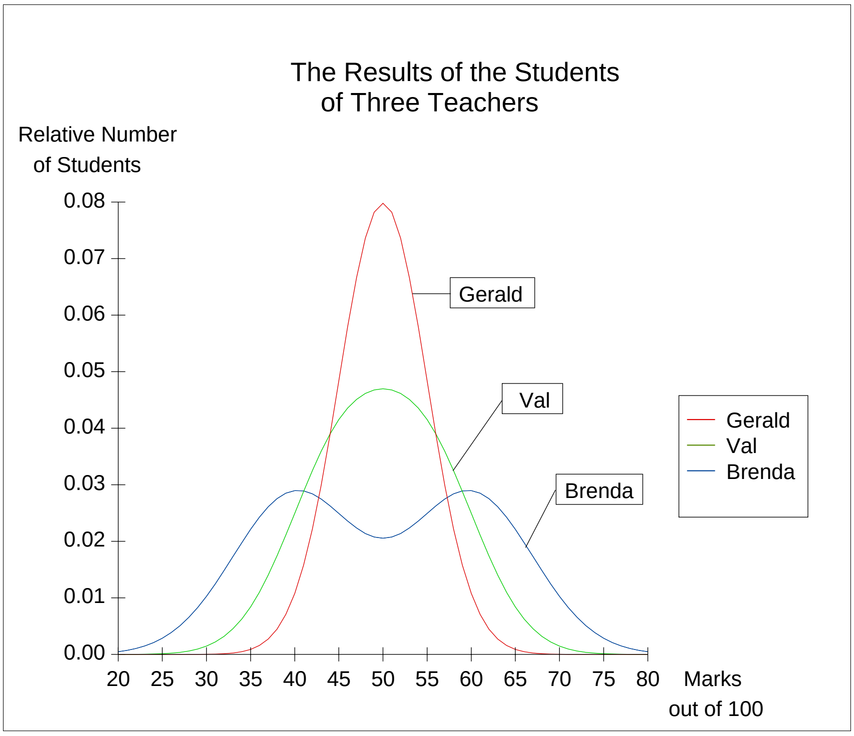
Gerald’s Column
by Gerald Fitton

The alternative to synthesising a Probability Distribution Function (PDF) using the concept of symmetry is to make many measurements. As with most things in life, every ‘solution’ to a perceived problem (eventually) introduces new problems—and so ad infinitum!

In the case of using measurements instead of synthesis to build up a PDF, the new problem is that of describing a massive amount of data using just a few numbers. Finding such simple descriptions for a PDF of measured data is called Descriptive Statistics.

Narrow or Wide

Having dealt earlier with the “Middle”, last month I introduced you to the concept of ‘width’; I used the example “Who is the ‘good’ teacher?” to demonstrate its importance. Last month I included the graph shown below. The “Middle” of all the three lines is at 50%. Brenda’s results are more ‘spread out’; the PDF of her results is much ‘wider’ than is the PDF of Gerald’s results. We need a number which describes this feature of a PDF.



Schema, Excel, Eureka, PipeDream, Fireworkz, indeed all but the simplest spreadsheets include a mathematical function which returns the “Standard Deviation” (Std). The Std is a single number which is small for narrow PDFs and large for wide ones. The units in which the Std is measured is always the same units as the statistic.

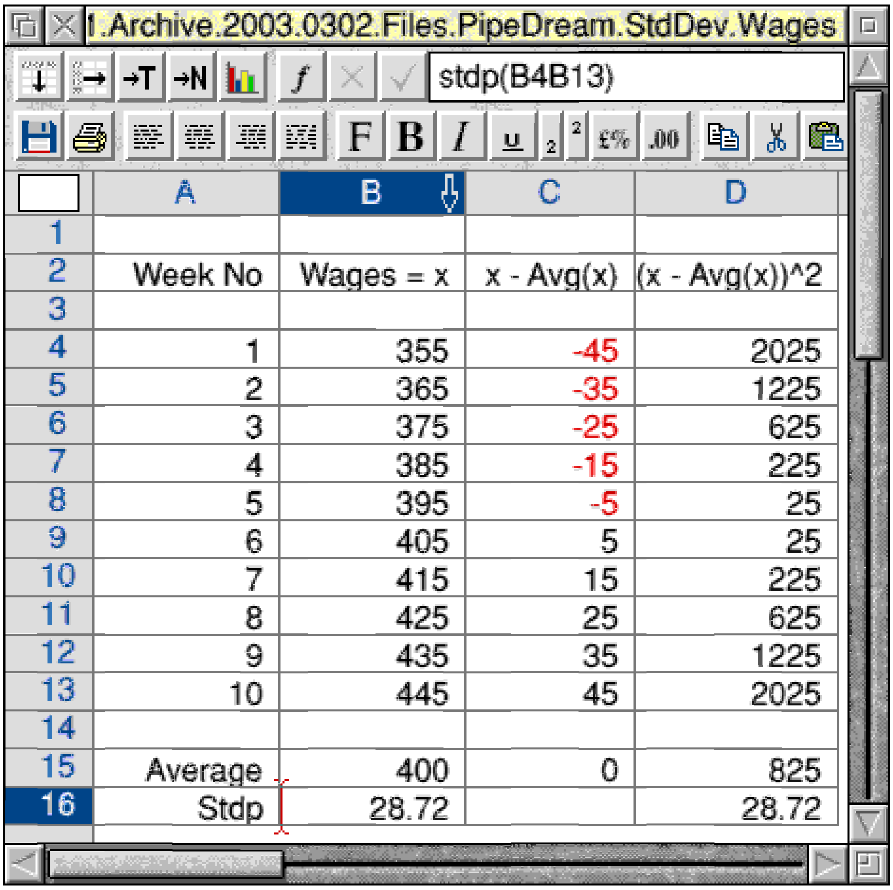
In the case of the ‘three teachers’ graph the values of the Stdp of the results of the students of Gerald, Val and Brenda are about 5%, 8% and 12% respectively. What I’d like you to do now is: Multiply the Stdp for each teacher by three to get 15%, 24% and 36%. Look at the graph and you’ll see that (for each individual teacher) not many students have a result which is more than 15%, 24% or 35% (respectively) away from the mean.

For example a band of results which is ±15% away from the mean of 50% spans the range 35% to 65%. This accords well with Gerald's results; most of Brenda's results fall within about two and a half Stdps of the mean.

Calculating the Std

Here's a definition. The standard deviation of a list of numbers is the square root of the average value of the square of the difference from the mean of the numbers in the list!

I shall now provide a simple example to help you understand this definition.



1.Archive.2003.0302.Files.PipeDream.StdDev.Wages

stdp(B4B13)

	A	B	C	D
1				
2	Week No	Wages = x	x - Avg(x)	(x - Avg(x))^2
3				
4	1	355	-45	2025
5	2	365	-35	1225
6	3	375	-25	625
7	4	385	-15	225
8	5	395	-5	25
9	6	405	5	25
10	7	415	15	225
11	8	425	25	625
12	9	435	35	1225
13	10	445	45	2025
14				
15	Average	400	0	825
16	Stdp	28.72		28.72

In might be in Euros or even pennies (you decide the units), the wages of our willing worker has varied between 355 and 445 on ten different weeks. The average of these ten wages (in cell B15) is 400 units. Column C shows the differences between each of these wages and the average wage. Of course, the average of column B is zero. The values in column D are the squares of those in column C. The average of these squared differences is 825 in cell D15. The square root of 825, approximately 29, is returned in D16.

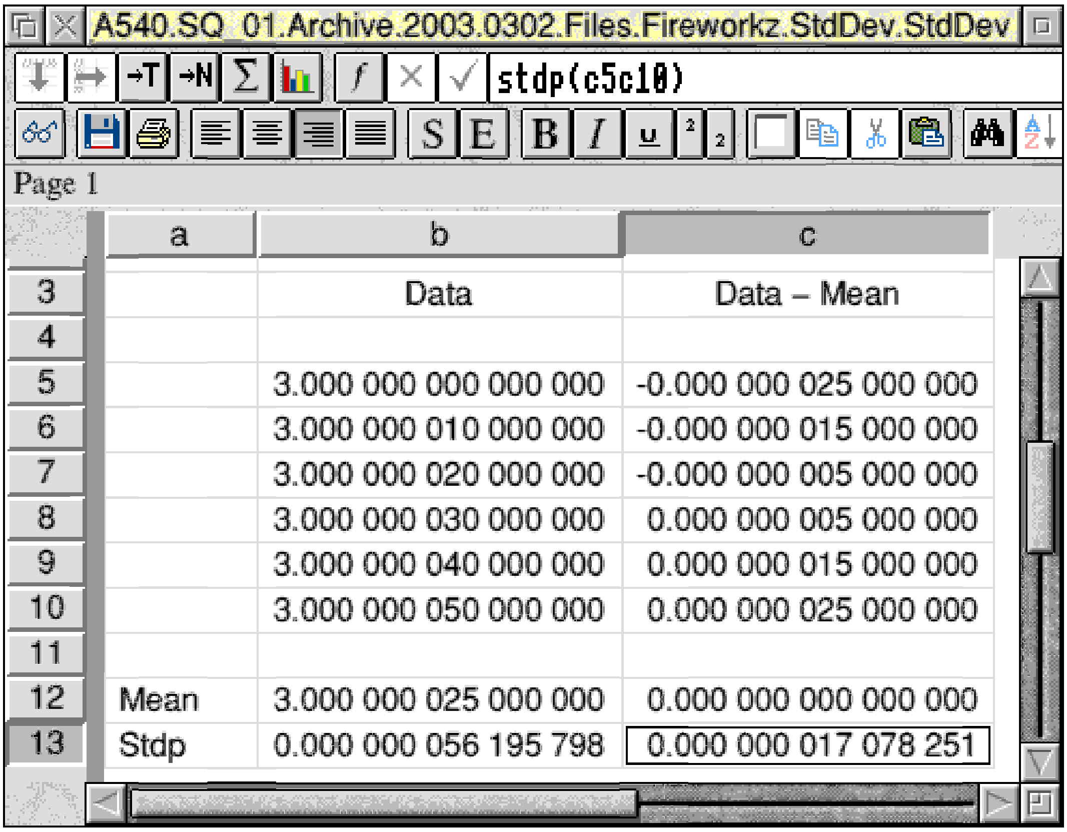
The Stdp of the numbers in column B is approximately 29. This is not a 'real' distribution of wages so you shouldn't expect most of the values to be bracketed by the rule "within three Stdp's of the mean"—even though they are!

The PipeDream formula for Stdp found in B16 is stdp(B4B13) which also returns 29.

Precision

I receive many letters about precision. One regular complaint is about “the failure of spreadsheets to calculate the standard deviation of certain sets of numbers accurately”.

One example I received was a set of numbers with properties similar to the set in the example shown below. I have deliberately chosen a ‘simple’ set of numbers so that you can see what is going on more easily.



stdp(c5c10)

	a	b	c
3		Data	Data – Mean
4			
5		3.000 000 000 000 000	-0.000 000 025 000 000
6		3.000 000 010 000 000	-0.000 000 015 000 000
7		3.000 000 020 000 000	-0.000 000 005 000 000
8		3.000 000 030 000 000	0.000 000 005 000 000
9		3.000 000 040 000 000	0.000 000 015 000 000
10		3.000 000 050 000 000	0.000 000 025 000 000
11			
12	Mean	3.000 000 025 000 000	0.000 000 000 000 000
13	Stdp	0.000 000 056 195 798	0.000 000 017 078 251

The data consists of six values ranging from exactly 3 to 3 + 5*10^{-8} .

You will see that the standard deviation calculated using the Fireworkz formula in the cell b13 is 0.000 000 056 195 798; a more accurate value (in c13) is a less than a third of this.

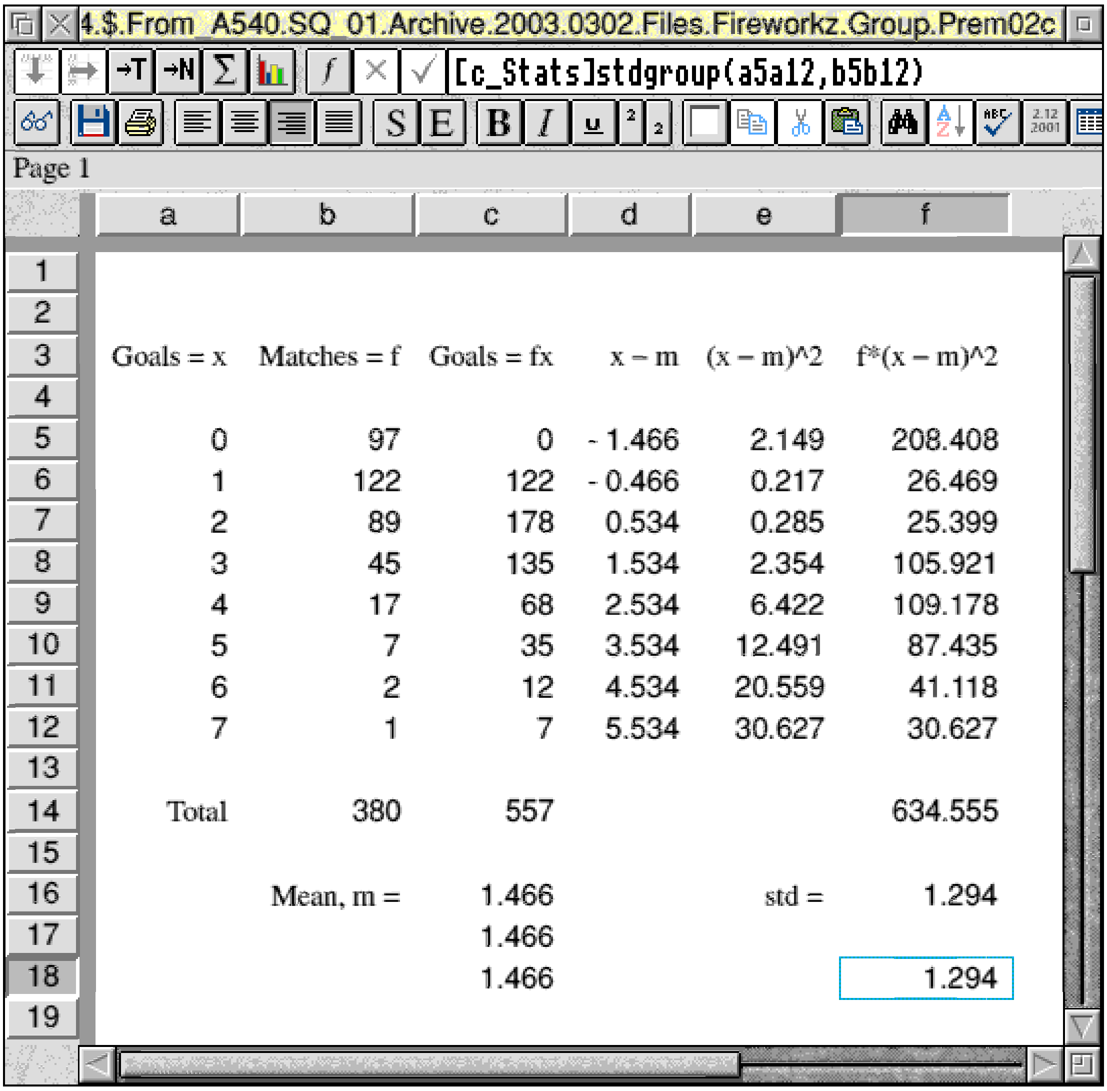
Most spreadsheets use the same algorithm to find the standard deviation of a list of ‘n’ numbers. That algorithm involves squaring the numbers in column b, subtracting n times the square of the average (in b12), taking the square root and finally dividing by n. If the differences between the original numbers are small as in our example then this algorithm fails because of the limited precision of the Floating Point Emulator. The FPE fails completely if the number 3 is changed to 4 so that the range is from 4 to 4 + 5*10^{-8}.

A more accurate algorithm is that which I show in the spreadsheet. Find the differences from the mean (as in column c) and then find the standard deviation of that set. The value in c13 is considerably more accurate than is the value in b13. Certainly it is correct to many more digits than I have displayed in the spreadsheet.

Grouped data

A couple of months ago I included a table showing the distribution of goals scored by the home team in the Premiership League during the 2001 – 2002 season. I have used this same table as my example of a calculation of the standard deviation.

As I described a couple of months ago, the value 1.466 in cells c16, c17 and c18 are values of the mean using three different spreadsheet formula to do the same calculation. The first of these is c14/b14, the number of goals divided by the number of matches. The second uses array multiplication and the third uses a custom function.



4.$.From A540.SQ 01.Archive.2003.0302.Files.Fireworkz.Group.Prem02c
Page 1

[c_Stats]stdgroup(a5a12,b5b12)

	a	b	c	d	e	f
1						
2						
3	Goals = x	Matches = f	Goals = fx	x – m	(x – m)^2	f*(x – m)^2
4						
5	0	97	0	- 1.466	2.149	208.408
6	1	122	122	- 0.466	0.217	26.469
7	2	89	178	0.534	0.285	25.399
8	3	45	135	1.534	2.354	105.921
9	4	17	68	2.534	6.422	109.178
10	5	7	35	3.534	12.491	87.435
11	6	2	12	4.534	20.559	41.118
12	7	1	7	5.534	30.627	30.627
13						
14	Total	380	557			634.555
15						
16		Mean, m =	1.466		std =	1.294
17			1.466			
18			1.466			1.294
19						

The standard deviation is returned in both f16 (the extended way) and f18 (calling a custom function to execute the same calculation but using array multiplication).

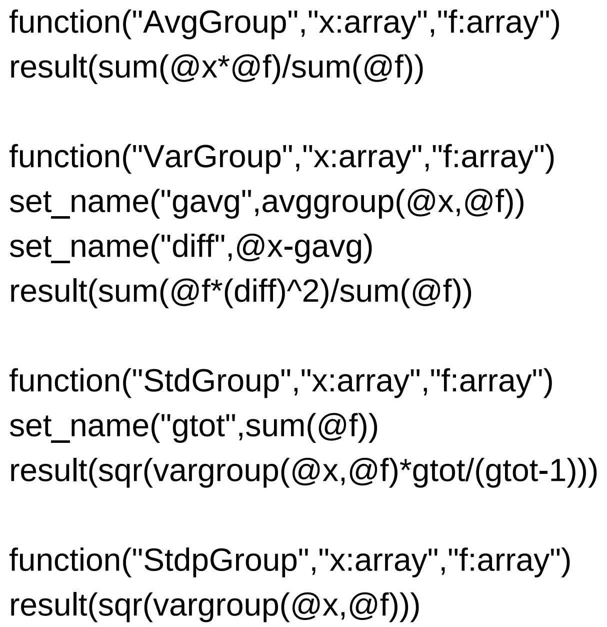
The values of (x – m), where m is that mean value, are found in column d. The values in column e are the squares of those in d. The values in column f are the values in column e multiplied by the frequencies (number of matches) from column b. The sum of the values in column f is returned in cell f14. Of course this is not the standard deviation. This sum has to be divided by the number of matches (well not quite—more about this another time) and then the square root is taken. The result in f16 is 1.294.

Custom Functions

You will appreciate that there is even more room for doing the ‘wrong thing’ when attempting to calculate the standard deviation than there is in attempting to calculate the mean of grouped (tabulated) data of this type. I shall not list all the mistakes which are sent to me with pleas for assistance. The most common one is to try to use functions such as std(list) which are built into all but the simplest spreadsheets.

In Excel, Eureka and Schema this same function is called stdev(list).

Let me emphasise that these std(list) and stdev(list) functions can not be used with grouped (tabulated) data. Either you must do it ‘the long way’ as I have described above for the value returned in f16 or you will need a custom function (sometimes called a macro).



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

function("VarGroup","x:array","f:array")
set_name("gavg",avggroup(@x,@f))
set_name("diff",@x-gavg)
result(sum(@f*(diff)^2)/sum(@f))

function("StdGroup","x:array","f:array")
set_name("gtot",sum(@f))
result(sqr(vargroup(@x,@f)*gtot/(gtot-1)))

function("StdpGroup","x:array","f:array")
result(sqr(vargroup(@x,@f)))
```

I have used the ‘Save as Drawfile’ facility from Fireworkz V 1.32/04 to create a graphic of a set of four Fireworkz custom functions. All four of these custom functions accept two arguments “x:array” and “f:array”. The x array is the array listing all possible values of the statistic; in my example this is the number of home goals from column a of the spreadsheet. The f array is the frequency with which the statistic takes particular values of x; in my example this is the number of matches from column b. For example, the entry in row 10 of the table means that in 7 of the 380 matches the home team scored 5 goals.

The four custom functions return the Average, the Variance, the Standard Deviation which I have called Std and the slightly different Standard Deviation I have called Stdp. These four custom functions can be used with data grouped into a table such as the one I am using for my example.

The custom function in f18 returns 1.294, the same value as that in f16.

Another warning! A very common error with grouped data is to get confused between the statistic (in column a) and the frequencies (in column b). A useful ‘tip’ is that the statistic is a list of (predetermined) possible outcomes and nearly always has units (such as ‘goals’) whereas the frequency is a number which you can imagine counting during the experiment.

Bill’s Problem

Last month I set you what I called Bill’s Problem. Bill has to improve the performance of the department. I have not received any answers yet. I am not surprised because that issue of Archive has not yet been published. I’m trying to get ahead in time for Christmas.

Summary

Standard Deviation is a measure of the width of a PDF. There are other measures but this is the most popular. The units in which the Std is measured is always the same units as the statistic. A band of values of the statistic about five or six standard deviations wide includes most of the data. Indeed, if a measurement falls outside this range then it is worth considering whether it is a mistake (or special case)!

Do not use avg(list) and std(list) with grouped data. Either use one of the extended methods I’ve described here (and two months ago) or, better in my view, use a custom function (macro) such as those shown in the graphic above.

All the files associated with this article are available on the monthly floppy disc, the annual Archive CD and on our web site: http://abacusline.demon.co.uk.

Next Month

Next month I shall explain why most spreadsheets include two functions with names such as std() and stdp(). The reason why there are two functions which return different answers is not simply to confuse you—even though confusion is a common consequence of this proliferation. The more observant of you will see that I have included two custom functions for standard deviation and that they will return slightly different answers.