Gerald's Column
by Gerald Fitton

One of the topics of 'conversation' on the Archive_42 Internet List is Statistics. I hope to get you excited about this intensely philosophical subject. Although I shall include a few spreadsheets I must repeat my warning from last month that much of what I have to say is about fundamentals. Computers help because they can do sums with precision and speed. What computers can not do is tell you what the answers mean. If you wish to understand the computer's answers then you might like to indulge me and my digressions.

Floating Point Hardware

But first…

John Barker sent me an email with a short test program which finds 2^53 correctly using BASIC 64 and the Floating Point Hardware he has in his A7000+. If you have the Archive monthly disc you will be able to test John's program for yourself.

An Anecdote

Last month I told you about something which really happened to me whilst I was teaching at the local College. As part of my lesson I took out a coin and flipped it three times. Each time it came down Heads. As I pointed out to my students, no coin is perfectly unbiased so we wouldn't expect exactly a 50:50 split between Heads and Tails.

Last month I left you wondering about the fourth throw. I asked "What is your best guess at the probability of the fourth throw returning the dreaded 'Head'?" I gave you a choice:

(a) The probability is still 50% (the coin doesn't know what happened on previous throws)
(b) The probability is less than 50% (it's about time our luck changed)
(c) The probability is greater than 50% (it's got stuck in a rut).

I received a couple of replies which argued as follows: Gerald would not be telling this story unless something surprising happened—therefore the result of the fourth throw was a Head. You are right. It was a Head.

You deduced it was a Head by using additional information, not information about the coin but information about why I was telling this story. To make use of such information is totally legitimate. Play your 'hunches'. They are right more often than 'chance' predicts.

So let's go on from here with a fifth throw. Remember this really happened. I asked the students to discuss what probability to ascribe to the chance of it coming down a Head for the fifth time in a row.

I'll tell you what happened later. Right now I want you to ponder on the three possibilities which I have listed as (a), (b) and (c) above and decide what you think about the next throw (and the next and the next…). In order to help you decide here is another little story. This story is something which did not happen—but it could happen to me (or you) one day. It is a 'Hypothetical'.

Buses

I am waiting at a bus stop. I know that the buses are due every ten minutes. The time table says so. Even if I have just missed a bus (I don't know if I've just missed a bus or not) the maximum wait should be ten minutes.

I wait ten minutes and there is no bus. Never mind, one will be along soon. Another ten minutes goes by. Still no bus. Yet another ten minutes goes by. Still no bus.

Now here's a question I want you to consider. Does the non appearance of the bus during my half hour wait make it more or less likely that a bus will come along during the next ten minutes? Will three buses all arrive at the same time? When I have put this to students they nearly all agree that a bus is more likely now that I've waited half an hour.

Please think about my question; compare the situation with the flipped coin experiment.

Some Definitions

In the course of our lives things happen—or they don't. In statistical jargon these 'happenings' are called Events. An Event can be something like a 'Happy Event' (the birth of a baby) or something of even greater import such as me flipping a coin in front of my students. What you have to realise about Events is that they are not something you measure—they are things that happen.

There are all sorts of things we can measure about an Event. In the case of the 'Happy Event' we can measure something which is a number such as the baby's weight. We could measure something which is not a number such as whether the baby is male or not. In the case of the flippin' coin we could measure whether it comes down Heads or Tails; or we could measure the number of rotations in the air, whether it was a left or right hand flip, the wind speed at the time or even whether person flipping the coin has blue eyes!

The thing which is measured is called the Statistic. A Statistic can be a number (such as 'the baby's weight is 8 lb') or it can be a label (such as 'Male'). Statistics which are not numbers are often called Attributes—we might measure (record) the Attribute of coming down Heads in our flippin' coin Event.

When an Event takes place we can make a measurement of the Statistic. Such a (single) measurement is called an Observation.

Very rarely are we interested in one single measurement—the Statistical Analysis of (what are called) Unique Events (such as The Creation of the Universe or the meltdown of a nuclear reactor) is a most interesting subject which I might be persuaded to discuss one day (it was part of the course which I taught) but for now, let's consider only Repeated Events.

One difficulty about Repeated Events which is not as trivial as it might first appear, is how to define the set of Events that we are interested in. The set of all the Events that we think are relevant is called the Population (of Events). An example of a Population of Events is the birth of all babies. Do we include only babies born in the UK? Should we include still born (dead at birth) babies?

What about the flippin' coin? Is the relevant Population all coins flipped by anybody anywhere at any time or is the Population the one session of coin flipping by me during that particular (unique in many respects) lesson?

Choosing a relevant Population raises an interesting philosophical point. Every Event is Unique. What we are looking for is the thing which is similar about all the Events which we deem to be relevant. The fact that I am telling the coin story has influenced your estimate of the probability of a Head. It has done so because people do tend to tell unusual stories rather than ones in which the Outcome is 'usual'. So what is a 'relevant' Population for our flipped coin? Certainly not the same Population as that which consists of throws of an unbiased coin!

For my next definition let's assume that we can't 'get at' the whole Population of Events. Indeed, it is usually the case that the Events which interest us most haven't happened yet. For example, in the case of the 'missing bus' the most interesting Event is the bus which hasn't arrived. As another example, what's going to happen to the next flippin' coin?

Since we can 'get at' only some of our Population of Events we have to make do with investigating those which are available. The set of Events which we select for further investigation is called the Sample (of Events). We hope that our Sample is representative of the Population of relevant Events. If it isn't then our deductions will be invalid.

The measurements of the Statistic for this subset of the Population is called the Data.

The Data rarely, if ever, contains the Observation which is of most interest to us. Usually what we are trying to do is use our Data to predict the Outcome of a future Event.

Before proceeding further please refer to the table below, read the summary and check that you understand the meanings of these definitions.

	One	*Many*	*All*
Happening	*Event*	*Sample*	*Population*
Measurement	*Observation*	*Data*	*Statistic*

To summarise these definitions: 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).

Probability

I have asked you what you think is the probability that the next throw of the coin will be a Head but I haven't told you what 'Probability' is. That's because it is quite difficult to define and even harder to understand. In my experience, rather like 'Infinity' and 'Eternity' people use the word 'Probability' with only a vague understanding of its meaning. I shall do my best to help you realise some of its properties and hence some of the things you can (and can't) do with it.

By the way, continuing with my anecdote, I flipped the flippin' coin for a fifth time and, would you believe it (?), yes you would (!); it came down Heads for the fifth time in a row. You believed it was going to be a Head because (like others) I tell only interesting stories.

Let me ask you again. What is the best estimate we can make of the probability that the coin will come down Heads on the sixth throw? Is it (a) about 50%, is it (b) less than 50% or is it (c) more than 50%?

What do I mean by 'Probability' in a case like this? What extra (hidden) information is relevant to making an estimate of this probability. Perhaps the story of the missing bus is relevant. Perhaps it isn't. Read again the section in which I introduced the (missing) bus.

Probability Distribution Function (PDF)

Computers are very useful tools for studying statistical problems. Usually the computer comes into its own once we have chosen a PDF to play with. Later in this article I'll use a spreadsheet to generate a PDF which is relevant to the flippin' coin problem (even if it not too much use when it comes to the hypothetical but nonetheless missing bus).

So what is a PDF?

There are many definitions and most of them are so circumlocutory that it would be counter productive for me to repeat them here! With apologies to all those who know a lot about PDFs here is a nearly accurate description.

Imagine that we have a Population of Events (such as flipping a coin). Suppose that we decide that our Statistic (the thing we measure) can take a limited number of values (in the case of our coin it can take two values, 'Head' and 'Not Head' where 'Not Head' includes such rare events as the coin disintegrating before we can measure whether it is a 'Head' or 'Not Head'). The (sought after) PDF is the relative frequency distribution of the different values which the Statistic can take (in our case the relative frequency of 'Head' and 'Not Head' is about 50:50). We can regard this relative frequency distribution (the PDF) as a step on the way towards understanding that rather more elusive entity, 'Probability'.

We might choose as a Statistic for our Population of 'Happy Events' the weight at birth (to the nearest ounce) of all babies born in the whole of the UK during the year 2000. The Statistic can take many values and we might even plot a graph of frequency against birth weight for our measurements. This graph would be a graphical representation of the PDF.

Finding a PDF

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 is Theoretical.

We could (and as a nation we do) spend a lot of time and money measuring the birth weights of babies and then describing the PDF of the birth weights we have found. I don't want to pursue this Measurement method right now. Let's leave it for another occasion.

Using Symmetry

Instead let's have a look at the other method. Usually what I have called the 'Theoretical' method of creating a PDF relies on symmetry and it is this feature which I shall use now.

Trials and Boolean Statistics

A Trial is a simple Event, the Outcome of which can often be represented by a Boolean Statistic. A Boolean statistic can take one of only two values, usually called Success and Failure, but sometimes called True or False. In the case of our flippin' coin we might regard Head as Success and anything else as a Failure.

I shall risk upsetting some of my readers by suggesting that in the case of a 'Happy Event' we might regard a male child as a Success and anything else as a Failure. Of course, my defence is that 'Success' and 'Failure' are only labels and should not be taken to have any 'value judgement' associated with them!

If the Trial consists of throwing a dice, then a Statistic chosen to represent the outcome could be the number shown on the upper face of the dice. That Statistic can take six different values and hence is not a Boolean Statistic. However, the Success or Failure of throwing a number greater than a four has only two possible Outcomes (Success or Failure) and is thus a Boolean Statistic. The distribution of such a Statistic is Bernoullian. It is named after Jacques Bernoulli. The notation for the probability of success is Pr(B=Success) = p and that for failure is Pr(B=Failure) = q. For the Bernoullian distribution the sum of p and q, the two complementary probabilities, is p + q = 1.

The Binomial PDF, Bin(n,p)

This PDF was described by Jacques Bernoulli in his treatise, *Ars Conjectandi* (the art of conjecture), published in 1713 (eight years after his death). To be fair I must mention that Blaise Pascal had produced numerical tables of this distribution in the 1650s. On the Archive monthly disc I have included a spreadsheet produced using Pascal's method.

It is the PDF of a Statistic representing the Outcome of an Event which is compounded from a set of 'n' Bernoullian Trials. By this I mean that the (compound) Event is not one isolated Trial but the combined Outcome of many, indeed 'n', Trials.

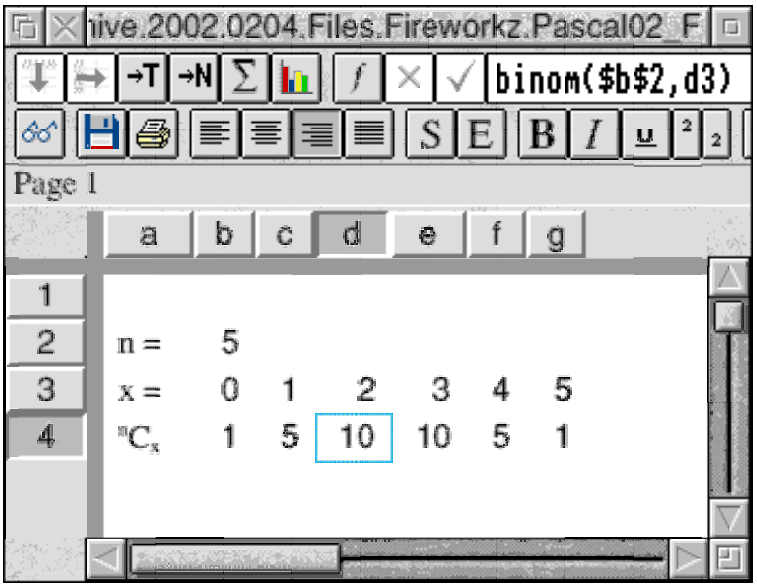
Let me refer back to my anecdote of the coin which, up to now, I have flipped five times.

You must stop thinking of these five separate Trials as five distinct Events but combine the whole lot into one (compound) Event. The Event is "Flip a coin five times". The Population of Events is the set of all such (five flippin') Events. n = 5.

The statistic, X, used in the Binomial PDF is obtained by counting the number of Successes which occur in the set of n Trials.

X, can take values between 0 and n inclusive. In the case of my five Trial (coin flipping) Event, the Statistic, X (the number of Heads) can take values between zero and five. There are six possible Outcomes for any individual (but compound) Event. We might have zero Heads out of five flips or 1, 2, 3, 4 or, as I did, 5 Heads out of five!

Before I frighten you off with formulae I shall invite you to look at the screenshot below. It is a Fireworkz spreadsheet which calculates the relative frequency with which we will get between zero and five Heads if we flip an unbiased coin five times.



Read the table this way. The number of Trials is 5; n = 5. The sum of the numbers in row 4 of the spreadsheet is 32. The (infinite?) Population of Events (each Event consists of five flips of the coin) can be divided up into groups of 32. Each group of 32 Events is identical and can be regarded as representative of the Population PDF.

The relative frequencies can be deduced using symmetry. The number 1 in cell b4 means that just 1 of the 32 Events has zero Heads. The number 10 in d4 means that in 10 of the 32 Events we have 2 Heads (and 3 Not Heads). In g4 you will see a 1. Getting five Heads in a row (as in my anecdote) happens once in every 32 (multi Trial) Events.

Now for the 'jargon'. The probability that the Statistic, X, takes the value x (between 0 and n) is written as Pr(X = x). The formula for the PDF is: $Pr(X=x) = {}^{n}C_{x}\, p^{x}\, q^{(n-x)}$. The coefficients ${}^{n}C_{x}$ which are generated as x varies between 0 and n, are the coefficients used in the expansion of $(p + q)^{n}$. It is from this binomial expansion that the statistical distribution takes its name.

In the formula line of the spreadsheet in the screenshot you'll see the function I've used to find the values of ${}^{n}C_{x}$. It is binom(n,x) where n is the number of Trials making up the Event and x is the number of Successes in the series of n Trials.

Flippin' Coin

Now let's apply this Theoretical PDF to the events which took place in my statistics lesson.

No coin is completely unbiased so no coin has an equal chance of coming down Heads or Tails. However, most coins are unbiased to an extent that we can apply this Theoretical PDF without risk of making a serious error.

The second requirement for applying this PDF is that when a Trial has been completed (let's say the first throw of the coin) then the Outcome of that Trial can not affect the Outcome of subsequent Trials.

Contrast this with a pack of red and black cards. If you pull out five black cards in a row and do not replace them, then the ratio of black to red cards has changed from 50:50 to a ratio which makes pulling out a red card next time more likely.

In the coin experiment do you think that the 'odds' on getting a Head is influenced by what happened on previous Trials? If so then we can not apply this PDF. If you think that coming up Heads five times in a row does reduce (or increase) the chance of a Head on the sixth throw then you can not apply this Binomial PDF.

If we make the assumption that the pre 1920 sterling silver half crown I was using is not seriously biased and that the Trials are (what is called) 'independent' then, for a five Trial Event there is 1 chance in 32 that I would get five Heads in a row. Put it another way: at the stage of five Heads in a row I have pulled off a 32:1 shot.

Let me ask you to reconsider yet again the forthcoming sixth throw. Using the same assumptions, six Heads in a row is a 64:1 shot which is double that of a 32:1 shot. Do you believe that the chance of me throwing a Head is 50%, less than 50% or more than 50%?

Buses

We've waited half an hour and still no bus. Does the fact that no bus has arrived make it more or less likely that a bus will arrive in the next ten minutes? Before you read the next bit please think about the poor elderly gentleman (me?) waiting at the bus stop. I've got to get into town because my library book is due to be returned. If I don't return it I'll get fined. I don't want to get fined. The Library will shut in another hour. Do I wait for the (ever more likely?) bus or revert to 'Plan B'? What would you advise?

I decide to wait. I wait a further ten minutes. I consult the timetable again. It is Saturday. The buses are due every ten minutes. I haven't made a mistake—or have I?

I wait and wait. An hour has gone by and no bus. I have half an hour left before the library shuts. Has 'no bus in the last hour' made it more or less probable that there will be a bus turning up shortly? I have to make a decision, a decision based on my estimate of the probability of a bus turning up. As an aside—what do I mean by 'Probability' in this case?

What does this 'hypothetical' tell you (if anything) about my forthcoming 64:1 shot?
The flippin' coin was not a 'hypothetical' but something which really happened.

Your choices are:
(a) The probability is still 50% (the coin doesn't know what happened on previous throws)
(b) The probability is less than 50% (it's about time our luck changed)
(c) The probability is greater than 50% (it's got stuck in a rut).

Before you decide on your answer let me complete the 'hypothetical' story of the bus. A friend with a car pulled up and asked me if I was waiting for the bus. "Yes!" "Did you know they have diverted the buses today and so they're not stopping here?" "No!" "How long have you been waiting?" "An hour." "You should have guessed something was up after half an hour. Why did you wait so long?". He gave me a lift.

That bus story is completely hypothetical. My anecdote about the flippin' coin is real.

Let’s try to analyse our thoughts about the probability of the bus arriving ‘soon’.

Whilst the missing bus is delayed by only a small amount of time this delay makes it more probable that a bus will arrive ‘soon’. After a longer amount of time has passed it becomes more probable that the bus will never come.

We are entering an interesting field of Mathematics called ‘Catastrophe Theory’ in which changes occur abruptly and dramatically. There is a crucial moment when our estimate of the probability (of a bus arriving) switches suddenly from “a bus is more and more likely to come soon” to “I’m almost certain that a bus will never come”. How do we decide when to revert to ‘Plan B’?

Now reconsider the sixth throw of flippin’ coin. Remember, it really happened and it happened in front of a class of students. Your choice of probability is (a), (b) or (c). I shall be most interested in your thoughts.

Email or write to me the address given in Paul’s Fact File.