# *Gerald's Column*
## *by Gerald Fitton*

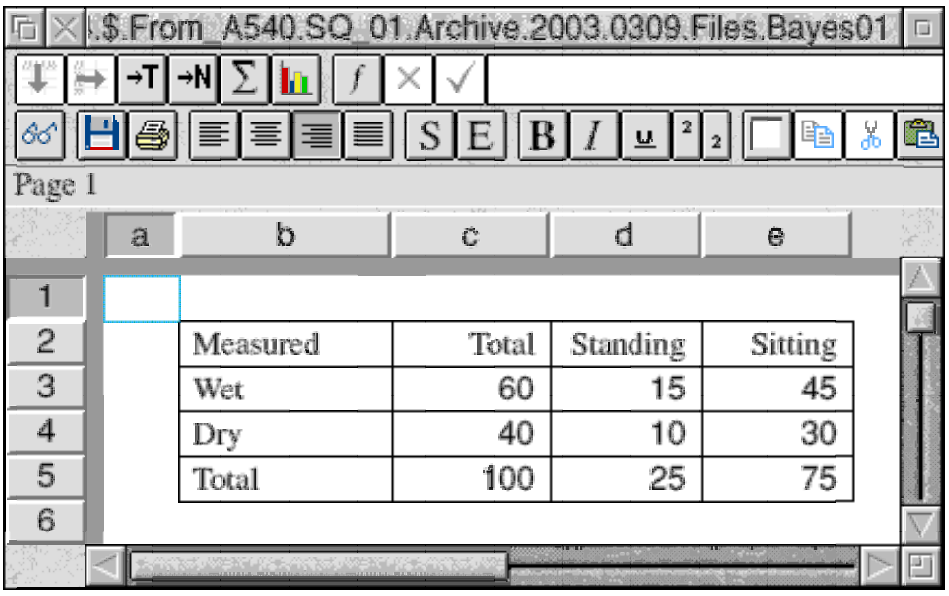
In the March 2002 issue of Archive I introduced you to the flippin' coin. Ever since then I have received at least one email (or letter) each month asking me to explain the almost mystical Bayes' Theorem and the incomprehensible formula which follows from it.

The formula is: $P(H_1|E_1) = P(H_1) * P(E_1|H_1) / (P(H_1) * P(E_1|H_1) + P(H_2) * P(E_1|H_2))$

### Standing in the Rain

My grandmother used to take me for walks. We had a set of 'standard walks'. It is probably a exaggeration (a slight one) to say that I remember with great clarity every blade of grass in every field on each of our walks—but it seems so. Most of our walks took us through fields in which there were cows. I was taught to leave gates in the condition I found them. Shut the gates that were shut and leave open the gates that were open. The only days that I can recall with clarity were sunny; but it must have rained now and again.

How do I know it must have rained on some of our walks? Apart from remembering that a certain bridge was a shelter from the rain and the borrowed super large black umbrella (my grandfather's), I do remember most vividly my grandmother's weather lore.




"If the cows are standing then it will rain; if they are sitting down then it will remain fine"

### The Raw Data

Before I get too much correspondence from those of you with meteorological knowledge I must tell you that the table is totally fiction. I made up the numbers. They are not real.

The weather is measured on 100 days. Look first at column c. Of these 100 days 60 are classified as "Wet" and 40 as "Dry". Now look at row 5. On 25 of these 100 days the cows were observed to be standing and on 75 days they were sitting.

There were 15 days when the cows were standing and it was Wet. There were 45 days when the cows were sitting and it was Wet. So far as Dry days go they stood on 10 days and sat on 30 days.

## Percentages

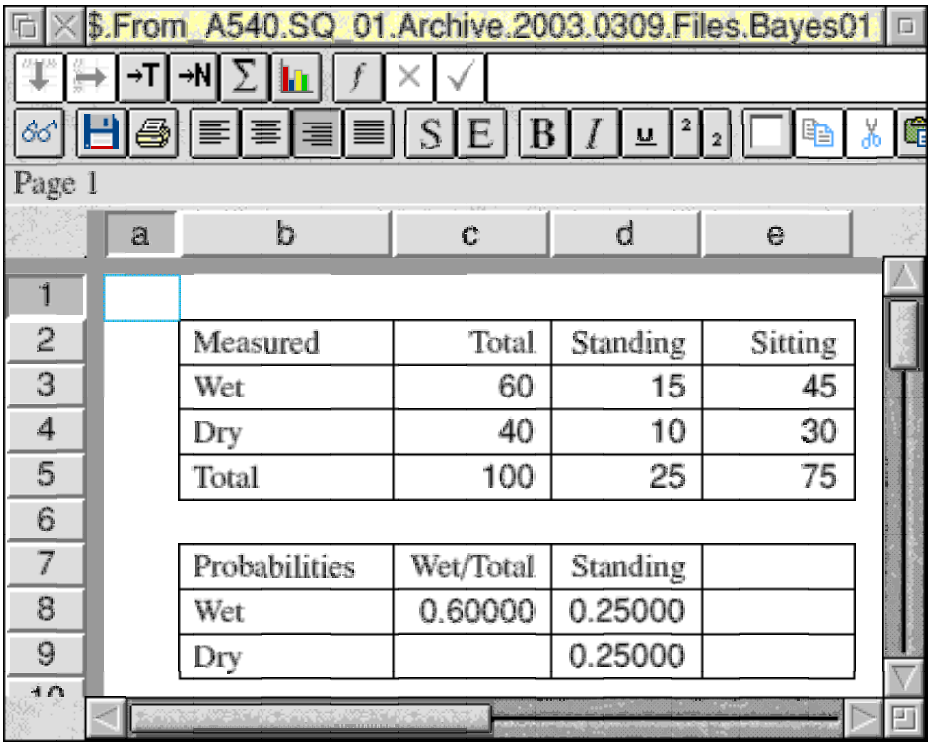
In the past I have told you that "probability" is not well defined. Indeed it means different things to different people at different times. In the context of converting parts of this table to probabilities I would ask you to take the simple view that a statement about observations such as "On 60 days out of 100 it is Wet" can be converted to "There is a 60% probability that any individual day will be Wet." Similarly, on 25% of days the cows will be standing.

You might find it a bit more difficult to express the contents of cell d3 as a percentage. The reason for the difficulty is not the arithmetic. The difficulty is nothing to do with deciding what number to use for the numerator, indeed it is obviously d3 = 15. The difficulty is knowing what to divide by. Relevant choices are c3 or d5 (and, in some circumstances, even c5). When do we choose c3 and when do we choose d5?

## Probabilities

Look at the next table. Probabilities are usually expressed as a number less than one rather than as a percentage. A probability of 60% can be written as 0.6.

In cell c8 you will see the probability that any day (regardless of whether the cows are standing or sitting) is a Wet day. The numerator for this calculation is in c3. The appropriate denominator is the total number of days (regardless of whether the cows are standing or sitting); this is the value in c5. The content of c8 is (simply) c3/c5.



The formula in d8 is d3/c3. This value, 0.25 (or 25% if you prefer), can be interpreted as: "Knowing that the day is Wet then, on what proportion of days will the cows be sitting?"

Similarly the formula in d9 is d4/c4. In spite of the smaller numbers, it turns out there is exactly the same probability, 0.25, that cows will stand on a day classified as a Dry day.

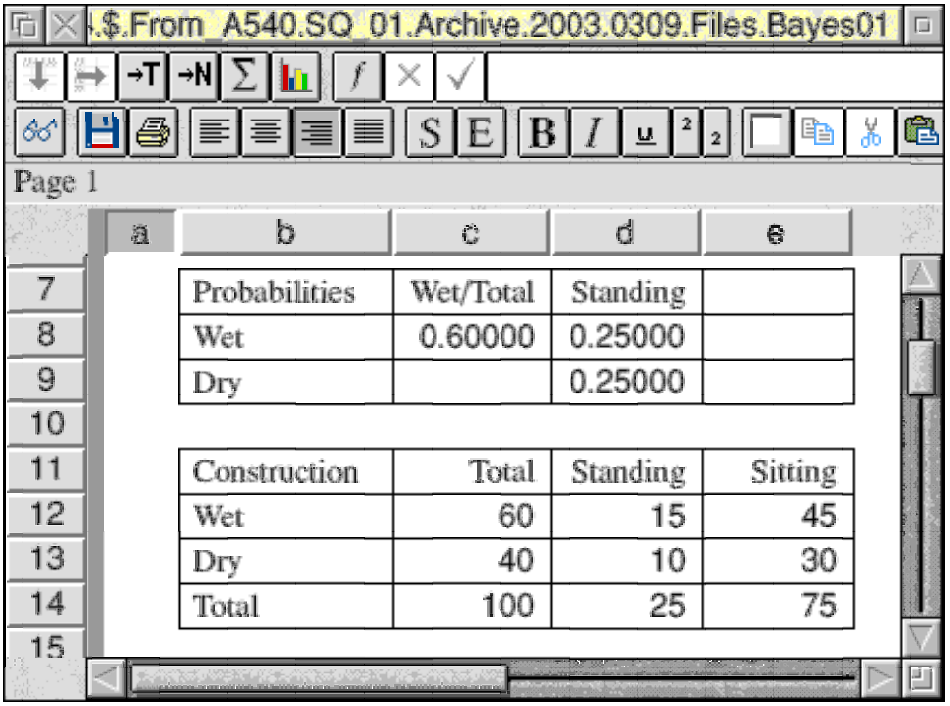
What you must note very carefully is this. We have started by considering whether the day is Wet or Dry and then looked at the proportion of those days on which the cows stand.

What I would like you to do is to look at the table in rows 2 to 5 and give consideration not to the Dryness of the day but, instead consider whether the cows are Standing. Select only those days on which the cows are Standing and then answer this question. On what proportion of those "Cows are Standing" days is the day a "Wet" day?

Here's a bit of help. The cows are standing on 25 of the 100 days. What proportion of these 25 days is it a Wet day? The answer is that returned by the Bayes' Theorem formula!

## Probability Table

In the screenshot below you will see that I have hidden the table which is in rows 2 to 5. In many Bayes' Theorem Problems this Raw Data table is not provided.




In a typical Bayes' Theorem problem the list of probabilities below is all you get.

| | |
|---|---|
| Probability of a Wet day | 0.60 |
| Probability of cows standing on a Wet day | 0.25 |
| Probability of cows standing on a Dry day | 0.25 |

The problem requiring Bayes Theorem is: "If you know that the cows are standing then what is the probability of the day being a Wet day?" This is exactly the task which I asked you to consider at the end of the previous section headed "Probabilities".

In order to ‘solve’ this problem your first need to reconstruct the table of rows 2 to 5 using only the information provided in the list above. I have carried out this reconstruction in rows 11 to 14 of the spreadsheet. The method is as follows:

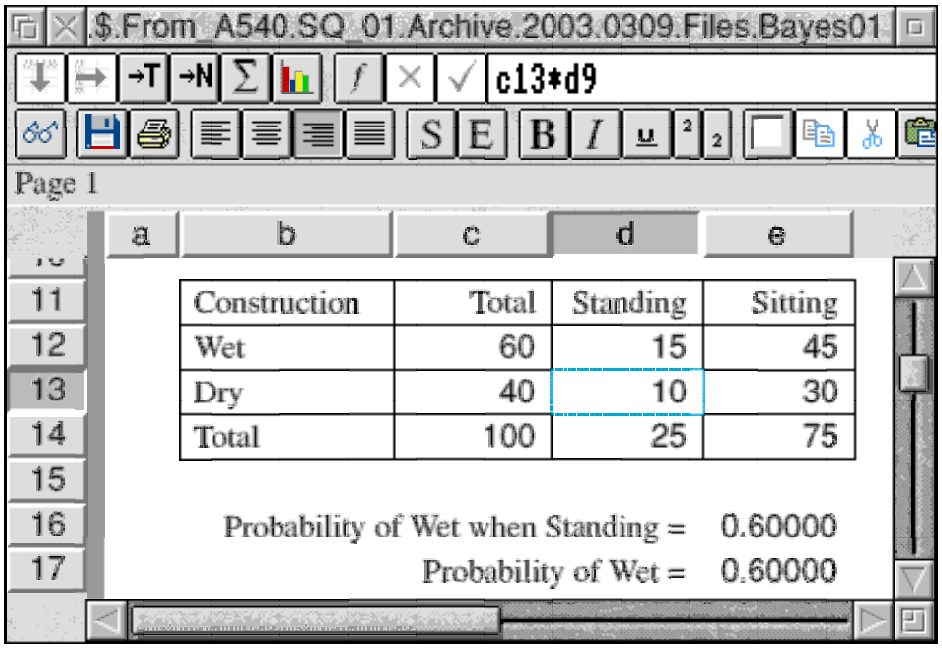
Enter a number in the grand total box, c14. This can be anything because the final answer, a probability, is a ratio. I have entered 100. The next value to calculate is the Wet total in c12. This is c8*c14. The Dry total in c13 is (c14 − c12). The value in d12 is (c12*d8) and the value in d13 is (c13*d9).

I am sure that you can decide what formulae are needed in d14, e12, e13 and e14.

## The Solution

The Bayes’ Theorem formula provides us with the answer to the question:
“If you know that the cows are standing then what is the probability that the day is Wet?”

In the next screenshot I have scrolled the spreadsheet down further and you will see that in cell e16 the value returned is 0.6 or, if you prefer, 60%. The probability that it is Wet on a day when the cows are standing is given by the formula d12/d14; its value is 0.6.




## What does the answer mean?

I must remind you that I have made up this set of figures. It does not represent the results of any real experiment. Indeed I have ‘doctored’ the values in order to get a specific result which is totally untrue. It must be untrue because grandmothers are always right! In e17 you will find another probability which is also 0.6. This ratio is the probability that the day is Wet irrespective of whether the cows are standing or sitting; it is the value from e8.

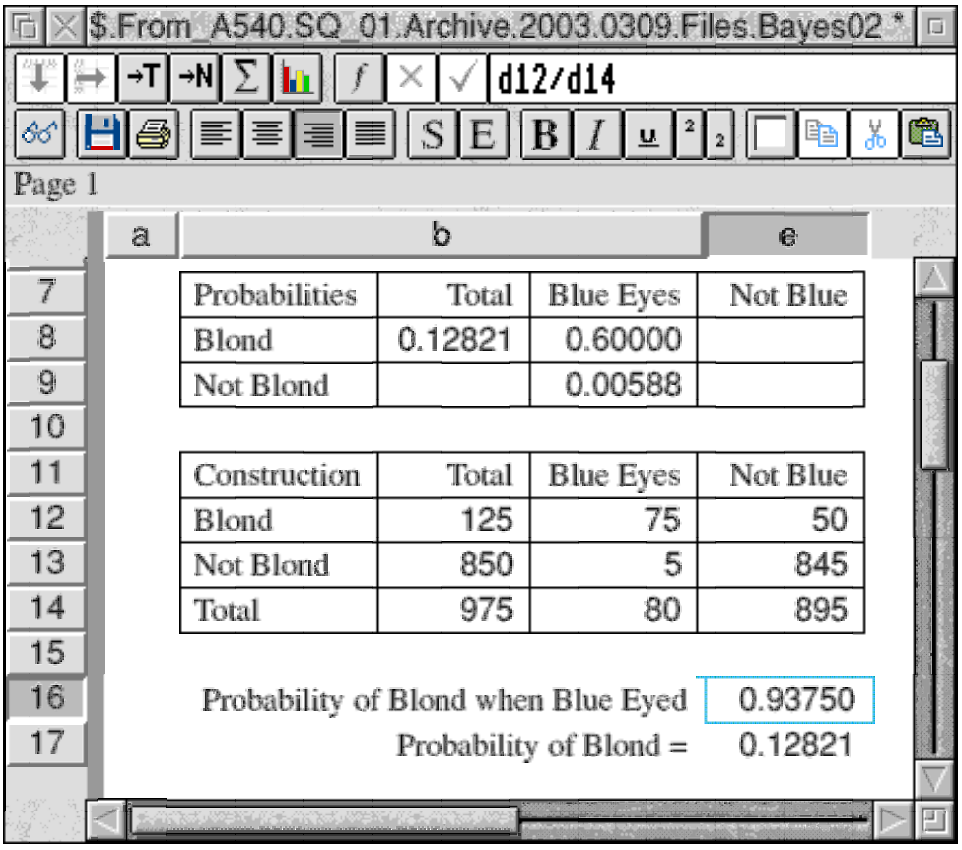
The conclusion from this exercise is this: Knowing whether the cows are standing or sitting has no effect on our estimated probability that the day will be Wet. Our belief that the day might be Wet rather than Dry is not changed by our knowledge about the cows.

## Blue Eyed Blonds

Have a look at the next spreadsheet. Once again I must warn you that I have made up the numbers in order to make my point. The data is not the results of a legitimate survey.

This time I shall state the question in typical Bayes' Theorem format.

In a survey of 975 people 12.821% were classified as Blond. Of these Blonds 60% were found to have Blue Eyes. The percentage of Blue Eyed people in the survey who were not Blond was found to be much smaller at 0.588%. Calculate the percentage of those with Blue Eyes who are Blond. Interpret your result stating whether being Blue Eyed is a valuable indicator when assessing if the person is likely to be Blond and comment upon whether being Blond is a valuable indicator for the likelihood that they have Blue Eyes.




You will see from the screenshot that my spreadsheet has done all the Bayes' Theorem calculations for you. The 'given' probabilities include the percentage of Blonds who have Blue Eyes (60%). Bayes' Theorem is a method of 'working backwards' through the data to find the percentage of Blue Eyed people who are Blond (93.75%).

Suppose for some reason we wished to select a Blond from a group and all we have is a list of names. If we simply choose one at random then the probability of our selection being Blond is 0.12821, about one in eight. Indeed, selecting at random this way we are more likely to select a Not Blond person (over 87%) than a (desirable) Blond.

However, if we have not only the list of names but also each person's eye colour then we should make our random selection from those who have Blue Eyes. The probability of us fulfilling our desire for a Blond is raised from 0.12821 to 0.93750, which is pretty good.

Having Blue Eyes is a very good indicator that this Blue Eyed person will be Blond.

As a by-the-way, if the person we desired to select had to have Blue Eyes and the only information available to us was their hair colour, then, by selecting a Blond, we would raise the probability of achieving a desirable result from just over 8% to 60%. Trying to select your Blue Eyes from the Blonds will give you a better success rate; but you will still fail 40% of the time. If it is important not to fail then 60% might not be good enough.

## Bayesian Epistemology

What would my articles be without a 'philosophical bit'.

Epistemology is a fancy name for knowledge. Well, not quite. Epistemology is not about the knowledge itself but more to do with devising methods for discovering (hidden) Truths.

Thomas Bayes (1702 − 61) was an ordained non conformist minister (one of the first six) who, like many others in the religious professions, was a Mathematician. In 1731 he became a Presbyterian minister in Tunbridge Wells.

He is principally remembered not for his good works with the Church but for a paper which was published posthumously in 1763. Somewhat archaically it goes by the title "Essay towards solving a problem in the doctrine of chances". Bayesian Epistemology is named after him because of the philosophical content of that paper.

I shall take the liberty giving a simplified (possibly oversimplified) description of Bayesian Epistemology. It is this: When we acquire information we can use it to change (what, in Philosophical jargon, is called) our "Degree of Belief". Let's use Blonds as our example.

We have a list of 959 names (such as those of our example) and we pick a name at random. Our "Belief" that the person we have chosen is a Blond is fairly low. Indeed this "Belief" can be quantified as just under 13%. There is an 87% probability that they are Not Blond.

If we are given additional information, not about whether they are Blond, but about their eye colour then we can use this information to increase this "Degree of Belief" (or, if you prefer, the probability) that the selected person is Blond.

If we are given the information that our selected individual has Blue Eyes then our expectation that this named person is Blond rises from about 13% to nearly 94%. If they definitely don't have Blue Eyes then our expectations fall from about 13% to about 6%.

What you must realise is this. Once the individual is selected then the (elusive) Truth about their hair colour is fixed. The colour of their hair is not changed no matter how much information we gather. What does change is our expectation of finding that this person is Blond. Until we actually see them (and, in some circumstances perhaps we never will) there will always be some lingering doubt about the colour of their hair.

All Scientific advances are the result of applying Bayesian Epistemology. Our "Degree of Belief" in a hypothesis changes (it increases or falls) as we gain information by doing experiments. Crucial Experiments are those which have a table with discriminating properties similar to that of the Blue Eyed Blonds. Cow type experiments are useless.

## SpamStamp

This statistical program looks at every word in every email message and calculates the probability that the message is Spam. It does so by using knowledge gained from earlier emails and being told whether those earlier emails are Spam or Not Spam. The program continues to 'learn' every time a message is manually marked as Spam or Not Spam.

Every word in every message is added to the database. The number of Spam and Not Spam messages is not counted. The database consists of words and the number of times each word has appeared in a Spam or Not Spam email.

After this initial 'learning' (or 'teaching' if you prefer) the database consists of lists of words with associated probabilities. These are equivalent to the probabilities which appear in column d of my two examples.

The two probabilities associated with every 'key' word in the database can be described as: "Knowing that the message is a Spam (or Not Spam) then this is the probability that the word appears". This is a different probability from that which can be described as: "Knowing that this word is in the message then what is the probability that the message is Spam?" What we want to know is the answer to the latter question; for each word our database contains a pair of probabilities corresponding to the former and not the latter.

In use, every email message is put through a Bayesian Inverter. By "Inverter" I mean that we have to turn the probabilities from the database around. The probabilities in the database relate to the frequency with which words appear in known Spams whereas what we want to know is the 'inverse' of this, namely the probability that the message is a Spam. The method used is based on the Bayes' Theorem formula and the computer processing required is an extension of that which I have demonstrated in my two examples.

Of course, until you actually read the message you won't know whether it is Spam or not. All that the Bayes' Theorem formula can do for you is increase the probability that those which you mark as 'Probably Spam' are actually Spam.

The Risk Analysis question for you to answer is whether you are going to open all your emails anyway because you can't afford to miss a vital email or whether you are prepared and willing to sacrifice a few genuine emails in order to avoid the delay associated with opening and then rejecting a quantity of Spam. If you are going to open all of them anyway then you don't need a Spam Filter. If you are willing to sacrifice a few genuine emails then a Spam Filter is a useful utility.

## Next Month – Bayes' Formula

Bayes' Formula is: $P(H_1|E_1) = P(H_1) * P(E_1|H_1) / (P(H_1) * P(E_1|H_1) + P(H_2) * P(E_1|H_2))$

## Communication

Please contact me by email (preferred) or by letter if you have any questions or comments.