Gerald's Column
by Gerald Fitton

In last month's article I started my explanation of the way in which SpamStamp works. This month I shall develop that theme and, as a byproduct, do something which I think Jan–Jaap has not done. In this article I convert the whole method of assessment to one which uses logarithms. I shall start with an introduction to logarithmic scales for graphs.

Where were we?

You will recall that the probability of an email being Spam can be found using the formula:

$$P = \frac{1}{(1 + r)}$$

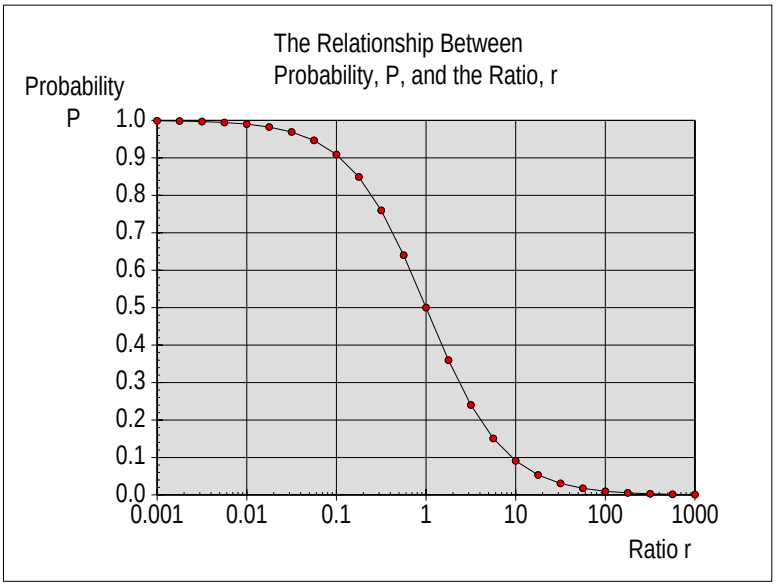
Simplified Bayes' Equation

In this formula 'P' is the probability that the email is Spam. 'r' is the product of many r_n: $r = r_1 * r_2 * r_3 * r_4 * r_5$ ... where each of the values of r_n can be found from the formula:

$$r_n = \frac{P(E_n|H_2)}{P(E_n|H_1)}$$

Calculating r_n

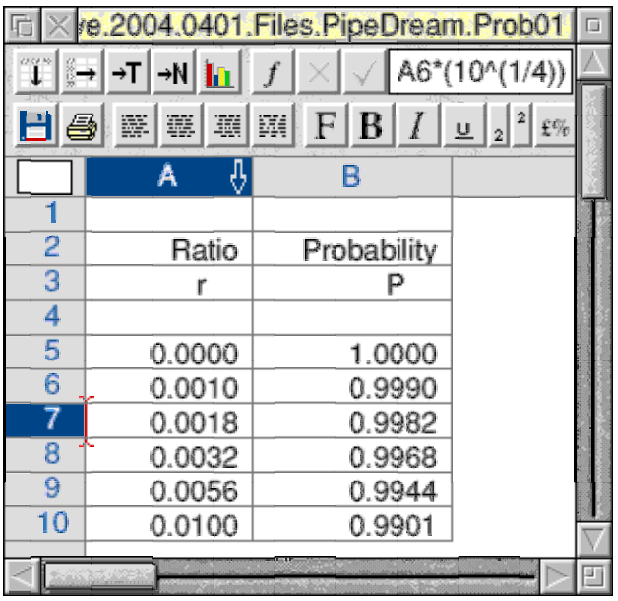
The graph of P against r is shown below:



Although values of r between 0.01 and 100 may be regarded as 'inconclusive', it will be a rare event for a real email to give a value of r within this 'range of uncertainty'. For nearly all emails P will be either close to one (Spam) or close to zero (Ham).

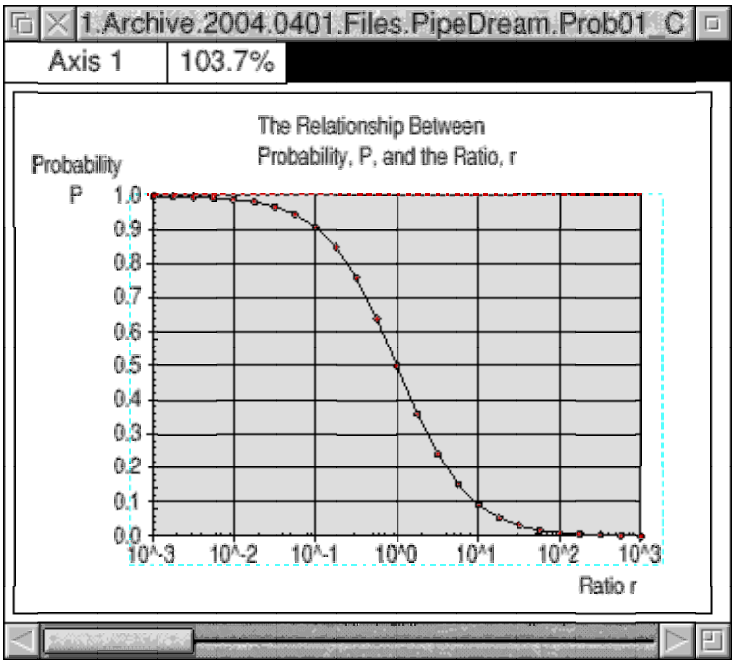
Logarithmic Scales

In passing I mentioned that the scale used for r in this graph is a logarithmic scale. I am not surprised that (once again) I have been asked how to create such a logarithmic scale. Have a look at the values in the spreadsheet [Prob01]. I have said that the data in column A (values of r) lends itself to the use of a logarithmic scale. Why do you think this is so?



	A	B
1		
2	Ratio	Probability
3	r	P
4		
5	0.0000	1.0000
6	0.0010	0.9990
7	0.0018	0.9982
8	0.0032	0.9968
9	0.0056	0.9944
10	0.0100	0.9901

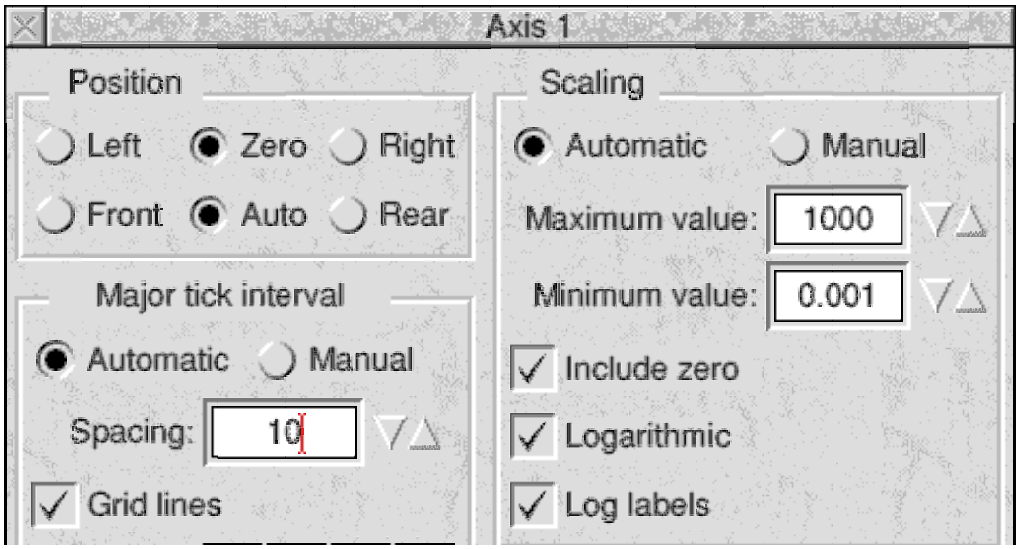
The formula in A6 is 10^12^(−1/4) or, if you prefer, 0.001. The formula in A7 is the value in A6 multiplied by the fourth root of 10. This formula is replicated down column A so that the value in A10 is 0.01 and that in A14 is 0.1. Using the fourth root causes every fourth cell to be ten times larger. A logarithmic scale is appropriate for graphs in which the values used on the axis increases exponentially (using a factor) rather than linearly.



Look at the labels along the x axis in the screenshot of the PipeDream chart [Prob01_C] (the earlier graph is a drawfile and not a screenshot). You will see that the numbers on the axis used for r (Axis 1) range from 10^{-3} to 10^{3} rather than from 0.001 to 1000. The use of powers rather than plain numbers such as 0.001 and 1000 is called 'using logarithmic (or log) labels'. This type of label can be achieved easily using PipeDream. Note that "Axis 1" appears in the top left box of this screenshot. We shall return to this in a moment.

I receive correspondence from people who can't get the Axis 1 (x axis) pop up menu on screen. When they attempt to do this they (inadvertently) go through a sequence which causes the Axis 2 (y axis) menu to pop up instead.

Look again at the screenshot of [Prob01_C]. You will see a blue dotted box around the x axis and all the points. Furthermore "Axis 1" appears in the box. To select Axis 1 you must click on the numbers which are along the x axis; then the blue box and will appear.



Execute the sequence <menu> – Selection – Axis and the Axis 1 dialogue box (shown in the screenshot) will appear on screen. You will see that I have ticked the box labelled "Logarithmic" and the box marked "Log labels". It is this latter box which changes the way in which the x axis is labelled from its default of 1000 to the 'Log label' 10^3.

Logarithms

You may remember that I said that the data in column A of [Prob01] lent itself to the use of a logarithmic scale.

One important reason for drawing this P against r graph with a logarithmic scale (with or without logarithmic labels) is because it makes the shape of the graph beautifully symmetrical. It is symmetrical because an r value of, say, 1000, exactly balances an r value of 1/1000. Indeed, you will remember my example (from last month) of the words "Viagra" and "Jan–Jaap" exactly balancing each other out. The symmetry of this graph with its logarithmic scale encourages you to see such relationships more easily.

I want to introduce a mathematical device which capitalises on this symmetry and the logarithmic nature of the relationship between r, and the Probability, P.

I shall introduce a new variable (I think not used by Jan–Jaap in his analysis), s_n, which is related to r_n by the equation $s_n = \log(r_n)$.

To make this mathematical device work I do not have to choose any particular base for the logarithms, it could be 10 or e or anything. I shall use base 10 logarithms for my article because I think it has advantages (for me if nobody else) in this particular case. For some applications the base e is more suitable and I have seen base 2 logs used to good effect!

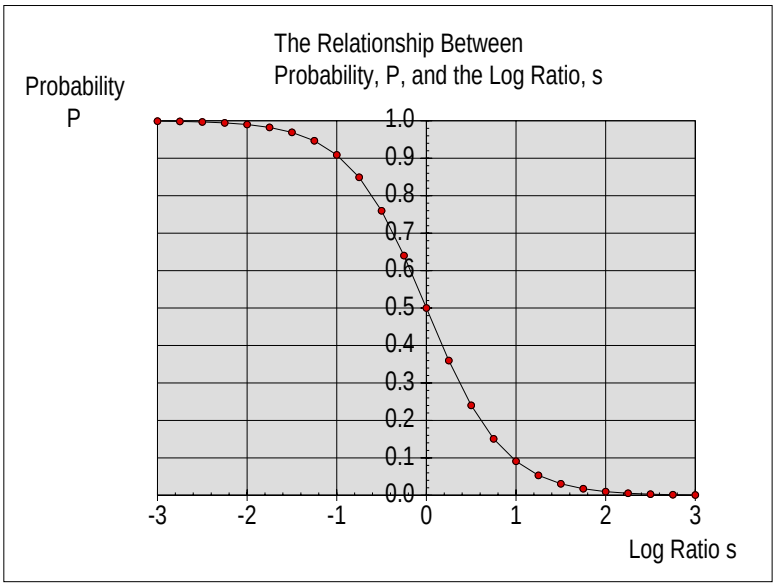
The inverse of $s_n = \log(r_n)$ is $r_n = 10\text{^}s_n$. We can interchange s and r as we please.

For those of you who might have forgotten how to use log tables as an aid to multiplication let's have an example. The identity we shall use is: $\log(r_n * r_m) \equiv \log(r_n) + \log(r_m)$.

When $r_n = 0.001$, $s_n = -3$. When $r_m = 1000$, $s_m = +3$.
These values of r are multiplied together to give: $r = r_n * r_m = 1$ and hence $s = \log(1) = 0$.
We could have added together the values of s to give: $s_n + s_m = 0$ and hence $r = 10\text{^}0 = 1$.

Instead of multiplying together values of r_n we can add together values of s_n. Generally, using s_n instead of r_n will give us a better 'feel' for the effect of a single word on the probability of the email being Spam. It will do so because (for most people at least) using the linear scale of s rather than the ratios of r is more intuitive.

Below is yet another graph which I created using PipeDream. It may look like the previous graphs but look again at the variable I have used for "Axis 1".



Significant Values of s

The range of probabilities which we may regard as 'uncertain' are those between 99% and 1%. When P > 99% we will regard the email as Spam and when it is below 1% we shall regard the email as Ham. This range, 99% > P > 1%, corresponds to a range of r which is $0.01 < r < +100$. In turn this corresponds to a range for s of $-2 < s < +2$.

I shall rephrase this:
If s is more negative than −2 then the email is labelled Spam.
If s is more positive than +2 then the email is labelled Ham.

Procedure

We can calculate a value of s_n for every word in the email using the formula:

$$s = log\left[\frac{P(E|H_2)}{P(E|H_1)} \right]$$

These values of s_n are added together to make $s = s_1 + s_2 + s_3 + s_4 + s_5 + ...$
If $s < -2$ then the email is Spam. If $s > +2$ then the email is Ham.
The values of s which lead to an uncertainty are $-2 < s < +2$.
It will be unusual for s to fall within this ‘uncertain’ range.

Last Month’s Examples

Let’s take a few of the special examples from last month and restate them in terms of this new (logarithmic) variable, s.

Missing Words – 1

Last month, in this section, I explained that SpamStamp allocates an r value of 0.001 to words (such as “Viagra”) which appear only in the Spam database. To use a bit of mathematical jargon, this value of r, r = 0.001, ‘maps’ to s = log(0.001) = −3. Similarly words (such as “Jan–Jaap”) which appear only in the Ham database have a value of r = 1000 which ‘maps’ to s = +3. Personally I find that adding together these two values of s, to give s = −3 + 3 = 0 much easier to understand than I do multiplying together the two corresponding values of the ratio, r.

The Bias Factor

In Jan–Jaap’s article he introduces a “bias factor” in his equation.
What this does is to change the equation $P_n = 1/(1 + r_n)$ to $P_n = 1/(1 + 2 * r_n)$.

As I explained last month, if the full bias factor is applied (to all 15 most significant words) then the effect is to increase the final value of r by 2^15. This factor of 2^15 is equivalent to (it ‘maps’ to) a value of s = log(2^15) = 4.515 (to three decimal places).

Since a single ‘spam word’ (such as “Viagra”) gives an s value of –3 you will see that Jan–Jaap’s bias factor will overpower one such word—but not two! Incidentally, a bias factor of 40% (rather than the full 100%) will give an s value of s = log(1.4^15) = 2.192.
This is not enough to overpower even one “Viagra” word.

Missing Words – 2

Words in an email which do not appear in either database are allocated an 'inverse Bayesian probability' of $P_n = 40\%$. If we convert this to a value of r_n using $r_n = (1/P_n) - 1$ then this works out to be $r_n = 1.5$. In turn this gives a value of $s_n = 0.176$.

If none of the words in an email appear in either database then the value of r for the email will be 1.5^15 which 'maps to s = log(1.5^15) = 2.641. One "Viagra" word with s = –3 together with 14 words which appear in neither database will sum to s = −3 + log(1.5^14) or s = −0.5347. This 'uncertain' value of s 'maps' to P = 1/(1 + 10^s) = 0.7740, a 77% probability that the email is Spam.

Nonsense Words

As I said last month, I have noticed that 'nonsense' words (such as "dlenciuqn", consisting of random characters) appear regularly in the subject line and body text of Spam emails. Often there are many such nonsense words. Generally these nonsense words will be ignored by SpamStamp because SpamStamp uses only the most significant fifteen words.

However, in the longer term, if such an email is classified as Spam, then these nonsense words get added to the Spam database diluting the significance of real Spam words such as "Viagra". Jan–Jaap has allowed for this by gradually allowing them to 'fade' from the memory of the spam and ham databases and be replaced with more frequently used words.

Summary

OK? Let me summarise where we have got to.

We might consider that there are four classes of words.

Words such as "Viagra", 'spam words' found (almost) exclusively in the spam database.
Words such as "Jan–Jaap", 'ham words' found (almost) exclusively in the ham database.
Words such as "dlenciuqn", 'nonsense words' found (usually) in the spam database.
Words such as "the" and "and", 'spam neutral words' found in both databases.

In assessing an email the words which are significant (is my 's' is for significance?) are those with large positive or negative values of s. The program SpamStamp uses only the fifteen most significant (large positive or negative) values of s. Almost exclusively these fifteen words will be found in the first two classes (above). Those in the fourth class ('spam neutral') will have very little (if any) effect on the classification of the email.

The value of r_0

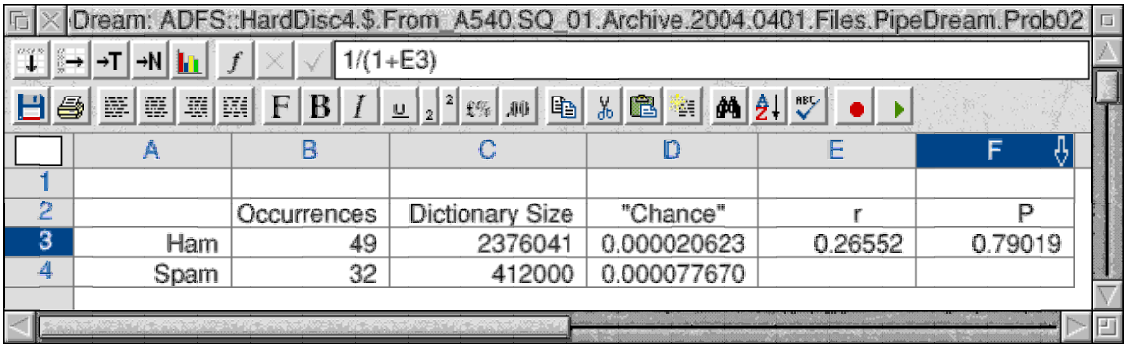
I have left this 'difficult to estimate' value until last.
What value shall we give r_0? What value shall we give s_0?
Perhaps you'd like to have a guess right now before I tell you what SpamStamp does.

Let me remind you that r_0 (and hence s_0) is totally different from all the other values of r_n. Indeed, all the other values of r_n are given by the formula:

$$r_n = \frac{P(E_n|H_2)}{P(E_n|H_1)}$$

Calculating r_n

In this formula for r_n the numerical values are determined objectively by counting the number of words in the spam and ham databases as I indicated last month. The spreadsheet which you can use to find these values of r (and P) is the one called [Prob02].



	A	B	C	D	E	F
1						
2		Occurrences	Dictionary Size	"Chance"	r	P
3	Ham	49	2376041	0.000020623	0.26552	0.79019
4	Spam	32	412000	0.000077670		

The formula for r_0 is different in kind and in its nature. It is:

$$r_0 = \frac{P(H_2)}{P(H_1)}$$

The values on the right hand side of the equation, $P(H_2)$ and $P(H_1)$, are not objectively determined numbers at all. They are the 'a priori' probabilities that the email is Ham or Spam respectively. They are chosen using either guesswork or prejudice.

The Extra Chance

Which reminds me ...

I have received only one response to my question "From whence the Extra Chance?"

Have a look back at the December 2003 edition of Archive and at the section related to the difference between Hypothesis Testing and the Bayesian approach to the advancement of knowledge (about the world of events in which we live).

If you are encouraged by the lack of response—then please give me some encouragement!

The Unknown Probability

I think I shall leave unto another occasion all the 'lyrical waxing' which I might indulge in when trying to describe methods of dealing with an 'unknown probability'. Since I am running out of space this month I had better leave it and 'get to the point' right away!

In Jan–Jaap's program SpamStamp, he chooses $r_0 = 1$ (which 'maps' to $s_0 = 0$). Effectively the program totally ignores all these difficult 'a priori' probabilities!

There is a great deal of justification for choosing $r_0 = 1$. For example, we could invoke the Principle of Equal Likelihood which (in essence) states that, since we have no idea whether the email is a Spam or Ham we might as well assume they are equally likely ($r_0 = 1$).

Of course it is not completely true that we have no idea what this month's 'spam rate' will be. I could try to argue that last month's 'spam rate' is no guide to this month's 'spam rate' but if I did that then I would find myself with a mail box filled with counter arguments!

I don't want to go down that road—well, not until I discuss "The Unknown Probability" more fully. Instead I shall blandly say that the most telling argument for regarding r_0 as being irrelevant is the original one devised by the most perceptive Reverend Bayes.

SpamStamp's r_0

The value of r_0 used in SpamStamp is 1. The corresponding value of s_0 is $s_0 = 0$.

You will see that if r_0 is given the value 1 then the formula r can be rewritten without any reference to r_0 as $r = r_1 * r_2 * r_3 * r_4 * r_5 *$ etc!

What an ignominious way for our 'a priori' probabilities to 'vanish' into obscurity!

Bayesian Epistemology

Obscuring our original uncertainty is inherent in the nature of Bayesian Epistemology.

In essence Bayes states that if we collect enough evidence then our starting point (the time when we were totally ignorant, when we had no knowledge at all) fades into obscure insignificance. Before we have collected evidence (data) we have to make a guess at the 'truth'. When we have collected sufficient evidence then, whatever our starting point (believer or disbeliever), after enough evidence, finally, we shall all believe the same thing!

Bayesian Epistemology leads us from ignorance and guesswork (or prejudice) to more secure knowledge and a more certain (ie supported by 'evidence') belief. An email which initially could be either Spam or Ham becomes (in our minds) either one or the other—and we treat it as such with few (if any) reservations.

As I have written before, there is no such thing as Certainty. Indeed, we must always be willing to change our minds as more evidence becomes available. This 'message' (of tolerance and flexibility) is quite apparent in the writings of the Reverend Bayes.

If we do not continue to collect evidence for our beliefs we shall eventually become slaves to prejudice. ‘Faith’, whether in the classification of emails using SpamStamp, in the potential warmth of tomorrow’s sunshine or in those personal relationships which are so very precious to us, this ‘Faith’ needs constant renewal if it not to become a form of destructive ‘Prejudice’.

I call such blind faith “destructive” because it prevents us from objectively seeking out and harvesting new information. Of course the new information may destroy our deeply held beliefs—but let’s be more positive about this search for information. Actively ignoring the possibility that objective new information will be useful to our quest for knowledge will automatically bar us from using it to further strengthen our cherished beliefs against wanton and even random attacks which, by their very nature, can not be anticipated.

In practical terms, from time to time you must reboot your Spam and Ham databases.
If you don’t do this then, eventually, SpamStamp will tell you ‘lies’!

Communication

Please contact me by email (preferred) at <Archive@abacusline.demon.co.uk> or by letter if you have any questions or comment. They are always most welcome. Particularly welcome will be comments about Hypothesis Testing versus Bayesian Epistemology and what I have called “The Extra Chance”!