithmetic average is their sum divided by n. Now suppose that we have the average of n numbers, A_n. We are given a new data point x and we would like to compute the new average of all n + 1 numbers, A_{n+1}. Many people simply add up all n + 1 numbers and then divide by n + 1. However, this seems inefficient since we added the first n numbers to get A_n. Also, what if we have A_n but we have lost the n numbers? And do we want to store all of those numbers anyway? What if there are a million numbers?

The solution follows from the fact that the sum of all n + 1 numbers is the sum of the first n numbers plus the last number. We can get the sum of the first n numbers just by multiplying A_n by n. Hence we have this recursive formula:

$$A_{n+1} = \frac{n \cdot A_n + x}{n+1}, \quad A_0 = 0$$

Recursive Formula for Averages

This formula is an excellent way to compute running averages. That is, as we collect the data we keep a running calculation of the average. In order to do this we start with $A_0 = 0$. The best

way to calculate the new average is to multiply n/(n + 1) times A_n. This way, we avoid actually computing the sum of the first n numbers which might be large and could cause overflow. Hence we write the formula as:

$$A_{n+1} = \frac{n}{n+1} \cdot A_n + \frac{x}{n+1}, \quad A_0 = 0$$

A Better Computational Form of the Recursive Formula for Averages

Since this form of the arithmetic average is computed continuously rather than at the end of the process, it is also called a *running average*.

Moving Averages

A popular method of tracking data is to keep moving averages. We fix a fairly small number as k, such as k = 9 or k = 5. As each data point is recorded we chart the average of the last n data points. We can do this recursively, but we must keep track of the last k points since each number will be deleted in turn from the sum. We follow the convention of representing the n'th number by x_n. Let A_n be the moving average after n points. The following box gives both the recursive and non-recursive formulas:

$$A_n = \frac{x_n + x_{n-1} + \ldots + x_{n-k+1}}{k}$$

$$A_{n+1} = A_n + \frac{x_{n+1} - x_{n-k}}{k}, \quad A_k = \frac{x_k + x_{k-1} + \ldots + x_1}{k}$$

The Moving Average of k Points (both Non-Recursive and Recursive Forms)

Exponential Averages

I have shown you moving averages because they are widely used. Exponential averages do everything that moving averages do, but usually do it better. The idea behind exponential

averages is to average all of the numbers, but to put more "weight" on the more recent data. The last data point is x_n and its weight is 1. The data point before it, x_{n-1}, we give weight p, with $0 < p < 1$. The data point before it receives weight p^2, and we continue in that manner. To get the exponential average of all n numbers we add up each number multiplied by its weight. In order to have a true average, we divide this sum by the sum of the weights. This gives us the following formula:

$$A_n = \frac{x_n + px_{n-1} + p^2x_{n-2} + p^3x_{n-3} + \ldots + p^{n-1}x_1}{1 + p + p^2 + p^3 + \ldots + p^{n-1}}$$

However, we know from the study of geometric series in Chapter 13 that since $0 < p < 1$ the denominator of the above expression is well approximated by $\frac{1}{1-p}$. This leads to a new formula:

$$A_n = (1-p)(x_n + px_{n-1} + p^2x_{n-2} + p^3x_{n-3} + \ldots + p^{n-1}x_1) = (1-p)x_n + pA_{n-1}$$

The last step follows after factoring out p from the last n – 1 terms.

For some reason it is the usual custom with exponential averages to interchange p and 1–p. Since $0 < p < 1$ it follows that $0 < 1-p < 1$ so replacing p with 1–p makes no real difference. So the usual formula for exponential averages is:

$$\begin{aligned} A_n &= p \cdot x_n + (1-p)A_{n-1} \\ A_1 &= x_1 \end{aligned}$$

The Formula for Exponential Averages

In this formula p represents the weight you put on the latest data point and 1 – p is the weight on all of the previous data. However, the weight on each data point is 1 – p times the weight on the next point.

Lastly, let me add that the best way to experiment with moving averages and exponential averages (with different p's) is by using spreadsheets. They make the execution of these formulas effortless and they provide instant graphing capabilities.

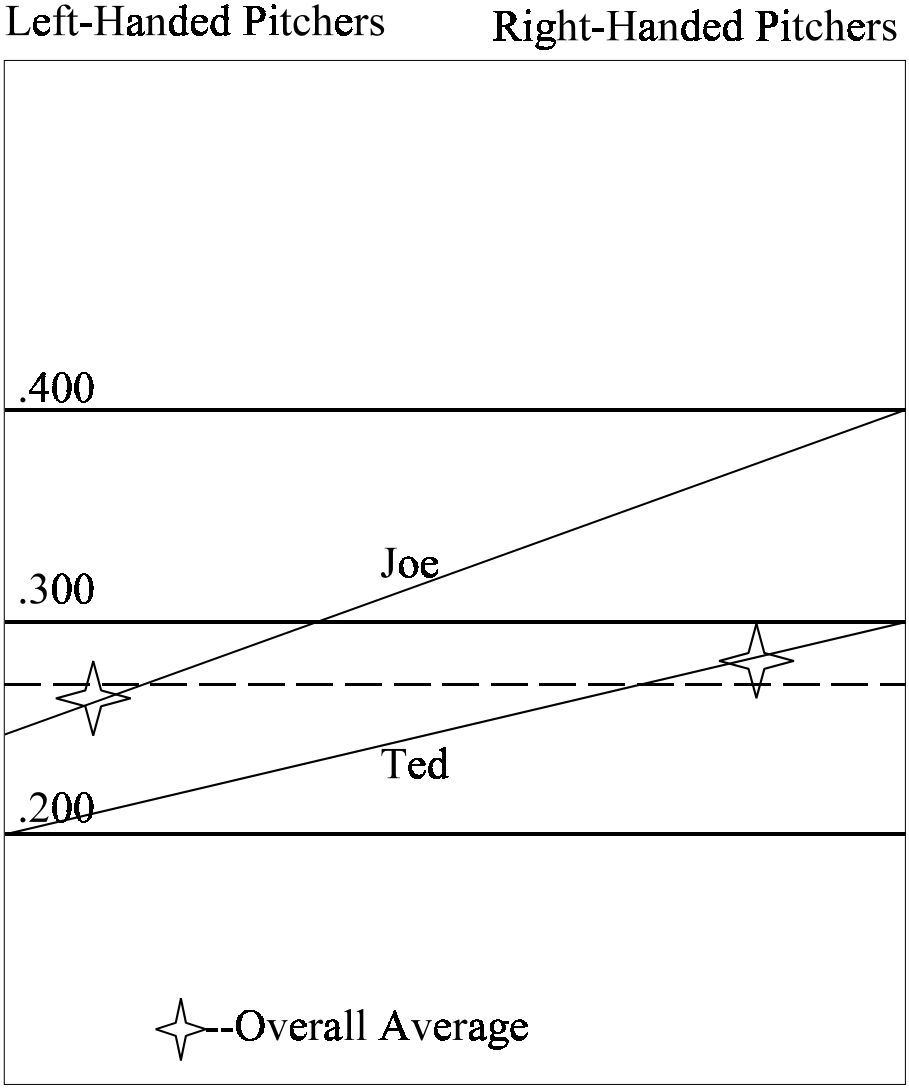
Simpson's Paradox

One of the consequences of weighted averages (that is of most averages) is Simpson's Paradox. On one hand it is something that most people think cannot happen but on the other hand happens in real life (and not just textbooks).

Consider the case of two pinch hitters (Joe and Ted). The following table shows their respective averages against right handed pitchers and left-handed pitchers, as well as their combined averages.

	Joe			Ted		
	Hits	At bats	Average	Hits	At bats	Average
Left-handed pitchers	25	100	.250	2	10	.200
Right-handed pitchers	4	10	.400	30	100	.300
Total	29	110	.264	32	110	.291

If you study the table you will see that Joe has a higher average against both left-handed and righ-handed pitchers than does Ted. However, Ted has a higher combined average. To understand why, one merely has to study the following figure:



Batting Averages for Joe and Ted