Gerald's Column
by Gerald Fitton

In last month's article I provided you with examples of the use of std and stdp for estimating the standard deviation of a set of data. In this month's article I shall give you a reason for accepting that std, and not stdp, should be used with measured data. I shall not prove that the correction factor used is the 'best' factor because the proof is a bit obscure.

But first …

PipeLine

I am not sure exactly when, but I think my first PipeLine column appeared in Archive during 1989 and the last last one in 1993. It happened like this.

After giving up my Acorn Master computer (which is still somewhere in the loft) and buying my first Archimedes, I continued to use WordWise (using an emulator) for my word processing and for a miscellany of articles for Archive. One day in 1989 Paul asked me if I would write a review of Colton Software's new integrated package, PipeDream 2.

In those days the operating system for the Archimedes was the single tasking Arthur (RISC OS 1); PipeDream 2 was single tasking. Even so, immediately I started to use it I was hooked. Wordwise (from Computer Concepts) wasn't a patch on PipeDream 2. That was only the start. I tried out some of the simpler spreadsheet functions and it was a total revelation to me how easy it was to solve arithmetical problems on a spreadsheet instead of writing a BASIC program. Slower but much easier.

Soon after my PipeDream review I started writing a regular PipeLine Column. PipeLine contained hints, tips and sometimes a detailed worked example of the uses of PipeDream. Archive readers sent me their PipeDream problems and I found (or tried to find) solutions. These worked solutions appeared in my longer PipeLine articles.

In July 1990, with encouragement from Colton Software, I started a User Group for PipeDream users with articles sent to members on floppy discs. PipeLine and ZLine (for Fireworkz users) were merged in 1998 to form GoldLine. Articles similar to those of the original PipeLine but covering a broader range of topics are distributed to members on CD.

I have been asked by a couple Archive-on-Line members to consider writing a few Archive articles about the use of PipeDream. I am certainly willing to do so (I might take some of the better articles from the GoldLine CD so it would be an easy option) but I'd like to know what you, the wider audience, think about that idea? Drop me a line.

Jargon

See if you can remember the definitions of the six words in the table below as you read the statement "Data is the set of Observations of the Statistic which we have made on a Sample of Events selected from a Population (of relevant Events)". If you need some clarification of the meanings of these words then look again at the April 2002 issue of Archive where I devote about 750 words to these definitions.

	One	Many	All
Happening	Event	Sample	Population
Measurement	Observation	Data	Statistic

Royal Command!

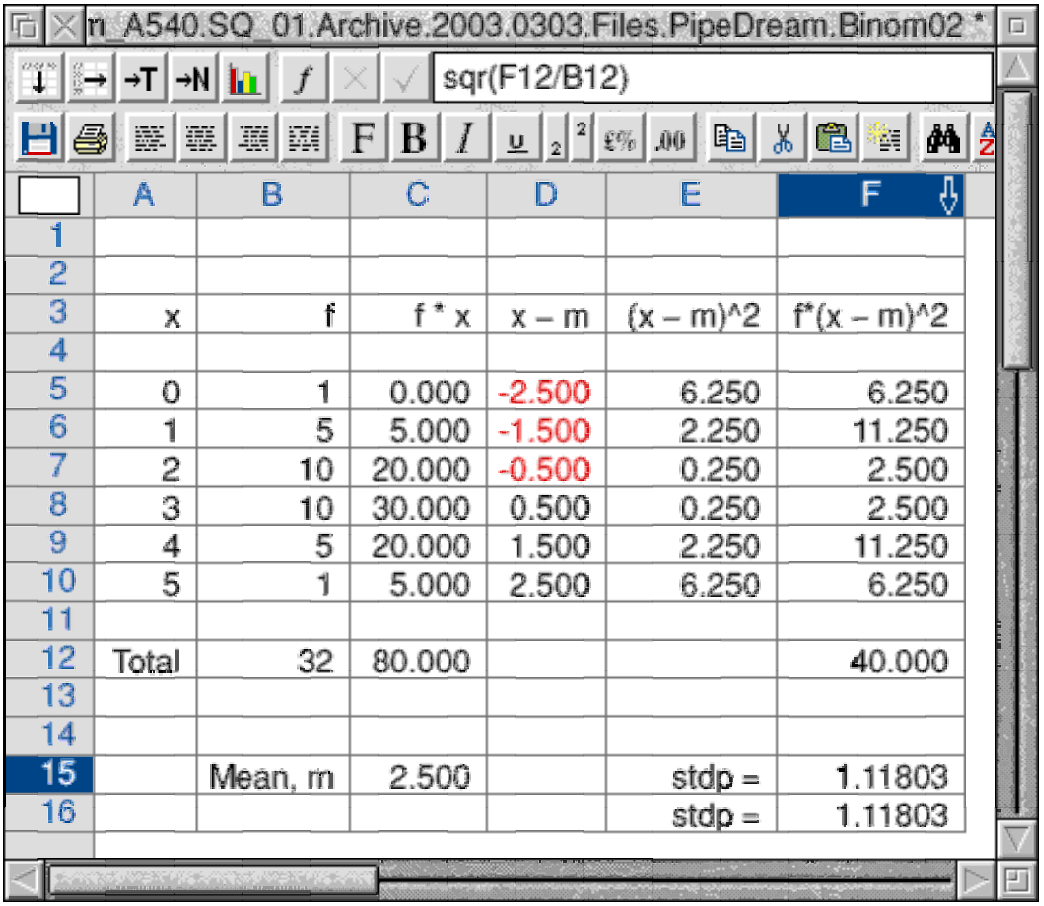
I must refer to the revival of Colin Singleton's Puzzle Corner.

When I looked at the contents page of Archive, I was delighted to see Puzzle Corner. That delight turned to fear when I realised that it was devoted to my "Royal Command".

I asked two questions. These were about a sample made up of 32 repetitions of the five flippin' coin event. The statistic for this event is the the number of heads (per event). The statistic can take values from 0 to 5 inclusive.

My 'simple' question was to find the chance that all 32 events in the sample resulted in five heads. In Colin's Puzzle Corner he gave us the 49 digit number (ending with a 6) which represents the chances of getting 5 heads in every one of the 32 repetitions of the five flippin' coin event—but not how to find it! Of course, Colin does know how to find it.

My question "worthy of a Colin Singleton Puzzle" was to find the chance that the sample of 32 five flippin' coin events would have the 'expected' outcome. This 'expected' outcome is such that the distribution of the statistic in the sample corresponds exactly to the Probability Distribution Function (PDF).



	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

In the screenshot the statistic is *x* (column A) and the relative frequencies are *f* (column B). The population of five flippin' coin events is infinite (which is different from very large) and the PDF (the probability of each of the six distinguishable outcomes) is proportional to the relative frequencies shown in the screenshot.

I had to make a rough estimate of the probability that a 32 event sample would have the outcome shown in the screenshot in order to reassure myself that I could describe this result as "highly unlikely". The value I estimated was about one chance in a thousand. Colin has sent me a letter containing a formula (and a number) representing the chances that 32 observations will have the same 1, 5, 10, 10, 5, 1 pattern as the population. His accurate answer, which is a bit below 1100 to 1, confirms my approximate estimate.

As I write this at the end of February 2003 I have received only Colin's solution to this more difficult puzzle. There is still 15 seconds of fame (not 15 minutes) waiting for someone willing to try! I am pleased that Colin has taken up this 'puzzle'—so please send your solution to him. I am sure that eventually Colin will provide an answer and formula.

Std and Stdp

Let me remind you that the std formula is used with measured data and the stdp formula is used when you know that your data is exactly proportional to the relative frequencies of the whole infinite population.

The slightly larger number is std.
The relationship between std and stdp is: $(std)^2 = (stdp)^2 * n/(n - 1)$.

The custom function [c_Stats]StdGroup(x_array,f_array) uses the std formula (and is used with measured grouped data). The custom function [c_Stats]StdpGroup(x_array,f_array) uses the stdp formula (and is used with grouped data for which the relative frequencies are exactly proportional to those in the infinite population).

Why use std and not stdp?

Now this is a most interesting question!

The very short answer is that the std formula (used with measured data) gives a more accurate estimate of the standard deviation of the (inaccessible) population than the stdp formula does. The n/(n − 1) ratio is called the Yates Correction—I know nothing more about Yates! This correction slightly increases the number which would be arrived at if the stdp formula had been used.

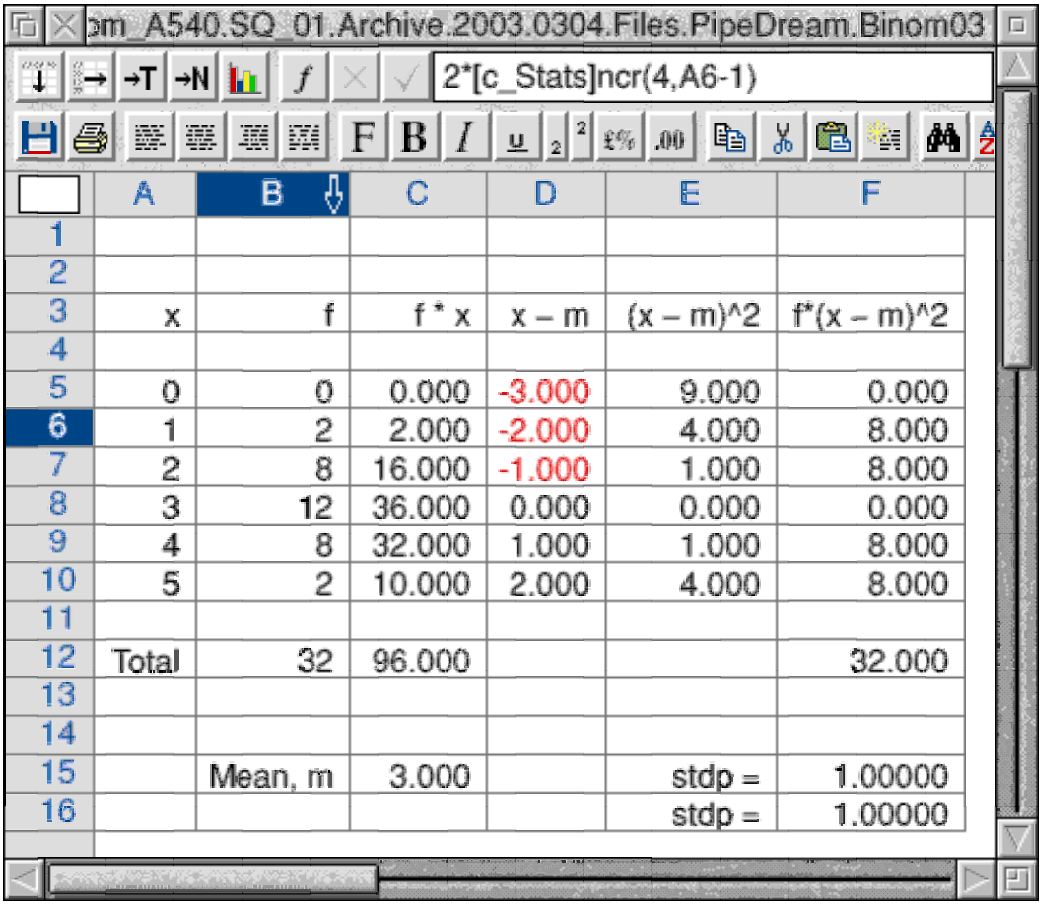
A Modified Event

The longer answer could involve lots of clever mathematics which I won't go into.
What I shall try to do instead is try to give you a 'feel' for why a correction is needed.

Let's consider a modified five flippin' coin event; let's repeat it 32 times to make our 32 event sample (drawn from an infinite population of similar five flippin' coin events).

I shall vary the experiment just slightly by secretly slipping one double headed coin into the set making four unbiased and one double headed coin. What a way to modify this flippin' event? The person recording the experiment doesn't know what I've done.

Now I've let you into this secret we both know that for this modified event the average number of heads will be 3.0 and not 2.5 heads per event. Furthermore we know that the standard deviation of the theoretical population is not half the square root of 5 (which is 1.118) but half the square root of 4 (which is 1). Have a look at the screenshot below and you'll see that the stdp of the modified experiment is indeed 1.



	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	0	0.000	-3.000	9.000	0.000
6	1	2	2.000	-2.000	4.000	8.000
7	2	8	16.000	-1.000	1.000	8.000
8	3	12	36.000	0.000	0.000	0.000
9	4	8	32.000	1.000	1.000	8.000
10	5	2	10.000	2.000	4.000	8.000
11						
12	Total	32	96.000			32.000
13						
14						
15		Mean, m	3.000		stdp =	1.00000
16					stdp =	1.00000

Now here is an interesting question which I shall not answer this month. Instead I shall invite you to answer it. The person doing the experiment thinks this is a 'standard' five flippin' (unbiased) coin event. Do you think that the results of a typical 32 event sample should arouse suspicion or not?

The results of a real experiment will not be those shown in the screenshot. The results will be biased towards the high end (the mean of our new experiment is 3 and not 2.5) and the calculated standard deviation will be a bit low (biased towards 1 rather than 1.118).

The Original Event

Now let us go back to the original sample of 32 five unbiased flippin' coin events.

Our statistic is still the number of heads in each event. We know that the mean of this statistic for the infinite population is 2.5. The mean of a sample such is this is called the ‘Sample Mean’. The sample mean of our (32 event) sample is unlikely to be 2.5.

If we have a sample for which the sample mean is a bit on the high side then, even if all five coins are unbiased, the calculated value of the sample’s standard deviation using the stdp formula will be a bit on the low side. It will be smaller than the population standard deviation because the data is squashed together towards the high end. You will remember that the standard deviation is a measure of the width of a distribution. The squashing together of the data will reduce the sample standard deviation from its population value.

The same applies if the sample mean a bit on the low side. The data will be squashed towards that low end reducing the width of the distribution and reducing the calculated value of the sample standard deviation from its population value.

The important point to grasp here is that the sample standard deviation is more likely than not to be smaller than the population standard deviation because the sample mean is unlikely to be the same as the population mean.

The Yates Correction of $n/(n - 1)$ corrects the calculated value of the standard deviation (making the value larger) so that the result is a better estimate of the population standard deviation. The Yates Correction is not just a number picked out of the air. Using a fairly complicated bit of mathematics it can be proved that this is the ‘best’ correction factor.

Summary

I know that I haven’t given you a rigorous mathematical proof of the Yates Correction but I hope that the ‘squashing’ effect I’ve described will make it seem more reasonable.

Remember to use the std formula when your data is not exactly representative of the population. Reserve the use of stdp for theoretical distributions.

To repeat my main point. The Yates Correction increases the value which would have been calculated using the stdp formula. It is the ‘best’ correction factor. It compensates for the squashing which occurs when the sample mean is different from the population mean.

This Month’s Challenge

If, unknown to you, I have slipped in one double headed coin then would 32 (typical) observations give you a good chance of discovering that something might be ‘fishy’?

Would it be sufficiently ‘fishy’ for you to wish to modify your initial ‘degree of belief’ that all five coins were unbiased?

Allow me to ask my question in a very specific way. Your theory is that the five coins are all unbiased. If your theory is true (whatever ‘true’ means in this case—no coin is completely unbiased!) then the population mean is 2.5. You do a single experiment. Your sample size is 32 events. You can not repeat the experiment. The data you get from your one experiment is all the information you have. It yields a sample mean of 3.0.

Here are my (two) specific questions.
The answer to the second is the key to answering to the first.

Is a sample mean of 3.0 sufficient evidence for rejecting theory that the coins are unbiased?

Assuming your theory (that the coins are unbiased) is true then what is the probability that you can get a sample mean which is 0.5 different from your (assumed) population mean?

Contact

Email is preferred. It has its disadvantages.

One is that I have lost my record of who it was who sent me a most interesting book about Outliers. I know that I had an email exchange about it and about one of the authors who taught my correspondent. However, I have lost the emails and I can't find my note which tells me who sent me the book.