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. Having collected a large mass of data we have to find a simple way of describing the results of our measurements. The branch of Statistics which deals with this is called Descriptive Statistics.

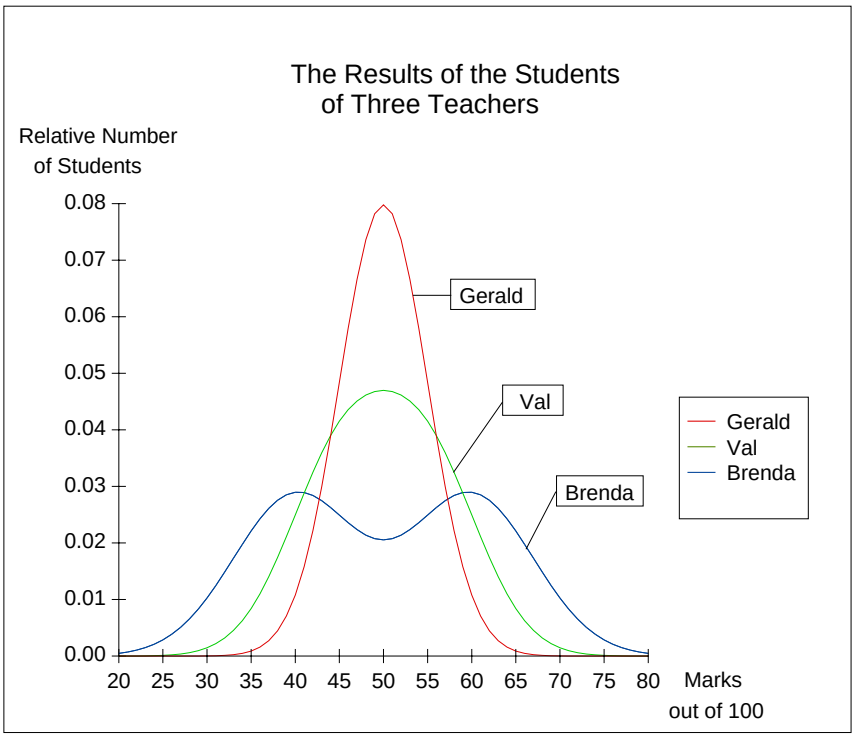
As I explained in earlier articles the two most important numbers used to describe a collection of measured data are the 'middle' and the 'width'. If the distribution is symmetrical then it is rare for there to be any ambiguity about where the middle is—but if the distribution is skewed then deciding which of many 'middles' to use can be difficult!

Last month I described the most common measure of the 'width' of a distribution, the standard deviation. All but the simplest spreadsheets provide two formulae for standard deviation; each gives a slightly different result. The existence of two different formulae for what is apparently the same thing is a common source of confusion. In this month's article I hope to make clearer when to use one and when to use the other.

But first ... **Bill's Problem**

A couple of months ago I asked what you might do if you had to improve the overall performance of Bill's department. I do like to acknowledge those who write to me but on this occasion I have had so many correspondents that it is impossible to name you all. I'll simply have to say "Thank you everyone!"

You will remember the graph below. It shows the results of three (of seven) teachers.



All who wrote to me were certain that the students should be streamed in such a way that Brenda taught only those who could keep up with her fast pace and that Gerald taught those students who otherwise might fail. Of course Bill did exactly that—but back in the 1980s streaming was considered to be very 'naughty'! We dare not admit to streaming.

Although I received many replies nobody followed them through to point out that when the students were streamed Brenda's Performance Indicators would improve and Gerald's would inevitably fall. The educational climate in those days was such that if Bill disclosed Gerald's new Performance Indicators he (Bill) would have been under great pressure to remove Gerald from the course and, maybe, even from the classroom altogether! A dreaded "Staff Training Course" would have been Gerald's minimum punishment!

So, to Bill's everlasting credit, he fiddled the results. When it came to Performance Indicators Bill insisted that all lecturers shared the honours equally. We each had two class lists, one was the list of the students we actually taught and the other a fictitious class list. This fictitious list was used for generating our individual Performance Indicators.

Perhaps someone who is into Education might let me (or Archive-on-Line) know whether streaming is still taboo or whether it has returned to favour? Are all Students 'equal'?

The Binomial Distribution

I shall use a Probability Distribution Function (PDF) called the Binomial Distribution to demonstrate how the two different standard deviation functions are used. One of these is usually called std and the other is usually called stdp; there are variants on these names.

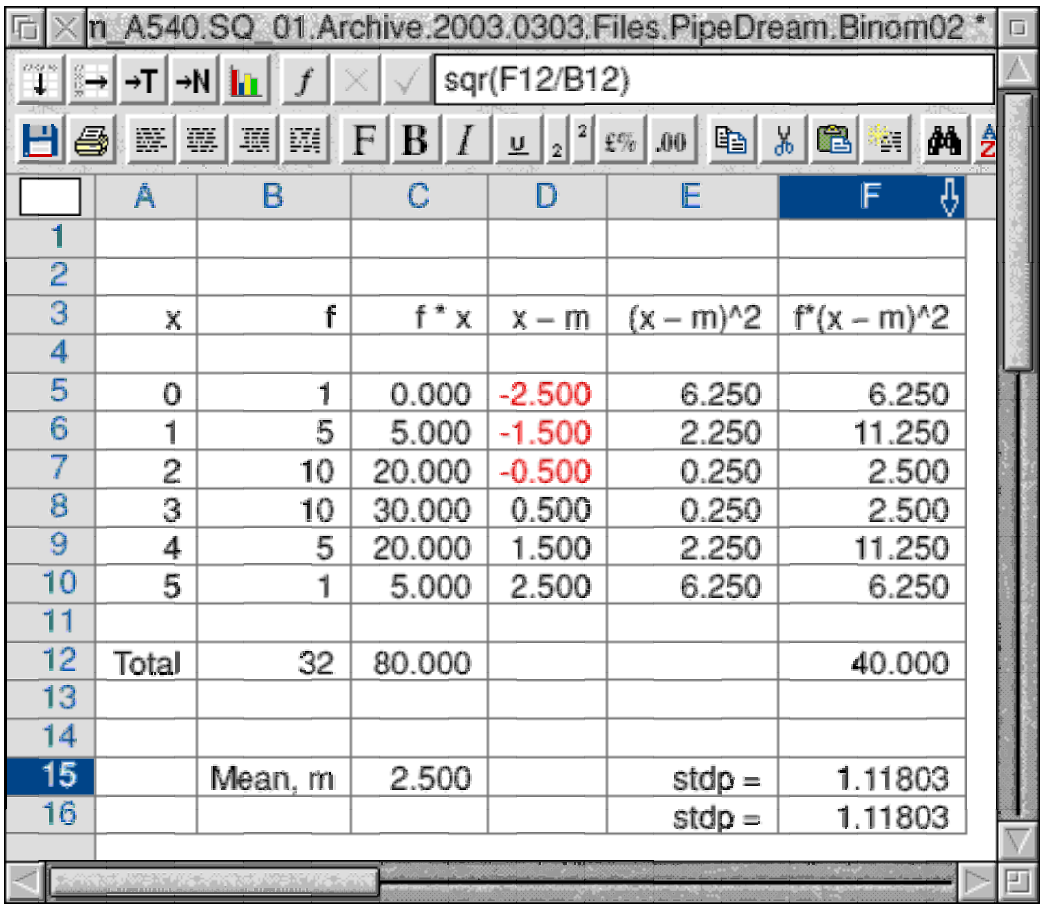
I would like you to consider the compound event of flippin' five unbiased coins. Of course you could flip one coin five times but I find it much easier to consider having five identical unbiased coins and flippin' them all at the same time. That way there is no confusion as to the exact definition of the event. It is a flippin' five coin event.

Now I'd like you to consider repeating this (flippin' five coin) event 32 times.

Of course you could get five heads as the result of every one of these 32 experiments—but that is highly unlikely. Indeed, without doing the sums, my guess is that a number expressing the the odds against this happening will be a number having about fifty digits (and the last one is a 6)! I do like writing things like that because I know that someone will fulfil my otherwise incomplete life by providing me (all of us) with the exact answer?

Rather than concentrate on that highly unlikely outcome let's go to the other extreme and think about a different but still highly unlikely result. For those who would like to be 'famous for 15 seconds' be warned that finding the odds against the outcome I am about to describe is a much harder sum—perhaps even worthy of a Colin Singleton puzzle?

The highly unlikely outcome I want you to consider is the one in which the result of these 32 experiments is exactly what you would expect if the 'law of averages' determined the result. The frequencies would be those shown in the table below where x is the number of heads (out of five) and f is the number of times (out of 32) that x takes that value. For example x is 0 (zero heads) in only 1 of the 32 experiments whereas 2 heads out of 5 'comes up' in exactly 10 of the 32 experiments.



	A	B	C	D	E	F
1						
2						
3	x	f	f * x	x – m	(x – m)^2	f*(x – m)^2
4						
5	0	1	0.000	-2.500	6.250	6.250
6	1	5	5.000	-1.500	2.250	11.250
7	2	10	20.000	-0.500	0.250	2.500
8	3	10	30.000	0.500	0.250	2.500
9	4	5	20.000	1.500	2.250	11.250
10	5	1	5.000	2.500	6.250	6.250
11						
12	Total	32	80.000			40.000
13						
14						
15		Mean, m	2.500		stdp =	1.11803
16					stdp =	1.11803

There is another way of looking at this table. If the set of five unbiased coins is flipped not 32 times but hundreds or even millions/billions/trillions/etc of times then, in approximately 10 out of every 32 experiments, the number of heads (out of five) would be two heads.

The number 10/32 could be regarded as the probability that the outcome of a single (flippin' five coin) event results in two of the five coins landing head uppermost.

The sum of all the probabilities is one.
(1 + 5 + 10 + 10 + 5 + 1)/32 = 1.

Stdp

Sometimes this standard deviation is called the Population Standard Deviation.
The 'p' at the end of 'stdp' stands for 'population'.

This is the function to use for the standard deviation when we use symmetry to create the PDF in such a way that the numbers in our table are exactly representative of the whole population of similar events. In the case of the flippin' five coin event the population of similar events is theoretically infinite (which is different from 'very large'); it includes not only all the occasions when five coins are flipped but all the future occasions when this will or even could happen! The distribution of the set of 32 experiments shown in the table is not any old sample of 32 experiments. No! This is a carefully constructed set which is exactly representative of the probabilities that x will take its six possible values (0 to 5).

The numbers in the column headed “f” (for frequency) is generated by a function found in many spreadsheets usually called “combine” or “binomial”. This function is built into Fireworkz but not PipeDream. Nevertheless it is relatively easy to write a PipeDream custom function which will find these values. I have collected together a set of statistical custom functions in a file called [c_Stats] which I keep in my PipeDream.User.Library. The cell [Binom02]B5 contains [c_Stats]ncr(5,A5).

My custom function “ncr(n,r)” returns the number of ways in which ‘r’ items can be selected from a set of ‘n’ items. There are 10 ways of selecting 2 items from a set of 5.

One of the most important points to understand is that the frequencies in the (x, f) table shown above are exactly proportional to those of the theoretical distribution you would expect (or achieve) in an infinite population of flippin’ five coin events. Consequently the standard deviation (using the stdp function) will be exactly the same for our carefully selected sample as it is for the whole (infinite) population fo flippin’ five coin events.

In the PipeDream file [Binom02] I have calculated the stdp (standard deviation) in two ways. The two identical answers are in cells F15 and F16.

Permit me to remind you that the standard deviation has the same units as the statistic. In this example the statistic is the number of heads (out of five) and takes values between zero and five. The standard deviation is a measure of the ‘width’ of the distribution.

The stdp is a rather difficult to explain average—but here we go! Find the arithmetic mean; find the differences from the mean; square this difference; find the average of these squared differences; finally take the square root! This square root (you might recognise it as the root mean square – ‘RMS’ – of the differences from the mean) is the stdp.

The various stages in this (difficult to explain) calculation are shown in the PipeDream spreadsheet. A complication in understanding how the spreadsheet achieves this is that the data is grouped. Persevere and you’ll get there—or, alternatively, believe me!

The number 40 in F12 is the sum of the squared differences. This is divided by 32, the total number of experiments from B12. The square root of (F12/B12) is returned in F15. As a by-the-way, the exact value of this stdp is half the square root of 5.

My second method is to use the custom function I introduced last month. The formula in [Binom02]F16 is [c_Stats]stdpgroup(A5A10,B5B10). The custom function “stdpgroup” is in the file [c_Stats]. If you want to use it regularly then I suggest that you, like me, place it in your !PipeDream.User.Library directory. It will be called from there whenever it is needed. On the Archive monthly disc you will find a similar [c_Stats] file which you can place in your !Fireworkz.User.Library directory. Of course this custom function returns the same value in F16 as that returned in F15 by what I might call the ‘longhand method’.

Std

This standard deviation (std without the ‘p’ for population) is the one to use if your PDF has been created without recourse to theory (or symmetry) but by direct measurement. If you have a set of measured data and you want to calculate a number which represents the ‘width’ of the distribution then it is almost certain that std is the standard deviation to use.

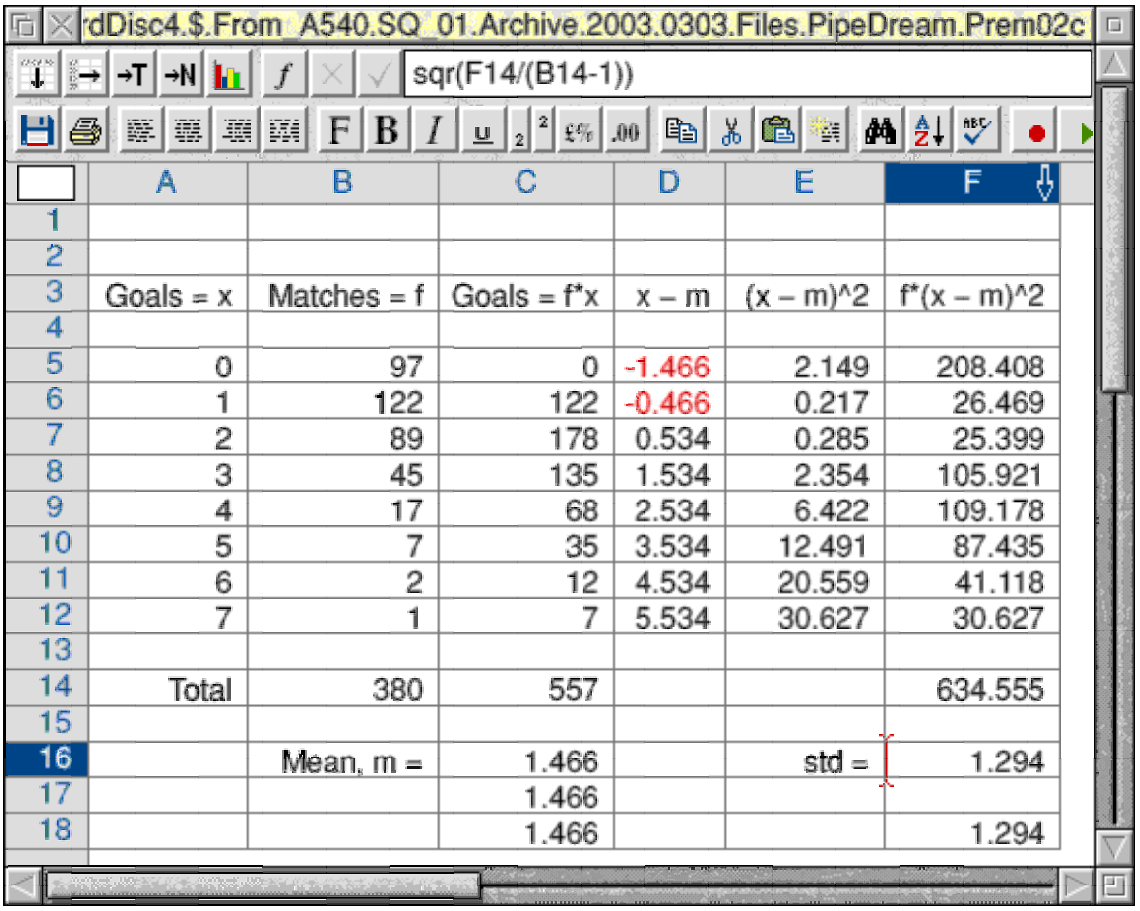
For example if you measure the heights of a hundred people and want to describe the 'width' of the PDF of these heights you would use the std and not the stdp function. How can you remember whether to use std or stdp? Nearly always you'll be making measurements rather than finding the standard deviation theoretically using symmetry so it is almost certain that you'll want to use std and not stdp.

Here is a 'tip'. The std function always gives a slightly larger value than does the stdp function. So, if you forget which is which, then, as a 'rule of thumb', calculate both std and stdp and (almost always) use the larger value!

The custom function [c_Stats]stdgroup(x_array,f_array) will calculate this std value for grouped data. In [c_Stats] I have included a reminder "Used with measured data" !

How do they differ?

The goals scored in premiership league matches is something which is measured. This table is not the result of a theory. Consequently the standard deviation to use is std.



	A	B	C	D	E	F
1						
2						
3	Goals = x	Matches = f	Goals = f*x	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

In the screenshot you'll see that the average (of the squared deviations) calculated in cell F16 is not one in which the divisor is the total number of matches (from B14). It is (B14 − 1). This slight reduction in the divisor (from B14 to B14 − 1) is what makes std slightly larger than stdp! When the total number of observations is large (380 is 'large') there isn't a lot of difference between std and stdp. The difference is important only when the number of observations is 'small', say 20.

Why use std and not stdp?

Now this is a most interesting question!
However, it is not one which I shall answer this month!
Have you any idea why the use of std is more ‘correct’ than using stdp for measured data?

Summary

Standard Deviation is the most usual measure of the ‘width’ of a PDF. The units in which it is measured is always the same units as the statistic. If a data point is more than two and a half or three standard deviations from the mean then it should be considered unusual.

Unusual data points are called ‘Outliers’ (I’m never sure how to spell “outlier”/”outlyer”). The most common cause of an outlier is that the data point has been recorded incorrectly! The usual advice is to discard it and not use it in the calculation of the standard deviation!

All but the simplest spreadsheets contain two standard deviation functions. Generally both of these are useless because the data you collect will almost certainly be grouped and these functions can not be used with grouped data. Your only two realistic alternatives are:
(a) the method I have have called the ‘longhand method’
(b) using a custom function.

There are two functions for standard deviation; generally you will want to use the one which gives the larger number. The one I have called “[c_Stats]stdgroup(x_array,f_array)” (without the ‘p’ for population) is almost certainly the one to use with measured data.

Files

All the files to which I have referred in this article including custom functions for the arithmetic mean and standard deviation of grouped data for both PipeDream and Fireworkz are on the Archive monthly disc and on my website at http://www.abacusline.demon.co.uk.

Those of you who regularly use either PipeDream of Fireworkz for analysing measured data should place the file [c_Stats] in your User Library. Functions such as those which find the average or standard deviation of grouped data will be ‘called’ automatically from your Library whenever they are needed.

If you are able to calculate the odds of getting either a ‘full house’ of heads from a set of 32 flippin’ five (unbiased) coin events and/or getting an exact ‘law of averages’ result then please let me know the values. Email is preferred to snail mail.