Gerald's Column
by Gerald Fitton

Computers can help with statistical problems because they do sums with precision and speed but they can't tell you what sums to do nor can they tell you what the answers mean.

In this series of articles I shall include a few spreadsheets containing statistical formulae and I shall try to explain to you what the answers mean. In addition and more importantly, I hope to help you understand the meaning of words such as 'probability' and discover how statistics can be applied to problems in which the outcome is uncertain.

Finding the largest prime

But first…

Professor John Greening (of Edinburgh) has sent me three Basic programs (on the Archive monthly disc) which he has used to find the largest prime number which can be held in double precision format by the Floating Point Emulator (FPE). One of his Basic programs will handle 50 digit numbers as a string—we don't need something that powerful. The one which is most interesting is called [!PrimeTes2] and, as written, it displays the nine largest primes below 2^53. It is easy to modify the program to find many more primes.

There is only one catch to this very fast program. It uses a module written by Nick Craig–Wood called [Numbers] and a routine called "Number–Probably–Prime". This latter routine (unlike Colin Singleton's technique) does not guarantee that a number returned by it is prime. The probability (what does that mean?) of the number being prime is 25%. By running the routine 30 times the probability of finding a false prime is reduced to one chance in 4^30 = 2^60. 2^60 is bigger than 2^53 so, if you use this method, it is unlikely that the 'probably prime' numbers which you find (amongst the 2^53 numbers) will be false primes. However, as I said, the method is not watertight nevertheless it is very fast and it has produced an answer which is 'probably' (almost certainly) right.

Statistical problems

Last month I wrote:

"The different ways of 'building up' a PDF for a real or hypothetical Statistic can be divided into two classes. One class of methods is Measurement and the other class is Theoretical."

This month let me draw your attention to three different types of situation which use mainly theoretical (rather than measured) Probability Distribution Functions (PDF).

The first is that which I might call 'a real problem'. An example is the flippin' coin.
The second is that which I called 'a hypothetical'. An example is the missing bus.
The third is the 'mathematical model'. An example is flipping an unbiased coin.

For a real problem we do not know anything about the PDF so we have to guess. For the mathematical model we control the parameters totally. Because of this we are certain that the PDF we create is relevant to our model.

The hypothetical bus problem is somewhere in between. First we shall chose a couple of mathematical models based on the assumption that the buses arrive at ten minute intervals. Eventually we shall reject those models in favour of one in which buses never arrive.

The flippin' coin

Let's return to my anecdote. Please remember that what I am describing really happened. Because this is a real situation (see above) you have to guess what is a relevant PDF. I'm better off than you are. I don't have to guess because I know what happened!

I asked you to consider the result of the sixth throw. If you have chosen as your mathematical model an unbiased coin for which the probability of coming down heads is 50% then getting six heads in a row is a 64:1 shot. It has come down heads five times already and that's a 32:1 shot. What do you think the (real life) experimental results of five heads in a row tells us about the probability of a head on the sixth throw?

Your options are:

(a) The probability is still 50%−well pretty close to 50%
(b) The probability is less than 50%—we can't keep on getting heads
(c) The probability is greater than 50%—tell me why

I have had to start writing this article rather early this month (grandparent duty calls) and last month's article hasn't been published yet so I must refer back to the replies I got last month. A couple of replies argued that I wouldn't be telling this story if it wasn't unusual. Those people forecast quite correctly that the fourth throw would be a head.

I have received more detailed replies along the same lines but with mathematical support and reference to something called the 'Central Limit Theorem'. Essentially these correspondents say that three or even four heads in a row is not so unusual that they would get excited about it. These more mathematical correspondents have decided (somewhat arbitrarily?) that when the odds on the compound (multi throw) event get to the equivalent of a 1 in 20 shot (5% probability) then they would regard that as "statistically significant". By this they mean that the outcome is unusual enough for them to become concerned.

Well, have I got news for you? The sixth throw came up heads! That's a 64:1 shot which, in probability terms, is around 1.6%. If millions of people flipped unbiased coins six times then (roughly) 1 in 64 of those people would find that they'd thrown six heads in a row.

I am trying to get you to understand this elusive probability thing. I assure you it will help you if now you decide (from the (a), (b) and (c) choices above) the probability that the seventh throw will come up head. Once you understand probability then it will seem easy!

Counting heads

Computers come into their own once we have decided on a PDF. Last month, without much explanation, I introduced you to the Binomial PDF. Let's study its construction in more detail. Have a look at the DrawFile below. I have created the table in Fireworkz as the file [Binom01F] and then Saved as DrawFile.

This ‘Save as DrawFile’ option has been introduced in V 1.32. However, V 1.32 is not on general release yet because there are a few serious ‘bugs’ in it so please don’t ask for your copy yet. One major problem with it is that the speed of calculation has slowed down unacceptably. It has been compiled using the 32 bit rather than the 26 bit compiler and we think it’s a problem with the C Library which is causing Fireworkz to slow down.

First	Second	Third	Count
Throw	Throw	Throw	Heads
Head	Head	Head	3
Head	Head	Tail	2
Head	Tail	Head	2
Head	Tail	Tail	1
Tail	Head	Head	2
Tail	Head	Tail	1
Tail	Tail	Head	1
Tail	Tail	Tail	0

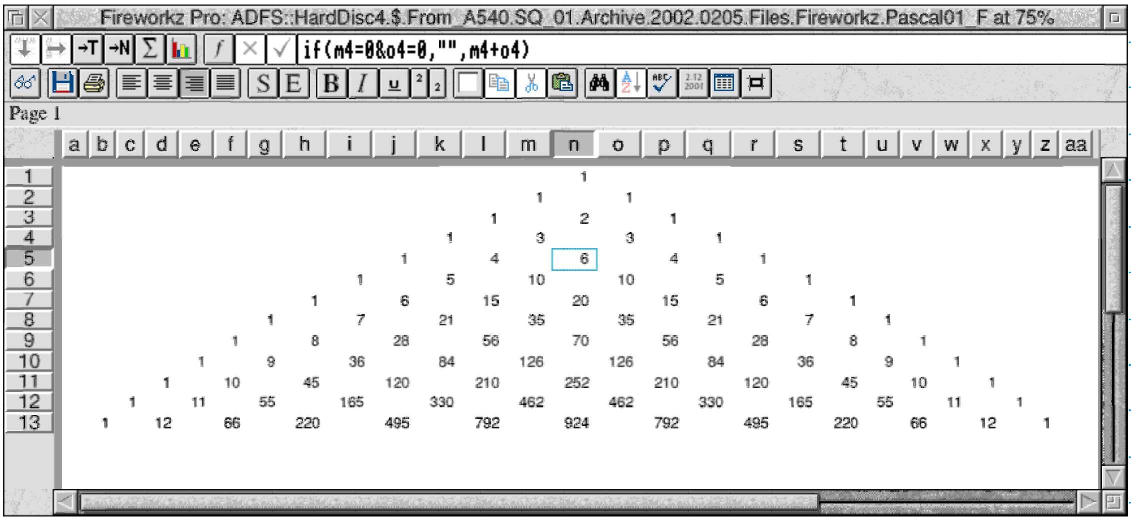
No of	No of
Heads	Times
0	1
1	3
2	3
3	1

The table represents an event consisting of three throws of an unbiased coin. There are eight permutations of the three throws. The first throw can be a head or a tail, four of each. Similarly for the second and third throws the outcome can be a head or a tail with equal frequency. Of the eight permutations, three have the outcome ‘two heads and one tail’ but the order is different in each of the three. You will see these 3 outcomes in rows 2, 3 and 5.

I have used the dcounta(range, condition) function to count the number of heads in each row and then, at the bottom of the table, I’ve used the same function to count the number of rows in which, 0, 1, 2 & 3 heads appear. The frequencies are 1, 3, 3 & 1 respectively.

Pascal’s Triangle

As I mentioned last month, Blaise Pascal (1623–62) discovered a simple technique for constructing a table such as the one above for three throws. Pascal’s method can be used not only for three but any number of throws. His triangle is built row after row from the top down by adding together the two numbers to the left and right and just above the cell for which we want the answer. For example the value entered into cell n5 in the screenshot below is the sum of the cells m4 and o4. I have introduced an if(,,) in order to make the spreadsheet look tidier; without the if(,,) there would be a lot of zeros in the table.



Fireworkz Pro: ADFS::HardDisc4.$.From_A540.SQ_01.Archive.2002.0205.Files.Fireworkz.Pascal01_F at 75%

```
if(m4=0&o4=0,"",m4+o4)
```

Page 1

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa
1														1													
2													1		1												
3												1		2		1											
4											1		3		3		1										
5										1		4		6		4		1									
6									1		5		10		10		5		1								
7								1		6		15		20		15		6		1							
8							1		7		21		35		35		21		7		1						
9						1		8		28		56		70		56		28		8		1					
10					1		9		36		84		126		126		84		36		9		1				
11				1		10		45		120		210		252		210		120		45		10		1			
12			1		11		55		165		330		462		462		330		165		55		11		1		
13		1		12		66		220		495		792		924		792		495		220		66		12		1	

The formula which you see in n5 is replicated throughout the spreadsheet. The last (visible) row of the spreadsheet corresponds to 12 throws of the coin.

Lucky seven?

An event consisting of seven throws of an unbiased coin has a distribution of outcomes represented by the numbers in row 8 of the spreadsheet. The numbers in that row are: 1, 7, 21, 35, 35, 21, 7, 1. There are 2^7 = 128 possible permutations of this seven throw event. Let's look at the number 21 in the table as an example. In 21 times out of 128 permutations the outcome will be 5 heads and 2 tails (but each time in a different order).

The probability of the event having an outcome of seven heads in a row is equivalent to odds of 1:128. Pretty long odds isn't it? Did it happen? What do you think?

The missing bus

This is a hypothetical situation and not a real one. You will remember that the buses were due every ten minutes. I waited half an hour—no bus. I asked you to consider whether the non appearance of the bus during that half hour wait made it more or less likely that a bus would come along during the next ten minutes.

I decided to wait. After an hour a friend with a car pulled up and told me that there were not going to be any buses—they had been diverted.

There is a crucial moment when our estimate of the probability (of a bus arriving) switches suddenly from "a bus is likely to come soon" to "I'm almost certain that a bus will never come". When this 'Catastrophe' occurs we switch to 'Plan B'—but when do we do that?

Choose a PDF

The essential starting point for such an analysis is a Probability Distribution Function (PDF) for (the statistic) the 'waiting time before a bus arrives'. What is a good choice?

Whenever anyone tells you that something is probable or improbable or starts telling you the probability of anything (except one used in a mathematical model), in their mind they will have chosen a PDF which they think applies to the situation. Allow me to be a little dramatic in order to drive home my point. In choosing their PDF they have made a guess! What is more, there is a non zero ‘probability’ that they will have guessed wrongly.

If my last sentence makes you feel a little uneasy then you are following me very well. Bear with me and I shall try to resolve this recursive paradox—but not this month.

Uniform distribution

Let us assume that all buses run exactly to their timetable. By this I mean that buses turn up every ten minutes on the dot. If this is the situation then the only unknown is the difference between our time of arrival at the bus stop and the time the last bus left it.

If this model of precision buses is accurate enough then my waiting time will have a uniformly distributed statistic. Every minute is equally likely to see the arrival of a bus. Each minute has a 10% chance of a Bernoullian Success (the bus arrives).

Every minute which goes by without a bus increases the probability that the next minute will bring me Success. During the second minute the probability of Success has increased to one in nine. During the third minute it is one in eight. If no bus has arrived for eight minutes then there is a 50:50 chance that a bus will arrive during the ninth minute.

The average length of time I have to wait at the bus stop is 5 minutes; the maximum is 10.

The selection of a Uniform distribution for the PDF is a reasonable guess for at least the first eight minutes of waiting and some would argue that a wait of nine or even ten minutes would not be a cause for concern.

Random buses

After ten minutes have passed it is quite apparent that the assumption that the buses run on time must be abandoned. Perhaps we should have abandoned it a couple of minutes earlier. After ten minutes we must guess again at the PDF.

It might be more realistic to give buses a bit of leeway. They never arrive spot on time but they might arrive within a few minutes of the scheduled time (most of the time). I do assure you that it is possible to combine a known (or guessed) PDF for the precision of bus timetables with the random selection of a time of my arrival at the bus stop. We could take that combined PDF and do some probability sums.

I’m not going to do that because the sums are hard and I have a better idea. I shall take a worst case scenario. I shall assume that buses might start out on time but due to all sorts of dispersion effects (such as variable traffic), by the time they get to my bus stop, their arrival time is totally random. Totally random in this hypothetical experiment still means that, on average, a bus turns up every ten minutes. In turn this means that for any and every minute the probability of a bus not turning up remains constant at 90%.

You will remember that in the case of the unbiased coin the probability of a head also remains constant from throw to throw so there is a similarity here. Oh, and by the way, on the seventh throw my sterling silver, pre 1920 half crown came down heads!

Have a look at the table below. In this table *q* is the Bernoullian probability of Failure where Failure means that no bus arrives during a one minute trial. This probability of Failure, *q*, remains constant throughout the multi trial hypothetical experiment.

The table is a DrawFile saved from the Fireworkz spreadsheet [Expo01F]. On the Archive monthly disc this spreadsheet (and all the others) is available in both Fireworkz and PipeDream format. The column headed “Probability of no bus” is a Cumulative Probability Function (CPF). I’ll explain how to read it with an example. The probability of no bus arriving sometime within the first ten minutes is 34.87%. This means that there is approximately a 65% chance that a bus has arrived in the first ten minutes.

q = 0.9000

Minutes waited	Probability of no bus
1	0.9000
2	0.8100
3	0.7290
4	0.6561
5	0.5905
6	0.5314
7	0.4783
8	0.4305
9	0.3874
10	0.3487
15	0.2059
20	0.1216
30	0.0424
40	0.0148
50	0.0052
60	0.0018

The value 0.3487 is 0.9^10. In general the probability of a bus not arriving within the first *x* minutes is *q^x*. The probability of a bus arriving within *x* minutes is *1 − q^x*.

What the answers mean

If our first guess of a Uniform PDF had been reasonable then, after ten minutes, a bus would have arrived. It didn’t. That guess for the PDF must be discarded after ten minutes.

Our second guess for the PDF is based on the worst case scenario of buses arriving randomly with an average delay of ten minutes.

It is reasonable not to discard this new guess after 10 minutes. After 20 minutes there is an 80% probability that a bus should have arrived. I would begin to get a bit edgy. After 30 minutes (half an hour) have gone by we're up to about 96% probability that a bus should have turned up. It still didn't turn up. Something seems to be wrong.

When in doubt—give up

The 96% chance of a Success hasn't happened. Let's look at this 96% probability another way. Instead of concentrating on whether I've been unlucky with the buses perhaps I should consider whether I've been unlucky in my choice of PDF! The 96% probability can be reinterpreted as: "There is only a 4% chance that my current theory is valid".

Perhaps I should give up my 'random distribution of buses' theory now (30 minutes). My friend with the car (had he done the sums?) said that he'd have given up after half an hour!

Finally, after waiting for one hour the chances are less than 2 in a 1000 that my 'random distribution of buses' is a valid model. Why did I wait that long? Perhaps I hadn't done the sums! Perhaps it was because I couldn't think of an alternative theory!

In summary

Let me ask you this: As a result of my hypothetical missing bus, have you subtly changed your mind a little about the meaning of probability? If so then I am well pleased.

To decide on the probability of a bus arriving we guessed a 'buses run on time' PDF for the time we had to wait at the bus stop. Then we waited at the bus stop for more than ten minutes. No bus came and so we rejected this simple PDF.

We developed another PDF based on the worst case assumption that the bus arrival times were randomised with a mean spacing of ten minutes. If our 'randomised bus times' model was a reasonable guess then, after 30 minutes, there is a 96% chance that a bus would have arrived—and no bus arrived. Now to the important bit. Instead of thinking about this 96% probability as the probability that the bus should have arrived (but didn't) you should regard this 96% as the probability that our 'random bus times' model is invalid.

It is this subtle shift in what the probability means which is at the heart of many scientific crucial experiments. Often in the text books I've read, this shift in meaning is very poorly explained. Possibly, or even 'probably', I've done better for you than that.

Finally

Now reconsider the eighth throw of flippin' coin. Remember, it really happened and it happened in front of a class of students. Your choice is still (a), (b) or (c). I shall be most interested in your thoughts. Email or write to me at the address given in Paul's Fact File.