

## *Gerald's Column* by Gerald Fitton

Last month I introduced you to what is called Descriptive Statistics. Instead of building up a Probability Distribution Function (PDF) by invoking symmetry, many measurements of the statistic are used instead. The name Descriptive Statistics derives from a need to describe the statistical distribution derived from those measurements.

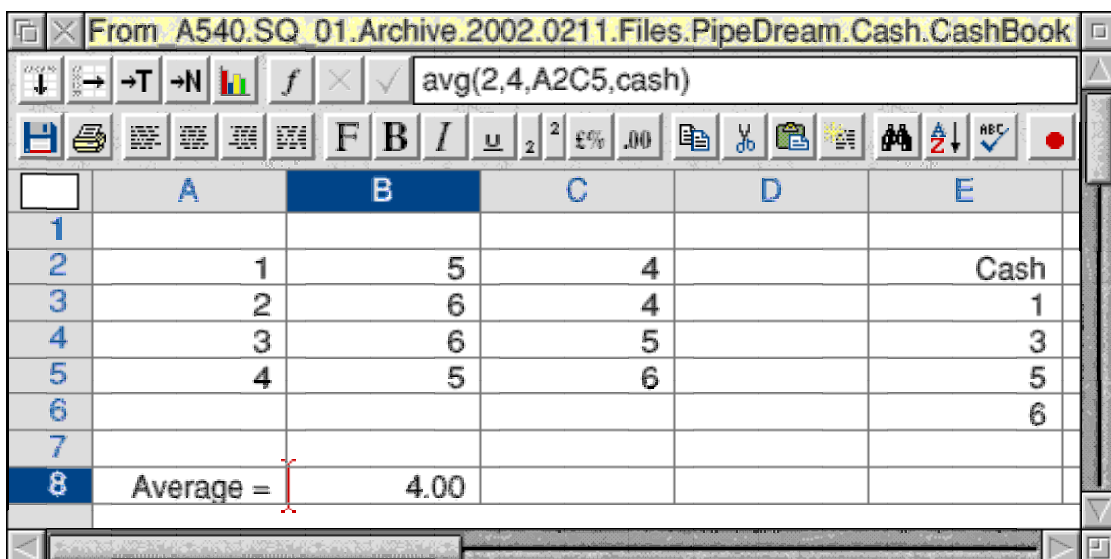
### **The Middle**

There is no doubt that one of the most important parameters of a distribution is the location of its centre. Last month I described the spreadsheet function “Average” the syntax of which is usually `avg(list)`. This spreadsheet function returns the arithmetic mean.

Don't forget that when the distribution is not symmetrical but skewed there is more than one measure of the centre. These include (1) the arithmetic mean, (2) the most popular value of the statistic (the Mode) and (3) the value of the statistic which splits the population or sample in such a way that 50% are larger and 50% are smaller than this value (the Median). The fact that there is more than one central value of the statistic does not imply that a middle doesn't exist. It does mean that you have to choose which of the 'middle' values (not limited to the three I have described) is best suited to your particular (analytical) purpose.

### **Individual Items of Data**

From time to time I am sent spreadsheet problems. I am always willing to help if I can. However, I would ask that you send me an example file with some data and describe your attempt(s) to solve your difficulty. If you send me such an example file then it saves me more time than you can possibly imagine—but that is not the most important reason I ask for an example. It is all too easy for me to misunderstand what it is that you are trying to do and where your difficulty lies. Sometimes I do lots of work solving the wrong problem either because you haven't explained it properly or because I have misunderstood.



The screenshot shows a spreadsheet window titled "From A540.SQ 01.Archive.2002.0211.Files.PipeDream.Cash.CashBook". The formula bar contains the formula `avg(2,4,A2C5,cash)`. The spreadsheet grid has columns A through E and rows 1 through 8. The data is as follows:

	A	B	C	D	E
1					
2	1	5	4		Cash
3	2	6	4		1
4	3	6	5		3
5	4	5	6		5
6					6
7					
8	Average =	4.00			

I am going to describe a common difficulty with finding arithmetic means. I used to get at least one of these every month but now the frequency is about half that. The mistake is to try to use the spreadsheet function `ave(list)` to find a mean when the data is grouped.

In my first screenshot I show the correct use of `avg(list)` to find the mean of 18 values. The function `avg(list)` finds the mean of a list of numbers. This list might be expressed as values, an array or Names (such as the Name 'Cash' in the screenshot) but it is always a list of individual values (some of which might be repeated). For example if the individual values were (2, 2, 2, 4, 4, 5) then each would have to be listed separately even though the number 2 appears three times. Note that the number 3 does not appear in the list.

### Misuse of `avg(list)`

Last month I included a summary of the Premier Division football results using as the headings for the rows and columns the goals scored by (for) and against the Home team.

In the DrawFile below (extracted from the newly release V 1.32/04 of Fireworkz) you will see the 'Home goals For' part of that table in a slightly different form. The first column is a list of the possible values taken by the statistic,  $x$ . The second column contains the frequency,  $f$ , with which the statistic takes those values. There were a total of 380 matches played in the Premier Division. In 97 of those 380 matches the Home team did not score any goals. In mathematical jargon in 97 of the 380 matches the statistic  $x$  took the value 0. As another example—in 7 of the 380 matches the Home team scored 5 goals.

Goals = $x$	Matches = $f$	Goals = $fx$
0	97	0
1	122	122
2	89	178
3	45	135
4	17	68
5	7	35
6	2	12
7	1	7
Total	380	557

Now let me describe the common mistaken use of the function `avg(list)`. The function `avg(list)` is applied either to the first or second column or even to the whole block. I can't count the number of times I've been asked why this technique is returning silly answers.

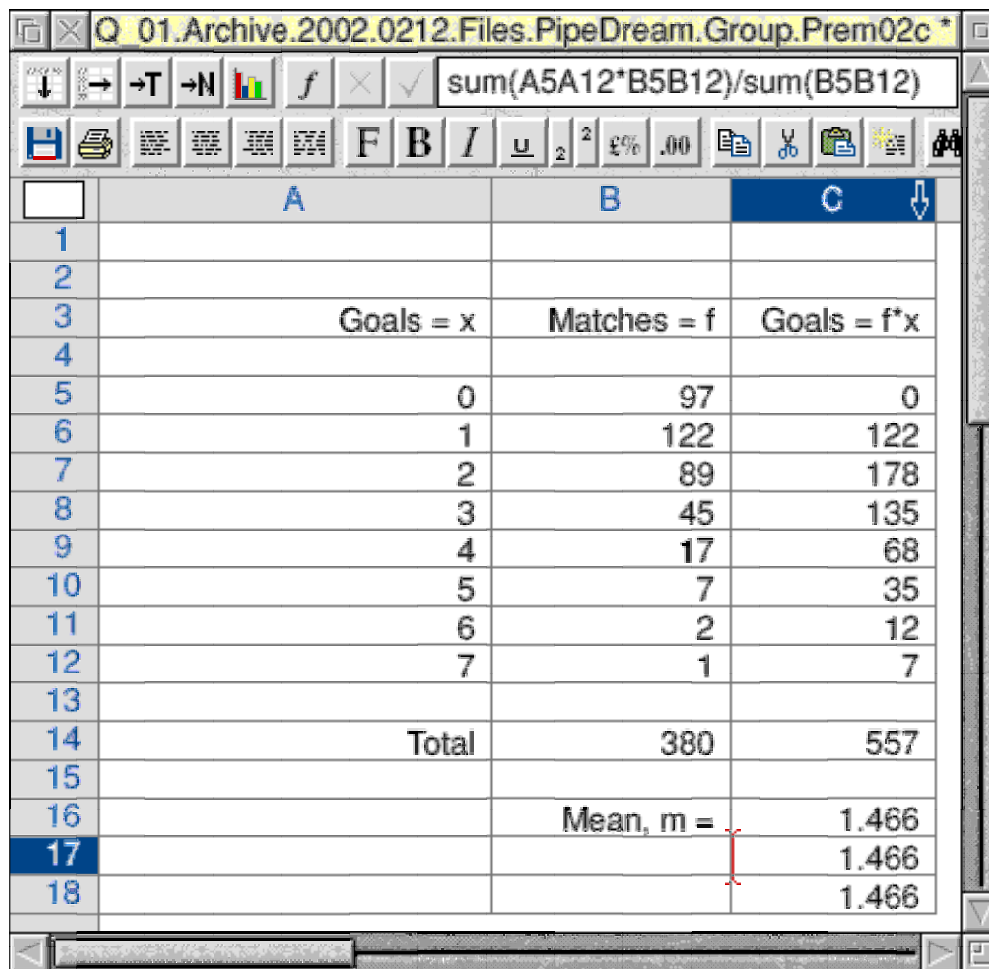
Let me remind you of something often forgotten. The average value is measured in goals! It is not measured in matches! The function `avg(list)`, listing all the frequencies, won't get you anywhere. It gives a totally wrong and meaningless answer. The average which we want is the total number of goals divided by the total number of matches. The third column gives the total number of goals. In the 'group' of matches for which the Home team scores 5 goals, the total number of goals scored is  $5 \times 7 = 35$ . The total number of Home goals scored during the season's 380 matches is 557 goals.

## Using Array Multiplication

The average number of goals per match is  $557/380 = 1.466$  (approximately). I have to say this again: there is no way of using the `avg(list)` function short of listing the results of 380 matches. The table does not contain such a list and hence `avg(list)` can not be used!

Now please refer to the screenshot of the PipeDream spreadsheet below.

One way of calculating the mean is by creating a column C in which the products  $f \cdot x$  are calculated and totalled in C14. In C16 the formula is  $C14/B14$ , the total number of goals divided by the total number of matches.



The screenshot shows a spreadsheet window titled "Q 01.Archive.2002.0212.Files.PipeDream.Group.Prem02c". The formula bar contains `sum(A5A12*B5B12)/sum(B5B12)`. The spreadsheet has three columns: A (Goals = x), B (Matches = f), and C (Goals = f\*x). The data is as follows:

	A	B	C
1			
2			
3	Goals = x	Matches = f	Goals = f*x
4			
5	0	97	0
6	1	122	122
7	2	89	178
8	3	45	135
9	4	17	68
10	5	7	35
11	6	2	12
12	7	1	7
13			
14	Total	380	557
15			
16		Mean, m =	1.466
17			1.466
18			1.466

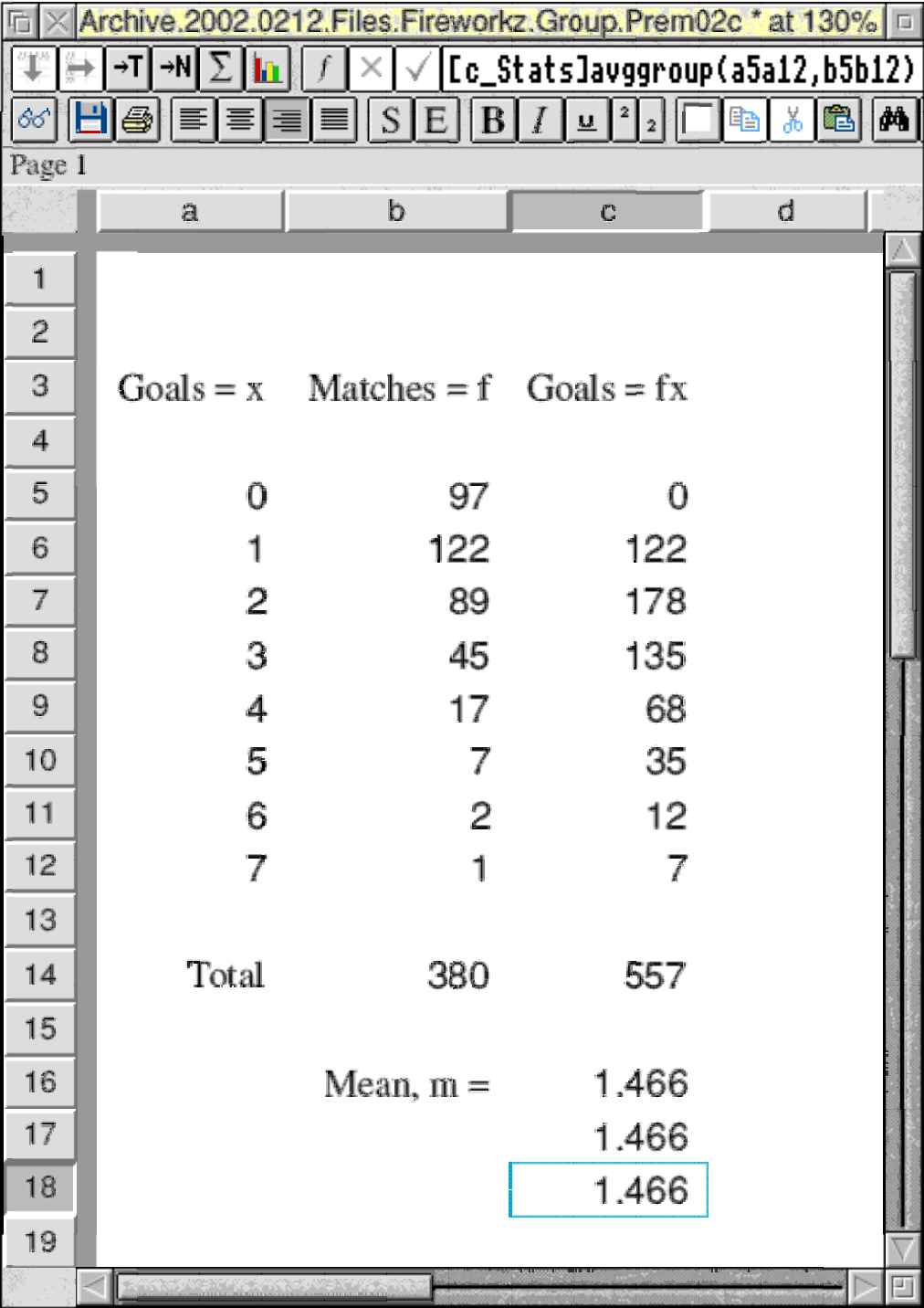
There is a slicker way which uses a facility found in many spreadsheets (including PipeDream and Fireworkz) called 'Array Multiplication'. Array multiplication is different from matrix multiplication (which I described a couple of months ago) even though this particular calculation could be expressed in terms of a matrix multiplication. In C17 I have used array multiplication to find the sum of the products of the x and f columns.

Let's look at the first part of the formula which is: `sum(A5A12*B5B12)`. The syntax can be written as `sum(range1*range2)`. The first range, A5A12 is a range of cells containing all the values of x; the second range, B5B12 is a range of cells containing all the corresponding values of f. The formula finds the sum of the  $f \cdot x$  products.

During array multiplication each value of *range1* is multiplied by the corresponding value of *range2*. The sum() function adds together these products to arrive at the total, 557. This total is divided by the total number of matches returned by the function sum(B5B12). Of course the result is 1.466 as before.

### Custom Function

PipeDream and Fireworkz have a facility called Custom Functions. Schema has a similar facility called Macros.



The screenshot of the Fireworkz file shows the syntax which I have used to call the custom function. The custom function is contained in a separate file called [c\_Stats]. The function is called "AvgGroup" (it is not case sensitive). I have used this name to indicate that it will find the mean of data which has been divided into groups. The function, AveGroup, takes two parameters, the range x (values of the statistic) and the range f (the frequencies).

```
function("AvgGroup","x:array","f:array")  
result(sum(@x*@f)/sum(@f))
```

The custom function consists of only two lines.  
You will see that it too uses array multiplication just like the longer form of C17.

This AvgGroup(x,f) function can be used to calculate weighted averages where f is the weighting function and x the list of values for which the weighted average is required.

## **Next Month**

You may wonder why I have gone to all this trouble of creating a custom function for such a simple calculation. Bear with me. All will be revealed next month when I discuss the second most important parameter used to describe a set of measurements under the heading "Measures of Dispersion (Spread or Scatter)". I use one custom function to call a second custom function, and that to call a third, all within the same [c\_Stats] document. This nesting of custom functions is a powerful feature of many spreadsheets.

## **Files Available**

All the files referred to in this article are available. If you want them and can't get a copy from the usual sources (web site, CD) then drop me a line with a blank floppy disc and I'll copy the files to your disc. Return postage and a self addressed label will be appreciated.