The background is a repeating pattern of the equation:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

AN INTRODUCTION TO
BAYESIAN FOR MARKETERS

Tom Lloyd
Founder and Director
MetaMetrics Ltd
September 2013

<MetaMetrics>

integrated marketing analytics

CONTENTS

ABOUT METAMETRICS

We are a new breed of agency. We were established in response to the blurring of lines between advertising, direct and digital marketing. As modern marketing becomes increasingly integrated we have broken down the silos that typically existed in analytics to provide insights across market, channel and customer.

We have brought together deep experience in brand marketing, econometric modelling, customer analytics, direct / digital marketing planning and technical innovation. This enables us to provide a fresh approach to solving our clients' challenges through analytics.

We are pioneering the use of Bayesian to help marketers deal with uncertainty and incorporate existing experience and learnings to create more accurate models.

Independent, impartial and innovative, we work with many of the country's best known brands.

AN INTRODUCTION TO BAYESIAN

If you search for Bayes' Theorem in Wikipedia the first thing you see is a warning "This article may be too technical for most readers to understand". Not the most encouraging start in the world.   And it's easy to see why… we are usually taken straight into the mathematical formulae. We will instead try to explain the significance of the theorem and its modern day application and leave the equations until later.

SO LET'S START WITH AN EXAMPLE

For example: "Suppose that we have two bags each containing black and white balls. One bag contains three times as many white balls as black. The other bag contains three times as many black balls as white. Suppose we choose one of these bags at random. For this bag we select five balls at random, replacing each ball after it has been selected. The result is that we find four white balls and one black. What is the probability that we were using the bag with mainly white balls?"[1]

Bayesian analysis at its simplest is about conditional probabilities – the probability that A will happen, given that we know B. In other words the joining of two pieces of information. If I had not taken out any balls I would have to assume the probability of having chosen the bag of mainly white balls is 0.5 (50%). However I have some extra information (I have taken out a sample of balls from one of the bags). Bayes' Rule is about joining these two pieces of information mathematically.

THOMAS BAYES' THEOREM

Bayesian refers to methods in logic and statistics named after the English mathematician and clergyman Thomas Bayes (c. 1702–61), in particular methods related to probability inference: using the knowledge of prior events to predict future events.

In the early eighteenth century, many problems concerning the probability of certain events, given specified conditions, were solved. Bayes instead turned his attention to the converse of such a problem an presented a solution to a problem of "inverse probability in "An Essay towards solving a Problem in the Doctrine of Chances". Sadly for Reverend Bayes this essay didn't see the light of day until it was read to the Royal Society after his death.

Bayesian probability is the name given to several related interpretations of probability, which have in common the notion of probability as something like a partial belief, rather than a frequency.

[1] Source: http://www.eecs.qmul.ac.uk/~norman/BBNs/Bayes_Rule_Example.htm

THE MATHEMATICS OF COMMON SENSE

What is interesting is that Bayes' Rule reflects how the human brain processes information. You don't need to be a mathematician to know from the balls example above that if you had achieved this result you would intuitively say "I am pretty certain I have got here the bag of mainly white balls. I know it COULD be the bag of mainly black balls but, it's not very likely." No maths here then, just common sense. But exactly how certain can you be… mathematically? Bayes rule allows you to calculate this probability. I won't go into why or how as there are lots of good examples openly available in the internet.

BAYES THEOREM IN PLAIN ENGLISH

Bayes uses certain terms that you may also hear such as: "Prior", "Likelihood" and "Posterior". These are important but simple concepts, so they are worth explaining:

LIKELIHOOD

The 'Likelihood' is simply the data that we have - in this case the likelihood I would choose the bag of mainly white balls if I knew nothing else (which is 0.5). If I had drawn no balls out that would be my answer.

PRIOR

This is the crux of Bayesian theory and is essentially why it differs from "normal" statistical methods. It represents the additional knowledge that we have. In this case it comes from taking a sample of balls from one of the bags – but the Prior could come from anywhere. It could be a hunch, data taken from another study, anything we feel could have a bearing on the calculation. Maybe we hold the bags up to the light to try and see if we can peek through the fabric. Maybe the person who filled the bags tells us "I think it's that one, but I'm not totally sure".

POSTERIOR

This is then our result, our mathematical calculation of how likely is it that I am holding the bag of mainly white balls given previous selections and the strength of belief in my prior.

So… a non mathematical way of writing this as a "formula" is:

PRIOR X **LIKELIHOOD** = **POSTERIOR**

| Strength of belief that I have the mainly white bag | The fact that there is a 50:50 chance of choosing the mainly white bag randomly | How likely it is that I have the mainly white bag |

In the extreme, if I had sampled 90% of the balls in the bag then I'd have a very strong prior and therefore I would be pretty confident I had the right one. If I had not sampled any balls and only had the evidence of the person who had filled the bags and he said…"actually, you know what, now you ask I'm not so sure…I thought I put them that way round, but Mary was over here messing about with the bags when I went to get some tea, she might have moved them". In this case I am not going to have a very strong Prior, so my posterior is going to be much more driven by the likelihood – the one thing I do know which is that one bag is mainly black and one white (50:50 chance).

Or think of it the other way. What if there were 100 bags of balls and only one was mainly white. My friend might say "I think it's probably that one", but I would have to think about the fact that there are still 99 mainly black bags and only one mainly white one, so even with that information I'd be lucky to pick it. In this case I have a strong prior, but this will be dwarfed by the strength of the very strong Likelihood – so my chance of picking the right bag is lowered by this.

PREDICTING THE FOOTBALL LEAGUE CHAMPIONS WITH BAYESIAN

I will give you an example that I think explains the principle in a way that is a little more relevant.

Let's consider the English Football Premier League table. Let's imagine that we are six weeks in and you are trying to predict the winners of the full 38 game season.

What would you do?

A   Base your predictions purely on the data from the first six weeks? But imagine last year's winners, Manchester United, had a poor start and are in 9th place and that newly promoted Hull were in fact topping the table having got off to a flying start. Traditional statistics would either say – well we have six matches' worth of real relevant data so we either use that as our sole basis for prediction and say that the newly promoted Hull will win the league

OR….

B   We must disregard that data from the six matches totally as too small a sample size and just use last year's final table as our best estimate. But since last season, Manchester United's hugely successful manager Alex Ferguson had retired, and Chelsea saw the return of Jose Mourinho… and what about the newly promoted teams such as Hull? – we have no past results for them at all in this division.

The truth is that there **is** value in that data from the six matches, as it gives us a barometer of current form, but I'm not completely happy to use it as my sole basis for making my prediction. In reality I would most likely blend the two pieces of information, giving a certain amount of weight to each. I might say that Manchester United will recover but probably won't quite win this year; the newly promoted team is likely to stumble (no promoted team has ever finished above 7th in the league) and that Chelsea look like very strong contenders for the title given the return of the Special One.

Last year's table and my knowledge about events in the world of buying and selling players is my **Prior**, the data from the six weeks is my **Likelihood** (my "real" data), and these combine to give my prediction - the **Posterior**.

Currently I'm giving a reasonably equal balance of weight to my Prior and the Likelihood. However, after 30 games of the season, I am naturally going to give less weight to my Prior, and almost all the weight to the Likelihood (the current league standings).

The interesting thing to reiterate once again about Bayesian analysis is that it reflects how the human brain makes judgements. It's just putting this into a mathematical format.

SO… HOW DOES THIS APPLY TO BUILDING MARKETING MODELS?

WHY BAYESIAN IS SO IMPORTANT FOR MARKETING

We have seen in our Introduction to Bayesian that the Bayesian approach blends Prior knowledge with the data we have (Likelihood) to produce a more sensible estimate (Posterior).

THE BAYESIAN APPROACH:

PRIOR
Existing knowledge or beliefs

X

LIKELIHOOD
The new data we have

=

POSTERIOR
Our estimate

This approach can be naturally applied to marketing analysis, and once again, reflects how business decisions are made in the real world: I have beliefs about how my business works (Prior), I get new information all the time (Likelihood) and I have to use both of these to make judgements about why things happened in the past and what might happen in the future (Posterior).

(The alternative statistical approach, which is known as **Frequentist**, simply produces a likelihood based on the proportion of times it occurred in the past).

One common tool regularly used is **Econometric Modelling**, sometimes called **Market Mix Modelling**. This approach is often used to explain past sales movements as a function of marketing and sales activities and other external factors such as weather, holidays, football world cups and so on. The model tells us which drivers are statistically significant and also, critically, by how much each one affects sales (or whatever it is we are seeking to understand).   By doing this we are able to understand why past sales performed as they did. This model can in turn be used for forecasting purposes – for example would dropping price by 10% drive more sales than doubling our ad spend? A mathematical model that hopefully closely replicates actual sales is created as a result. See Overleaf:

A mathematical model to replicate sales:
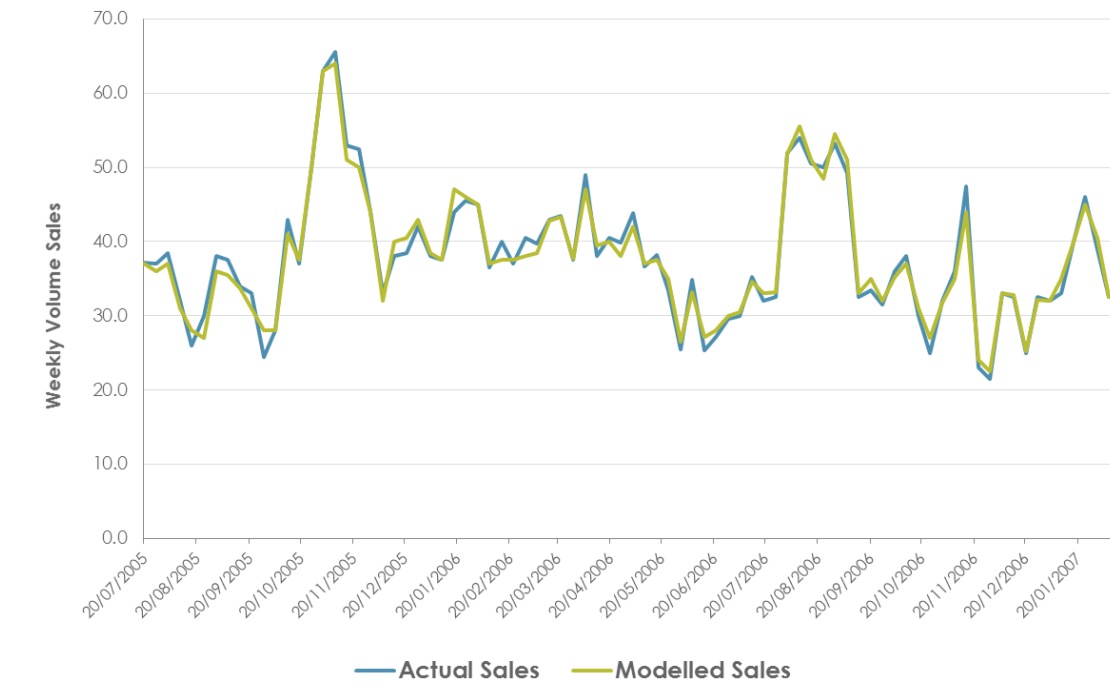
The variables and weightings within this model can be used to decompose sales and create a "due-to" report that explains WHY sales changed from year to year. In this example there were more price promotions that boosted sales but that was offset in part by loss of distribution and a reduction in Advertising GRPs.
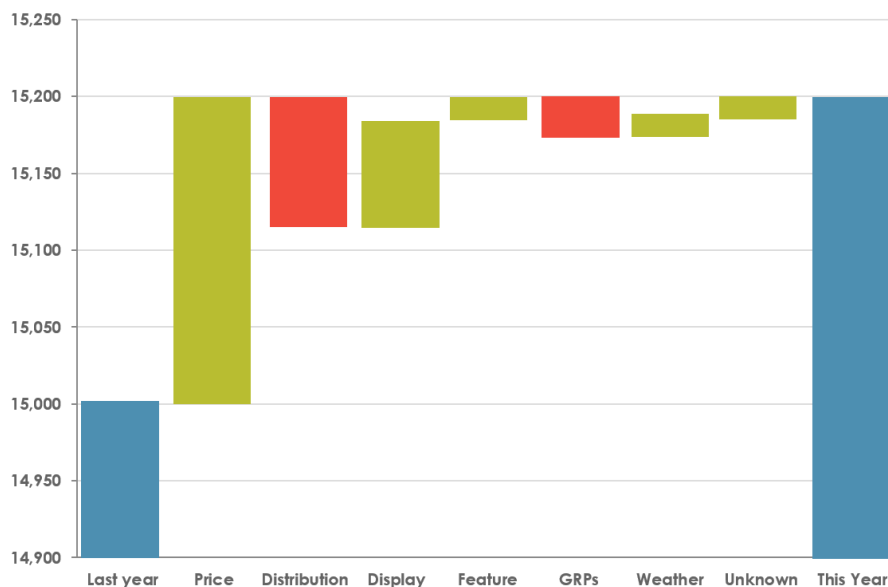
SO FAR SO GOOD. I HAVE A MODEL THAT EXPLAINS MY BUSINESS PERFORMANCE. HOW CAN BAYESIAN PRINCIPLES HELP ME?

To understand this, we need to take the lid off the black box that is the Multiple Regression process that powers econometric models.

How are these models built? The statistical formula that builds these models is essentially "dumb". It doesn't know that when price goes up, sales would normally go down. In fact the statistically "best" model that any statistics tool automatically creates (technically speaking that gives the highest R2) is almost never the most sensible solution. The variables selected are often not logical, have weightings that are too high or too low, or in the wrong direction (+ve when they should logically be –ve or vice versa). It is a purely mathematical process.

Why is this the case? In most cases the data we are modelling is imperfect. Of course if we had perfect data that was beautifully structured just for the purpose of building models, this would not be a problem at all. We would have all possible combinations of events and the model would be able to detect the impact of all the different variables perfectly. Marketers oddly enough, do not structure their activities with the sole aim of making it easy to model (I wish…). Events are often run at the same time, in all regions at once, with none of the nice variations that models need to function well. Data can be incomplete, there can be measurement errors; even simply being large does not mean a dataset is perfect.

It is therefore the job of the skilled analyst to carefully enter, modify and remove the different variables until the model "feels right". The problem is that the analyst, however skilled, can only do three things with a variable in the model: He (or she) can force it in, leave it out, or fix the weighting of it at a precise level. If that were not hard enough, adding or removing one variable can often change the weighting estimates for all the other variables. Model building is a skilled balancing act.

> *The answer is rarely just "in the data"*

The reality is that once you move away from the one absolute statistically "best fit" model, we open ourselves up to a myriad of other model solutions for that set of data. Hundreds of judgemental permutations are possible, with different combinations of variables and weightings and all with better or worse fit to the real data.

So the answer is very rarely just "in the data". We are already applying judgement to create a sensible model. By just relying on the data (especially imperfect data) we also run the risk of our models being unstable over time. This is sometimes called "overfitting". For example: one year we conclude weather does have an impact on sales, but when we come back to remodel it next year, the model says it doesn't.

That's probably not a genuine change in the impact of weather, just that the model could, with statistical significance, detect an impact one year from the data set under examination, and one year later it couldn't because we had added new data points to the model and maybe other things were going on that masked that effect and made it statistically non significant.

THE BAYESIAN APPROACH TO MODELLING

In the Bayesian approach we accept that we will be applying judgement to our model. But there are two key differences:

1 This is agreed and shared explicitly **before** the modelling process begins rather than during it.

2 The judgement is applied and specified and described mathematically.

WHAT KIND OF PRIOR KNOWLEDGE MIGHT WE HAVE AND HOW CAN IT BE REPRESENTED "MATHEMATICALLY"?

We might have a range of Prior information that we may want to represent in our model. This often falls into two areas:

1 Logical, common sense constraints. We might (or might not) all agree that certain variables should have certain **directions** to their weightings. Do we all agree that price elasticity is usually negative (as price goes up, sales go down?). We are not necessarily specifying the exact weighting of that relationship…the model can tell us that…just that it can't be the wrong direction.

2 Valuable information from other studies. For example, we might have done a huge amount of pricing research and know pretty well what the price elasticity of our product is. We want to represent the level of our certainty in that figure in the model. We might have done a huge weather study and know that weather definitely has an effect on sales and have a feeling for the approximate magnitude of that.

We are in effect "helping" the model; in other words telling it where to start looking for a solution. Here are some examples of how we might decide to enter our level of Prior knowledge about price elasticity into a model. It is done in the form of probability distributions:
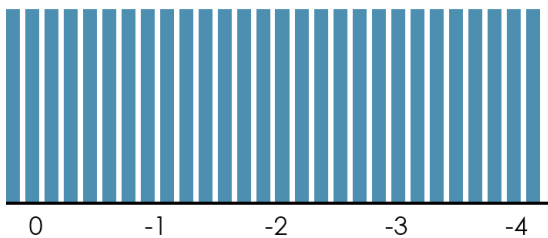
We believe price elasticity is around minus two with some variation either side, but cannot be positive.
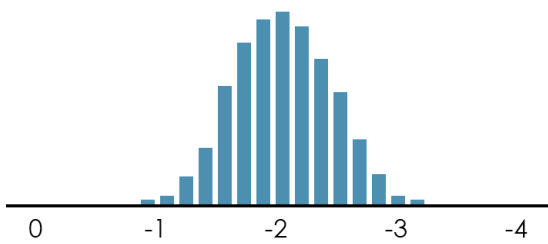
We have very strong evidence that price elasticity is close to minus two.
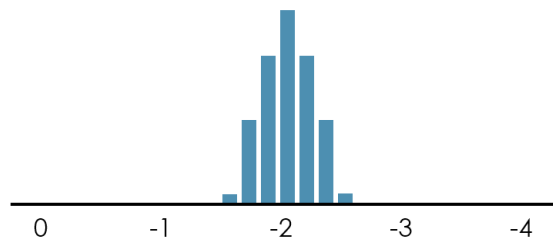
We have no idea of the true value, but it cannot be positive or more than minus four. All values are equally likely.

We believe minus one is the most likely value. It cannot be positive and there is some chance is could be as high as minus four.
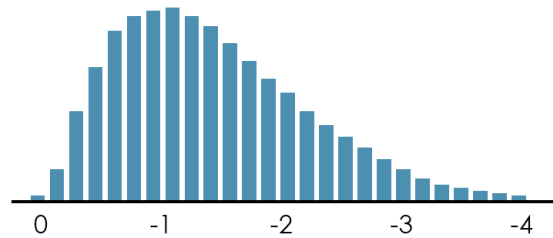
We would not suggest putting Prior constraints on all variables – some would normally be left to float freely. And, of course, the model has to fit the data to an acceptable degree. So a typical discussion might go along the lines of… well if you really strongly believe that weather and pricing have a large impact on sales then, in order for the model to fit the data, you have to believe that advertising has little or no effect. This can be a very valuable discussion to have.

We need to check the model fit, of course. If our Prior knowledge just simply doesn't fit to reality and we can't get any kind of sensible solution then we need to revisit those assumptions. Critically, the Bayesian approach allows us to challenge and reject management hypotheses if necessary in a structured approach.

Again, if we had very high quality data, we could just let the model answer all of these questions. In Bayesian terminology, our Likelihood would be very strong and dwarf the strength of any Prior we had. 30 games into a 38 game season, your start of season belief (Prior) that a certain team (now horribly struggling in the football league) would win the championship is now massively challenged by the real data (Likelihood). We need to continually ask ourselves about the balance of the strength of the data and our Prior. In contrast if we had very poor data but very strong Prior beliefs, we might choose to let our Prior have more weight over the data.

The process of gathering and agreeing how we represent Prior knowledge in the model is called "elicitation" and is an important stage in any Bayesian analysis. It is also an ideal opportunity before any model is run to get the key stakeholders in a room and discuss what we know, and with what level of certainty and proof. This helps to separate myths and hunches from proven findings and assesses to what extent beliefs are shared by different people in the organisation. This by definition tends to lead to greater management buy in to the model results. It is rarely helpful to have model findings heavily challenged for not fitting with what is already believed during the final presentation. It is much better to have the discussion of Prior knowledge, well... prior to that.

We should also consider **why** we build models. I would suggest that we build models firstly to obtain the variables and weightings of those variables that describe our sales. So for example I want to know what my price elasticity is. However, a good model should help the organisation **learn**. It should be the repository for all our knowledge about how our sales are impacted by activities and events.

> " *Essentially all models are wrong, but some are useful.* "
>
> George E. Box

Our uncertainty should be explicit and openly shared and the process of model building should stress test and strengthen our learning. Finally a good model should also be **useful** to management to help manage the business. A model that is not believed will not be used. As George Box famously said: "All models are wrong, but some are useful".

CAN YOU GIVE ME AN EXAMPLE?

Imagine a model where we have variables such as distribution (the % of stores stocking our product), the number of stores that have a chiller cabinet, as well as the usual variables of pricing, advertising, weather etc. Our first pass at modelling this data might give a very high elasticity for distribution. It would imply that if we increased the number of stores stocking by 1%, we would increase sales by 4%. That would be a good trick, but is unlikely! This was really driven by the nature of the data set and distribution was in effect "soaking up" the impact of other activities and therefore giving illogical results.

The other prior information available might be from a very comprehensive test and control store study of the effect of chiller cabinets that gave a very accurate read as to the uplift. So we might specify a common sense Prior distributed around a distribution elasticity of 1 (a 1% increase in distribution should give roughly a 1% increase in sales) and specify another prior around the known impact of chiller cabinets. The result would be a much more sensible looking result – we would hope that the other variable weightings would change to be within much more logical ranges. This is in fact what we find in the majority of cases.

SO IF THIS IS SO GREAT WHY AREN'T BAYESIAN APPROACHES IN WIDESPREAD USE ALREADY?

There are three main reasons for this:

1 It's difficult technically. The formulae used to do this are so difficult to solve that actually it's not practical to solve them. Instead the calculations are done through running many computer generated simulations. Imagine you wanted to know the probability of tossing at least 700 heads out of 1000 coin tosses. You would have to do a huge piece of maths to calculate the probability of getting 700 heads, plus the probability of getting 701 heads, plus the probability of getting 702 heads etc. Actually it's much faster to get a computer to mathematically "toss" a set of 1000 coins say 10,000 times and just observe the result. With enough iterations, the outcome will be almost identical to the formula result. The development of Markov Chain Monte Carlo methods (MCMC) in the 1980's helped make this approach feasible.

2 Linked to the above, the available computational power was simply not capable of doing even these simulations until relatively recently. Suffice to say the computer we use to run these analyses is affectionately known to us as "The Beast".

3 Lastly, there is a debate going on, and has been for many years, between the Bayesians and the traditionalists (more accurately called "Frequentists"). Here is not the place to get into that but it's a largely philosophical debate over the use of Prior information. The upshot of this is that the Frequentists held sway in statistical "royalty" for many years – Ronald Fisher, who in many ways invented the process and language of traditional statistics you recognise today, was strongly opposed to Bayesian principles for most of his life.

IN CONCLUSION

Bayesian approaches are becoming more widespread in many areas, but have yet to be widely used in market modelling. Some companies are realising that, especially where data is poor (for example in developing markets), better models that are more useful for management can be built using Bayesian approaches – for example applying knowledge from similar countries or other brands in the category in order to help the model find a sensible solution. Not every modelling situation is suitable for the Bayesian approach but it is a very powerful and little understood tool.

I hope this inspires you to find out more about Bayesian theory and think about applications to marketing problems. Most of what's written about it is largely impenetrable to the non-statistician unfortunately. Hopefully this paper helps to demystify it somewhat.

If you are interested in exploring how Bayesian approaches might help your business decision-making or marketing planning then please get in touch.

ABOUT THE AUTHOR

Tom has worked in research and insights for 24 years and is the founder of new breed marketing analytics agency MetaMetrics. Prior to this he was a research head at Nestlé and held senior roles at Kraft Foods as joint head of UK Consumer Insight and European Analytics Director followed by a Global Analytics role for SABMiller.

Tom is a Full Member of the Market Research Society and leads the training programmes for Econometric modelling and advanced statistics.

ABOUT METAMETRICS

We are a new breed of agency. We were established in response to the blurring of lines between advertising, direct and digital marketing. As modern marketing becomes increasingly integrated we have broken down the silos that typically existed in analytics to provide insights across market, channel and customer.

We have brought together deep experience in brand marketing, econometric modelling, customer analytics, direct / digital marketing planning and technical innovation. This enables us to provide a fresh approach to solving our clients' challenges through analytics.

Independent, impartial and innovative, we work with many of the country's best known brands.

< **MetaMetrics** >
integrated marketing analytics