An Introduction to Bayesian Analysis Russ Lavery

INTRODUCTION:
This booklet is intended to support a seminar on introduction to Bayesian analysis. It is intended to ***not*** be mathematical but, instead, to focus on insight and making the concepts understandable to people who were not majors in mathematics or statistics.

BAYES' LAW PROVIDES USEFUL, AND COUNTER-INTUITIVE, RESULTS

The data in this next example is real and most doctors get this answer wrong. We are not going to do much explanation in this section. This is just one worked example that, hopefully, will get you interested in the subject.

A mammogram has an 87% chance of detecting breast cancer when it is present. A mammogram will falsely signal breast cancer 12% of the time (when it is not present). About 1% of women between 40 and 60 have breast cancer.

If a mammogram signals cancer, what is the chance of the subject having cancer? To get the answer we should use Bayes' law.

$$P(C \mid \text{Positive Test}) = \frac{P(\text{Positive Test} \mid \text{Cancer}) * P(\text{Cancer})}{P(\text{Positive Test})}$$

$$P(C \mid \text{Positive Test}) = \frac{.87 * .01}{(.01*.87)\ (.99*.12)} = \frac{.0087}{.1275} = .0682 \text{ or } \sim 7\% \leftarrow \text{does this surprise you?}$$

Most people guess a number closer to 87%. The false positive percentage I used is the upper limit on the actual percentage (7% to 12%). If a woman were to get 10 yearly mammograms the chance of at least one false positive result is about 50%. Thankfully breast cancer has been declining since 2000.

HISTORY:
The Bayes formula is not new. The original insight was created by Thomas Bayes who lived between 1702 and 1761. A little history might help in understanding how we got to where we are.

In 1662 the English Parliament passed the act of uniformity which required all English subjects to use the Book of Common Prayer. Thomas Bayes father objected to this and was what was called a "non-conformist" minister. Nonconformists were any non-Anglican or non-Christians (Methodists, Baptists, Congregationalists, Puritans, etc.), groups that are pretty mainstream today in America. Nonconformists could not get university degrees in England for over 150 years and so Thomas Bayes got his degree in Scotland.

He never published a math paper in his life but was elected to be a fellow of the Royal Society. The formula we know as Bayes' law was discovered in his papers after his death and was published by a friend. I have heard that Bayes was trying to discover a formula to prove the existence of God. It is possible he never saw the full practical applications of his concept. The formula that we know as Bayes' law was formalized by Pierre-Simon the marquis de Laplace in *Essai philosophique sur les probabilités* (1814). Laplace wanted to apply his new idea, that he called inverse probability, to things like courts of law. I think that you can see some of this influence in our famous statement "a man is assumed to be innocent until proven guilty beyond a reasonable doubt." The reason people think this is Bayesian is because Bayes' law models how we update our beliefs in the presence of new information/evidence.

For a long time Bayes' Law was not used very often because calculations are difficult and because of opposition to priors.

R. A. Fisher (1890-1962) has been described as "a genius who almost single-handedly created the foundations for modern statistical science" and "the single most important figure in 20th century statistics". Oddly, Fisher was primarily an agricultural researcher and was concerned with improving food production. He was the first to use diffusion equations to attempt to determine the distributions of different forms of genes (allele) by analyzing genetic links applying MLE to populations. He has been called the greatest of Darwin's successors.

Fisher was a very powerful statistician and Fisher hated Bayesian Analysis. He wrote "The theory of inverse probability is founded upon an error and must be wholly rejected." He attacked people who supported Bayesian analysis and he was a powerful deterrent.

Neyman and Pearson helped put Bayesian Analysis on the back burner when they helped formalize/systematize the methodology of research. They developed the methods of hypothesis testing and confidence intervals that revolutionized applied and theoretical statistics. It was not victory of competing theories but of practical application. Bayesian calculations were difficult and Fisher, Neyman and Pearson had good practical methods for doing analyses.

A recent book, "The Theory that Would Not Die" asserts that there was lots of Bayesian analysis done during World War II but that the projects were all classified – during the war and for a long time afterwards. The book asserts that Bayesian analysis was used to crack the Enigma code. Bayesian analysis is asserted to have been used to determine the number of German tanks.

The book also cites a story about using Bayesian analysis to discover a lost nuclear bomb in the ocean. I have read several accounts of this and would like to fill in some of the information that's sometimes is mentioned and sometimes not. Bayesian analysis is supposed to use prior information. A local fisherman reported seeing something fall into the ocean right next to where the bomb was found. That information seems to have been ignored until late in the search.

The ocean bottom had been divided into fairly large squares and each square was assigned a calculated probability of containing the bomb. As each square was searched, and found to be not containing the bomb, the probabilities for all of the squares had to be recalculated. There was relatively little computing power on the search vessel and so the updates of probabilities took quite some time. The great amount of time required to update the probability for a particular square had forced the squares to be fairly large.

The bomb was found, very close to where the fisherman said it had landed, in a low-probability square but near the border between that low probability square and a higher probability square. After reading several accounts of this process I have an odd feeling that if they had listen to the fisherman at the beginning they would've found the bomb quite quickly.

However; Bayesians claim this as a success. I have listened to many Bayesian statements and I think the claim that we are all Bayesian's is founded on the idea that we all use prior information in making decisions. We learn from our mistakes and, hopefully, don't repeat them. When a Bayesian sees a mouse learning to find cheese at the end of the maze the Bayesian say "mice are Bayesian". I leave this discussion to other people.

I would like to take a bit of this paper to explain what I have gleaned from readings about cracking the German code in World War II. Alan Turing was once asked, of a technique he had invented, “Isn't this really Bayes' theorem?" His reply was “I suppose so.” and I find that a lukewarm confirmation.

The code cracking problem was that the Germans had a mechanical coding machine with several “swappable” rotors”. The number of possible combinations of rotors in the German code machine was astronomical. Applying brute force methods, just trying a particular combination and seeing if it worked, would have taken too long.

However; several German radio operators violated common secrecy practices allowing the British to have “cribs”. British universities used to focus on translating Latin and Greek into English. A “crib” was used by students to help them in their translation homework. It was basically an English version of the Latin text. Students would pass down completed translations, “cribs”, to younger students and these “cribs” greatly helped struggling students.

Weather ships often sent the same information every day in the same way (longitude, latitude, rain, wind direction, wind speed, cloud cover etc.). One German radio operator was in love with a girl named Cecilia and, every day, he would set the code wheels to the settings for that day and then broadcast to the world his girlfriend’s name. The code-breakers knew the contents of these messages and this was a great help. Some books suggest that the codebreakers didn’t do much work on a day until the “cribs” arrived. One book has a code breaker asking “have we heard the silly yet” as a reference to the coded version of Cecilia.

Having the cribs greatly reduced the number of “rotor setting” combinations that had to be considered but that number was still a very large number. I’ve heard that the cracking method was to take all the possible combinations of “rotor settings” and apply each setting to a large number of messages. If the combination was wrong, the message would be nonsensical. If the combination was right then the message that came out would make sense to a human. But there were too many combinations for a human to try and read the output.

The ingenious trick I heard about involved letter frequencies. Each language has frequency of use for each of the letters. In English, E is the most common letter and Q is relatively infrequent. German has a letter frequency that is different from English. Letter frequency was key to cracking Enigma.

After the British computer produced a translation for a hopeful setting of the coding machine wheels a little program calculated the relative frequencies of the individual letters in the message. The observed relative frequencies were compared to the relative frequencies to be expected from German military messages. Messages where the sum of observed minus expected was small were sent to human to read.

If this sounds like the chi-squared formula to you, you are correct. So maybe chi-squared cracked the Enigma. Maybe it’s unfair to say that Bayes’ Law cracked the code produced by the Enigma machine simply because we used prior knowledge contained in the cribs.

Stories being heard now are that the military continued to apply Bayes’ law after the war but kept all the success stories confidential. In summary: the difficulty of computing answers, the opposition by giant figures in the field of statistics and the lack of publically available success stories about how Bayes’ law had been applied contributed to Bayesian analysis being a minor activity.

Most writers suggest that the increase in Bayesian use came about when the MCMC algorithm was matched up with the increased computing power that started becoming available in the 1960's. This made it practical to solve Bayesian problems – removing one of the main factors in the dominance of frequentists.

The rise of Bayesian Statistics, while benefitting from MCMC and computing power, was a complex process and many people contributed. A fine history is "When Did Bayesian Inference Become Bayesian"? by Stephen E. Fienberg https://projecteuclid.org/download/pdf_1/euclid.ba/1340371071

YOU ARE A BAYESIAN.

People who are proselytizing Bayesian analysis will say we all are Bayesians. I've even read that dogs are Bayesian. The support for this statement is that people who make statements like this believe that anyone who uses prior information is using a Bayesian process. I do think that we all update our plans when we hear new information but am not sure that we all apply, in the neurons of our brain, a formula that's like the Bayes' Law. I don't think our minds are that simple and would suggest that reading a book called "Thinking Fast and Slow" is well worthwhile.

BAYES LAW: THE FORMULA AS IT IS USED BY A FREQUENTIST

The Bayes' law formula has many looks, and this can make it a bit confusing. We will endeavor to use worked examples to give concrete illustrations of the words that are used in Bayesian statistics and the parts of the formula.

P(A|B) * P(B) = P(B|A) * P(A)
is the basic formula and, to the author, offers no insight. We ask for the reader's indulgence while we state some theory.

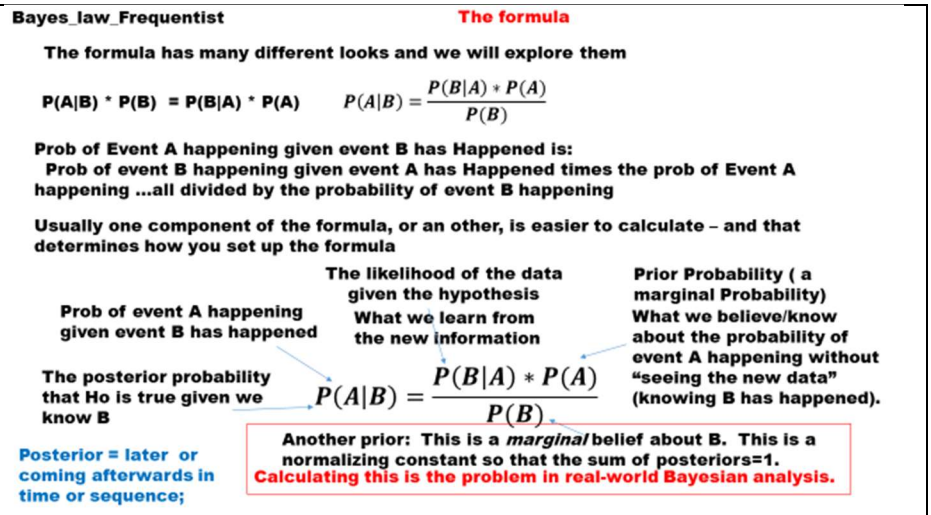


Figure 1

I suggest a quick read of the theory – then a study of the examples – and then looping back to the theory again.

Bayesian analysis is the updating of our beliefs in the presence of new information. We represent our prior belief by P(A) and the new information by P(B). The most commonly seen version of the formula is shown below.

$P(A \mid B) * P(B) = P(B \mid A) * P(A)$ → very simple math → $P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$

$(A \mid B)$ is the posterior probability- our belief after we have seen the new information. It says that the probability of event A happening, given that we know event B has happened, is equal to the right-hand side of the equation.
$(B \mid A)$ is the likelihood of event B happening given that event A is happened. It is the likelihood of the data given that our hypothesis is true (more on this later). This is an inverse probability, at least inverse to the posterior probability.
(A) is the prior probability. It is our belief about the probability of A happening without knowing about B.
(B) is a normalizing constant. This is a prior belief about B occurring. It makes the result of the calculation a true percentage and will need to see some more information before we can explain why that is true.

NOTE: In formulas, I will flip between using A and B in formulas and other letters. A and B show up in books and we should see that notation. However; I find that using S for shy and A for actuary makes things easier to follow.

Let's use an example to make some of these concepts more concrete. Imagine going to a university event that is a mixture of marketing and actuaries (apologies to all). Your +1 says that they met a shy person. You think that actuaries are shy and wonder what was the chance that this person was an actuary.

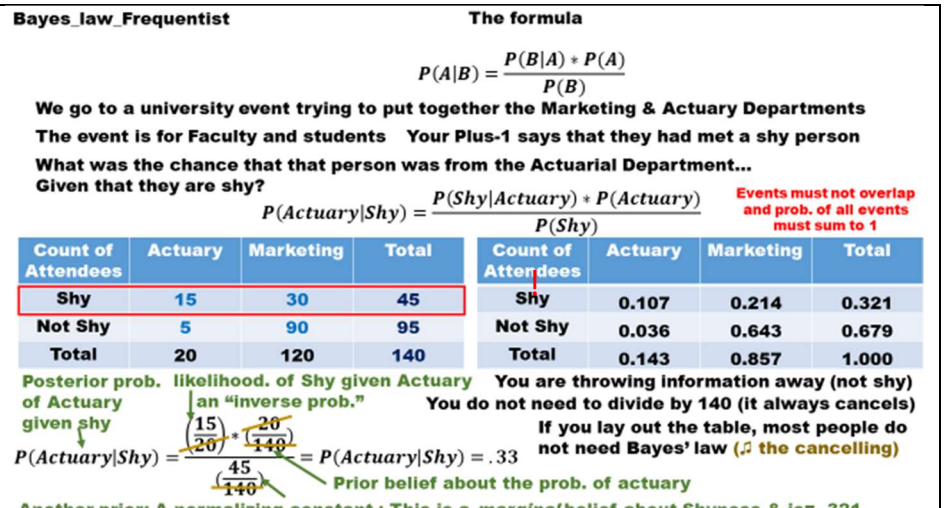


Figure 2

The table, in Figure 2 allows us to work out Bayes' Law. In fact, if the problems are laid out in tables most people do not need Bayes' Law to get the correct answers.

The basic equation of Bayes' law can be seen in Figure 2.

The probability of being an actuary and shy (see exclamation point) is .107 <u>of the total number of people</u>

and, importantly, we can get to that .107 two different ways: P(A|B) * P(B) = .P(B|A) * P(A) = (A ∩ B). This is evident from 2x2 tables as seen in Figure 2

The next two lines illustrate the equivalence of P(A|B) * P(B) = .P(B|A) * P(A)

Prob(S | A) * P (A) ← = → P(M | S) *P(S)
(15/45) *.(45/140) = .1071 ← = → (15 / 20) * (20 / 140) = .1071

Above is the basic logic. Using simple division on P(A|B) * P(B) = .P(B|A) * P(A) gives Bayes' law below.

$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$ ← so we have a proof of this ☺ and can go on and use the formula.

That .107 number is the percentage of shy **and** actuary as a percentage of the total number of people. Our question is "what's the probability of being actuary given that we know the person was shy". Attendees at the meeting were a mixture of shy and "not shy".

Please look at the red box in Figure 2. Given that we know you're shy we throw away part of the information in the table. Given that we know you're shy we only are concerned with the information inside the red box. Bayesian analysis always throws away information and, if we look at the red box, we are throwing away the information about people who are not shy.

What does the denominator do for us? *The answer only can be seen if we do TWO calculations*.

$P(A \mid S) = \frac{P(S \mid A) * P(A)}{P(S)}$ → $(15/45) = \frac{P(15/20) * P(20/140)}{P(45/140)} = \frac{.75 * .1428}{.3214} = .333$

$P(\text{Not } A \mid S) = \frac{P(S \mid \text{Not } A) * P(\text{Not } A)}{P(S)}$ → $(30/45) = \frac{P(30/120) * P(120/140)}{P(45/140)} = \frac{.250 * .857}{.3214} = .666$ 1.00 ☺

The denominator is a normalizing constant that makes the probabilities of the two possible outcomes (actuary or not actuary) sum to one and this means this calculation returns a true percentage and probability. Remember, if you sum the probabilities of occurrence, for all things that can occur, the sum must equal one.

Bayes_law_Frequentist - practice **The formula**

The formula has many different looks and we will explore them

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} = P(F|needs\ help) = \frac{P(B|A) * P(A)}{P(A \cap B) + P(\hat{}A \cap B)} = = \frac{P(needs\ help|F) * P(F)}{P(F \cap Help) + P(M \cap help)}$$

You are always throwing information away – 20 people need help

Count of Students	Needs Reading help =YES	Needs Reading help =NO	Total
Male	15	30	45
Female	5	90	95
Total	20	120	140

P(F | Yes) = $\frac{P(Yes \mid F) * P(F)}{P(Yes)}$ = _____ =

P(M | Yes) = = _____ =

P(Yes | M) = = _____ =

P(No | M) = = _____ =

P(Yes |F) = = _____ =

P(No | F) = = _____ =

Figure 3

I think this formula is easy to calculate but kind of confusing and therefore worth a couple of examples. Figure 3 shows another 2x2 table and gives us some space to do some calculations.
The answers are below. Arrows point to formulas with the same denominator and these should sum to 1.

Answers for Figure 3				
P(M|Y)	= 0.75		P(Y|M)	= 0.333
P(M|N)	= 0.25		P(Y|F)	= 0.053
P(F|Y)	= 0.25		P(N|M)	= 0.667
P(F|N)	= 0.75		P(N|F)	= 0.947

Let's look at how this formula can quickly become confusing.

P(help) = 20/140 and that can be written as P(help ∩ M) + P(help ∩ F) ←these are just cell counts of 5 and 15. We could also write this as P(help ∩ Male) + P(help | NOT Male) and some books do,
The situation is: Many formulas and little insight.

Here is a major hint – it changes Bayes' law into English words and provides insight/

$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$ → P (A|B) = the probability of A and B occurring together divided by the probability of all the ways B can occur.

This can be helpful in word problems. The numerator is the probability **of A and B both occurring**. The denominator is just the **sum of the probabilities of all the different ways that B can occur.**

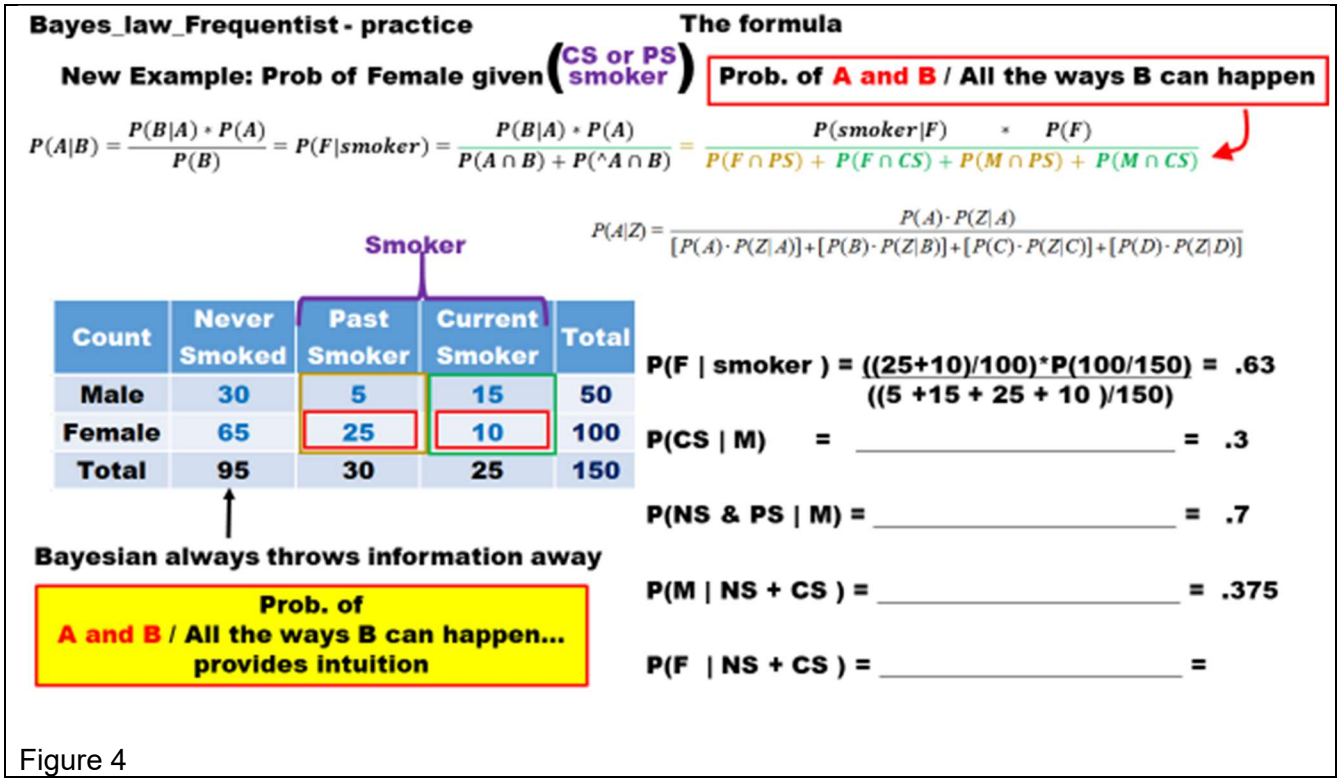


Figure 4

Figure 4 is another chance to do some calculations and emphasizes the helpful hint mentioned above. Here we have three categories and this allows us to emphasize that the denominator is the sum of the probabilities associated with all of the different ways that B can occur.

I think saying that Bayes' law is the probability of A and B occurring together divided by the probability associated with all of the different ways that B can occur allows me to check my logic in more complicated tables.

The red rectangles shows the probability of A and B occurring at the same time – that is the probability of being a female and a smoker (past or current). The orange and the green boxes show all the different ways you can be a smoker (past or current). Some of the calculations requested above are a bit odd but they will allow us to explore the Bayesian formula on something other than a 2 x 2 table.

Some books explain Bayes' law using a Venn diagram and I think that's not the best way. Please look at the Venn diagram in Figure 5 and compare it to the cross tab in Figure 5.

The Venn diagram shows the shy and actuary and the intersection of the two characteristics.

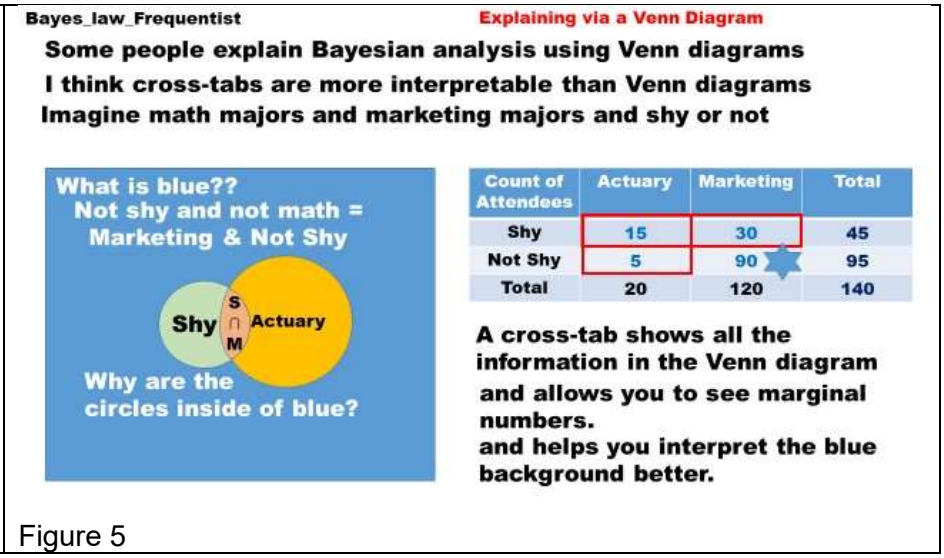


Figure 5

I think a cross tab is superior because it shows the same information as the Venn diagram and more. The crosstab shows the actual numbers that are associated with the characteristics of the people. The crosstab also shows marginal totals and information about the problem that is not shown well in the Venn diagram. Marketing people that are not shy are the blue area in the Venn diagram, but that is not obvious.

I would suggest that, if you have a 2X2 table and you want to make a Venn diagram, draw little boxes on the table itself. That will be more informative than creating a Venn diagram.

A SHORT DIGRESSION ON THE IDEA OF A LIKELIHOOD

Bayesian analysis often uses likelihoods as a way to avoid calculating the denominator in the Bayes' Law formula. It would be good to spend some time thinking about the difference between a likelihood and a percentage.

In the upper right corner of Figure 6 we see the formula for the height (the PDF) for the normal curve.

The red denominator in the formula is a normalizing constant. It makes the area under the normal curve integrate to one.

The fact that the area under the curve integrates to one is a requirement, if areas under the curve are to be interpreted as a percentage and a probability.

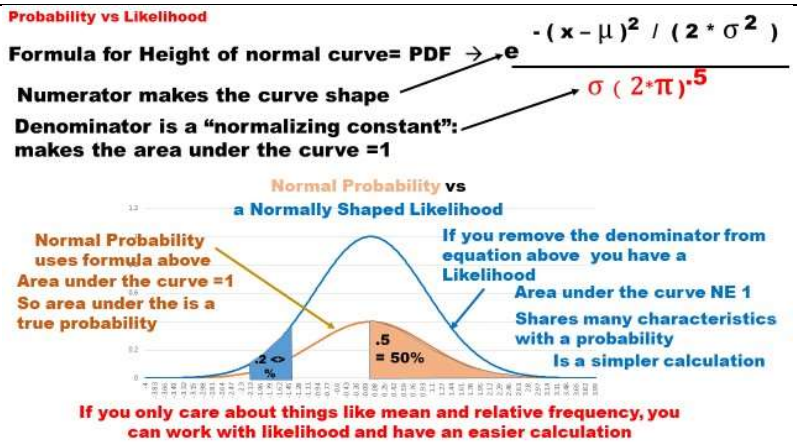


Figure 6

The fact that the area under the tan curve integrates to one allows us to interpret the number we get when we integrate under that curve– the one area number that is the output of the integration - as a percentage and a probability. If we integrated under the tan curve from Z equals 0 to Z equals positive infinity the area would equal .5 and because the area under the curve sums to one we can interpret that area as a percentage and a probability.

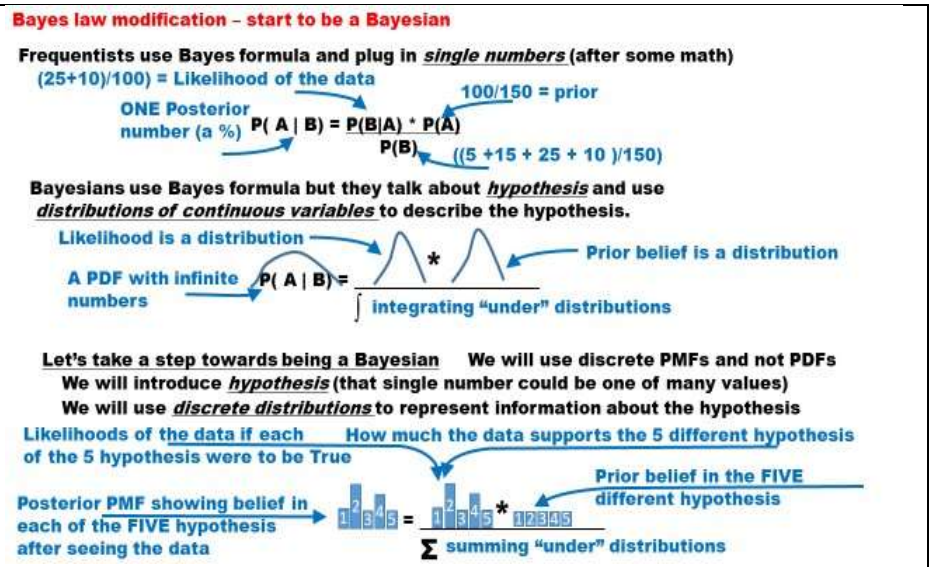
If we remove the denominator from the "normal" formula we get the blue curve in Figure 6. If we integrated under that curve we might get a number - say 10 as the area under the curve. If we integrate to find the blue shaded region we might get a number like .8. We cannot say that the probability of getting a number in that range shown by the blue distribution is 80%. Results of integrating under the blue curve cannot be directly interpreted as percentages or probabilities.

However; the two curves have a lot of similarities. Their centers are the same and they have the same shape. If that denominator is difficult to calculate, and all we need are a few simple characteristics of the curve, it is sometimes easier to work with the likelihood curve (blue) then the normal curve. It just saves us some calculations.

Remember that calculating the denominator in the formula for Bayes' Law is usually the difficult part of the formula. The denominator is a normalizing constant so that the area under the curve sums to one and so that the areas can be directly interpreted as percentages. Will see this in the examples to follow.

BAYES LAW: USING THE FORMULA MORE LIKE A BAYESIAN WOULD USE THE FORMULA

Example 1

Frequentists use Bayes' Law quite often and use it as we have shown above. They tend to use single numbers (or the results of simple calculations) as each of the components of the formula. They are certain of the values that they are entering into the formula and therefore only need to enter one, scaler, number. .	 Figure 7

Bayesian's have a different approach and are less certain about their numbers Because they are less certain about their numbers Bayesian's will describe their prior belief in someone being shy as a distribution. A Bayesian might say the following.

"I think 32% of the students at this university are shy but I want to have some uncertainty. My best guess is 32% but it might be 30% or maybe 35%. I'm pretty sure it's not 50% and very sure it's not 60% or 5%. Rather than providing you one number about my belief I'd like to provide a distribution."

That is what we are working towards but we want to get there in baby steps. There are some mind–stretching concepts to get through before we can apply Bayes' Law as a Bayesian. It's going to be useful to do several examples so that we can see pictorial representations of these new concepts.

The top formula in Figure 7 is Bayes' law as used by a frequentist. Each part of the formula is one number and this implies we are very certain about this number. Frequentists look only at the data that they have and so they can be very certain about the characteristics of the data that they collected.

The middle formula in Figure 7 is Bayes' law as used by a Bayesian. Each of the parts of the formula are continuous distributions. The divisor is actually an integral over part of a distribution and this is where Bayes's Law gets really difficult. In the numerator we are going to be multiplying two distributions. In the denominator were going to be doing an integral over a joint distribution. Astronomers, when they're using Bayesian analysis to try and find a planet around another star, might have 15 to 20 variables describing the joint distribution in the denominator. We mentioned before that the Bayesian calculations were difficult. It's the calculation of the denominator, the integral over many different X variables, which makes Bayesian calculations so difficult. In fact, there are only a few exact, or closed form, solutions to Bayesian problems. Modern Bayesian analysis uses approximation algorithms, not closed form equations.

So frequentists use the formula with only one number for each of the components. Bayesian's use the formula with continuous distributions as each part of the formula.

Our next step is shown in the bottom of Figure 7. We are going to take a “baby step” towards using a continuous distribution by using a discrete distribution. Instead of taking integrals, we will be able to sum numbers. This will simplify calculations but still give concrete examples of the difficult concepts.

Imagine we have several dice in a cup. You can buy four sided dice six sided dice eight sided dice 12 sided dice and 20 sided dice and they look like the picture in the upper right-hand corner of Figure 8.

We put the dice in a cup and allow someone to, without showing anyone else, pick a die and then roll it. That person tells us the number on the dice and says it is a 12.

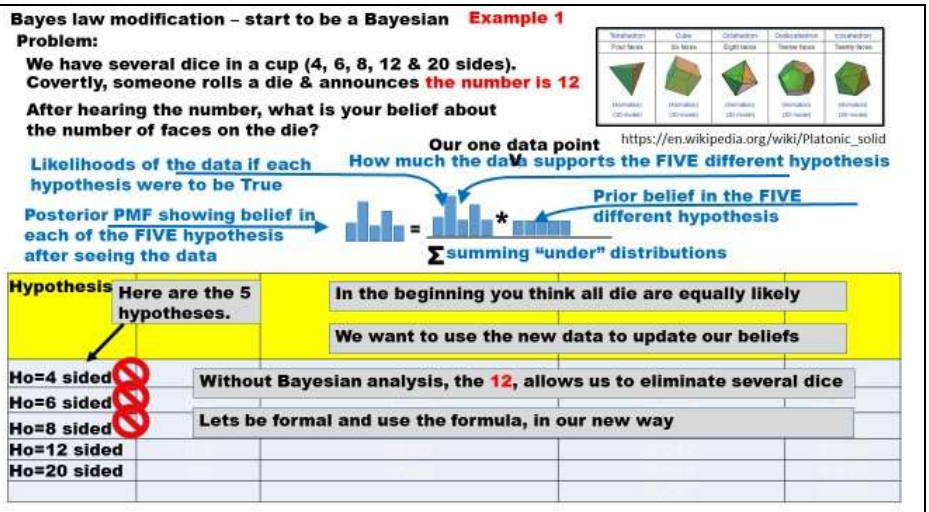


Figure 8

Figure 8 shows a shell of a “calculation aid” table. We’re going to gradually fill in the columns in a way that mimics the formula for Bayes’ Law.

THE HYPOTHESIS: (in example 1)

We would like to use Figure 8 to help understand the concept of a hypothesis. Hypotheses can be true or false and we are setting up five competing hypotheses. Our competing our hypotheses are shown in the leftmost column in Figure 8.

Since we only have five dice we are going to use a discrete distribution. If we could create dice that would roll a fractional number like 3.1725 we could make this into a continuous distribution – and then we would have to do different math. Let’s use discrete distributions, in the next few examples, so that we can develop the concepts.

In Figure 9 we add our prior belief - how likely we think it is that each die was picked.

This example illustrates us having no idea, before hearing about the 12, which die would be picked.

Our idea is that without any information, and five dice in the cup, we would think that each die has a 1/5 chance of being selected.

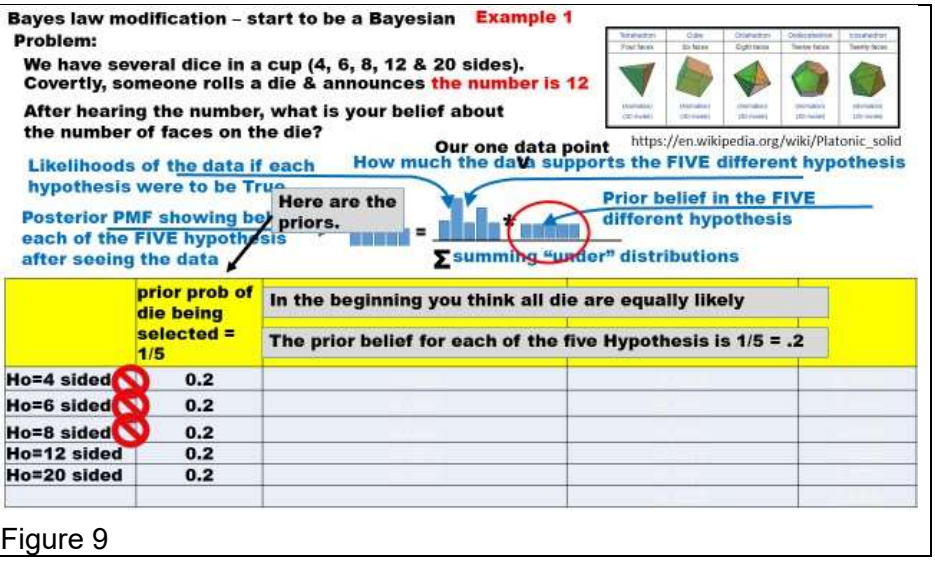


Figure 9

That is our prior belief in the probability of each hypotheses.

Figure 9 shows a prior belief in the different hypotheses that would be called a flat prior or an uninformative prior. That prior describes our belief in how likely each die is to be selected – our belief in the truth of each hypothesis. We don't have any idea which die would be selected. This column corresponds to the prior distribution and that part of the formula is indicated with a red circle. Notice that, in the picture, the prior distribution, is flat.

This is our prior belief in the hypotheses and we want to update our belief in the hypotheses based on new information (the 12). I'm sure you realized, and didn't need Bayes' Law to understand, that if the dice rolled the 12 it could not be one of the dice with four sides or six sides or eight sides. We didn't need Bayes' Law for that. We want to use Bayes' Law, and formalize, the understanding that you already have.

In Figure 10 we add what are commonly called "the likelihoods" and I'd like to add some more words to that much too short phrase. The column we added shows the likelihoods of getting the data we saw if the Ho on that row is true.

The likelihoods part of the formula is circled in red on the slide.

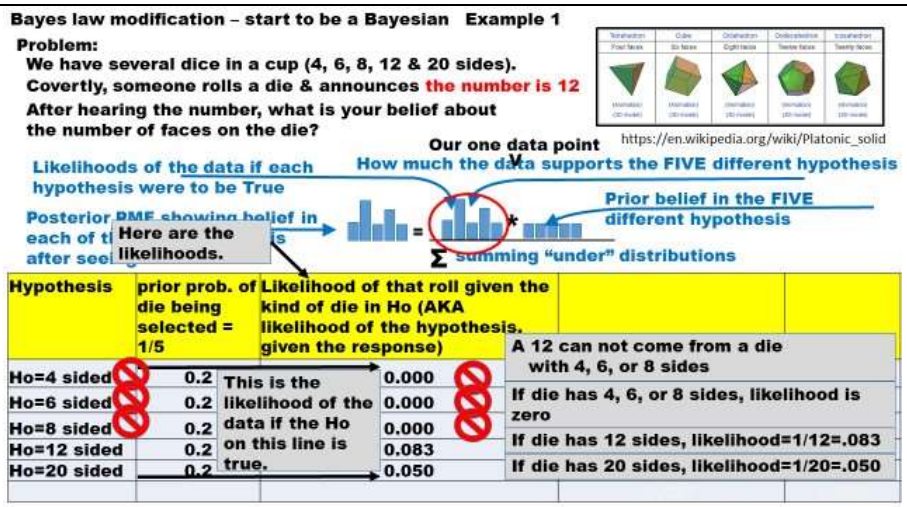


Figure 10

The heights of the bars are only illustrative of the concept and are not proportional to the likelihoods. I will use this blue graphic in many places and updating the sizes tiny boxes too tedious. However; if we were to update this graphic to match the data in this slide the first three bars would have zero height.

In the Bayes' formula, this part of the formula is P (B|A) and this stands for the likelihood of the data occurring if the hypothesis is true. If you have a 20 sided die, any particular number has a 1/20 chance of occurring. If you have a 12 sided die, any particular number has a 1/12 chance of occurring. These are the likelihoods of getting a 12 given that you have a 12 sided die or a 20 sided die.

Figure 11 adds a column that does the multiplications for the numerator of the Bayes' Law formula. This action is indicated by the red oval on the graphic. These numbers do not sum to one and so they are not percentages or probabilities.

They do show our relative belief in the hypotheses but they cannot be interpreted as percentages or probabilities.

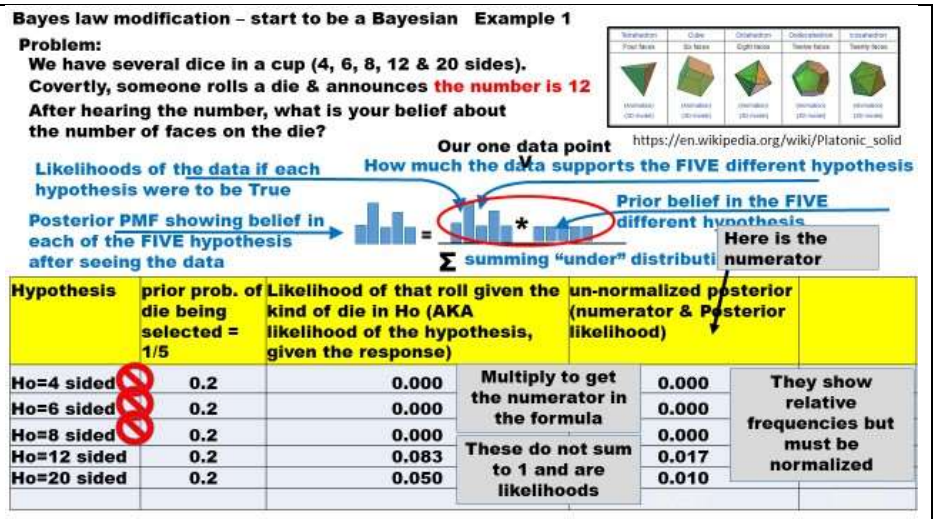


Figure 11

To do that we must bring in the denominator – the normalizing constant.

Figure 12 shows the completed calculations. If we add up the column called un-normalized posterior we get .27 and that is the normalizing constant. We use that to create the posterior percentage. As examples:

.017÷.027 equals .625. and .01÷.027 equals .375.

These numbers sum to 1 so we can talk about the ***percentage chance*** of these hypotheses being true.

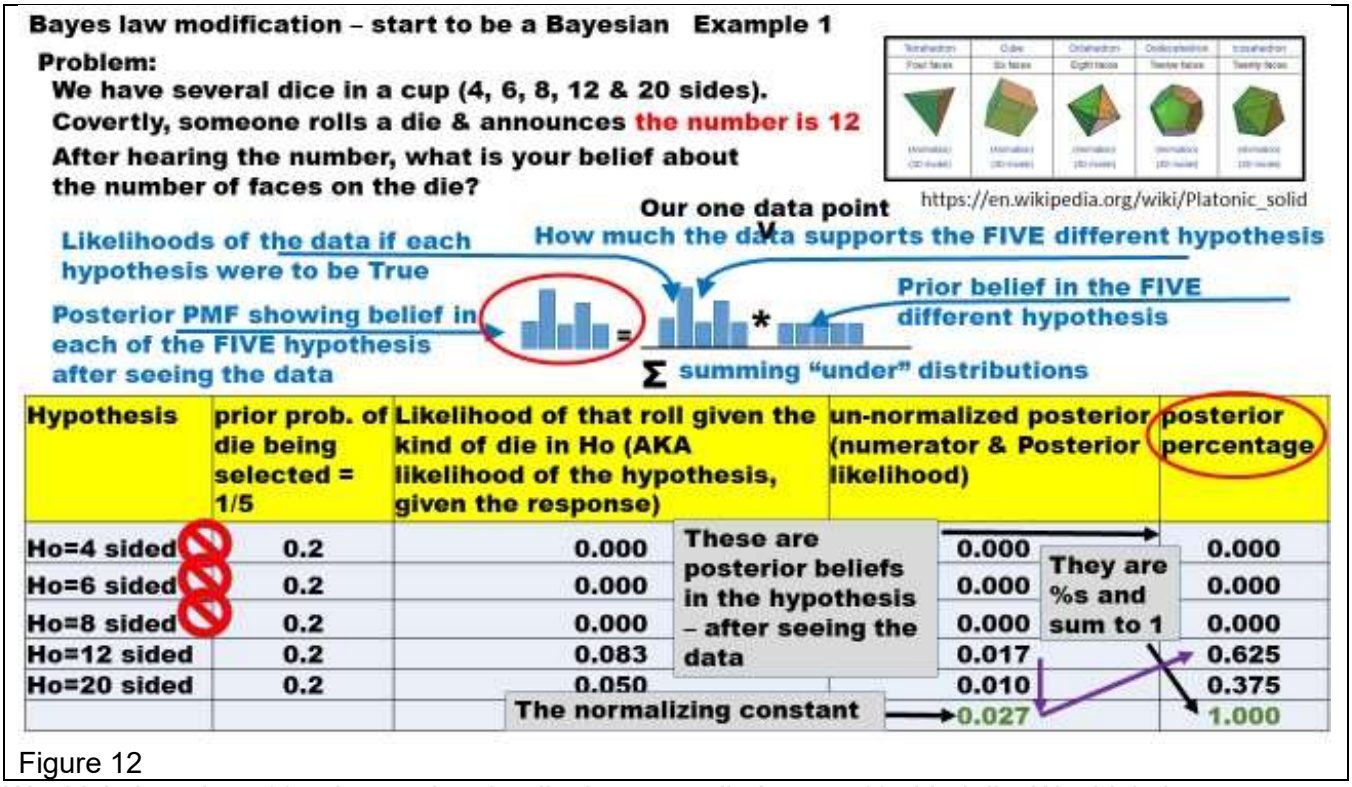


Figure 12

We think there is a .625 chance that the die that was rolled was a 12 sided die. We think there was a .375 probability that the die that was rolled was a 20 sided die.

The 12 sided die only has 12 possible outcomes. The 20 sided die has 20 possible outcomes. If we roll a number that can be produced by either of the two die it's more likely to have come from the die with a fewer number of possible outcomes. That's the end of this example and will do a few more.

Example 2 (exploring the likelihoods)

Example 2, shown in Figure 13, uses the same "set up" but the difference is that the number rolled is a 4.

A 4 sided die has a 25% chance of rolling a 1, 2, 3 or 4. A 6 sided die has 6 possible outcomes and has a lower likelihood for each of those possible outcomes.

The same logic applies to 8, 12 and 20 sided die.

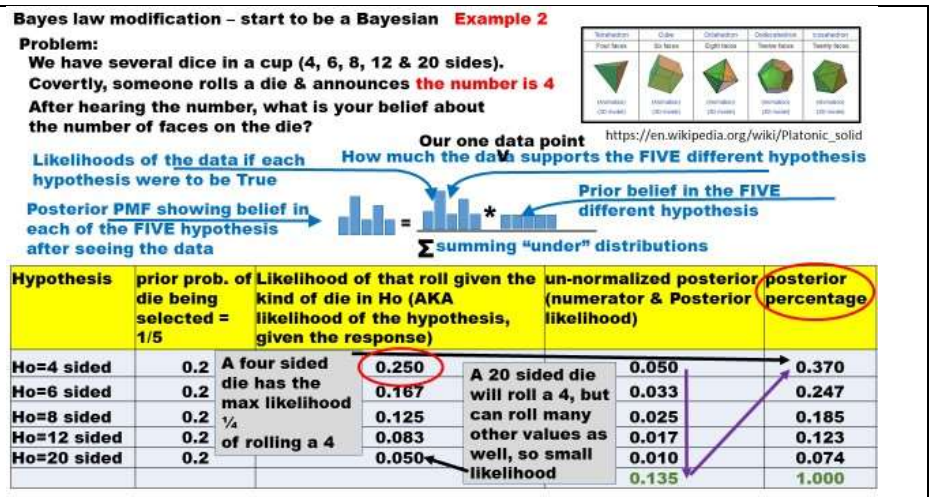


Figure 13

It is suggested that a reader might work through the calculations on this page. The steps in the table can be mapped directly to the graphic that illustrates the Bayes' Law formula.

Example 3 (exploring the likelihoods)

In Figure 14, we add a little complexity to the likelihoods.

We roll two die and see that the numbers are a 4 and a 7. If these are independent rolls then we can multiply the probability of each roll to get the probability of the two different rolls.

If the 4 sided die is picked the probability of a 4 is .25. If the 4 sided die is picked the probability of a 7 is 0.

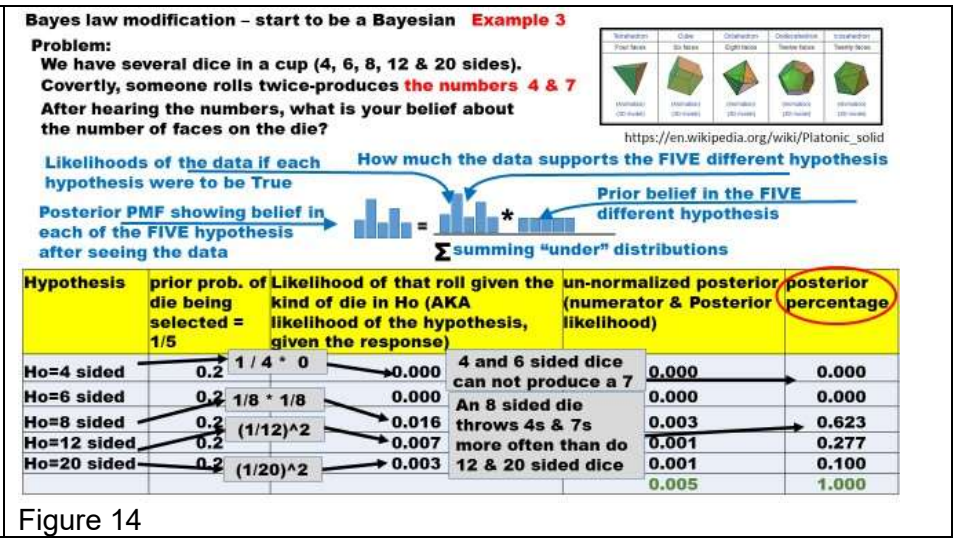


Figure 14

The probability of a 4 and a 7 is the product of those two probabilities and is 0. This logic can be applied to calculate the likelihoods for all of the other die. After calculating the normalizing constant the un-normalized posterior numbers can be changed into posterior probabilities by dividing by the normalizing constant and we can then talk about the chance of each of the hypotheses being true.

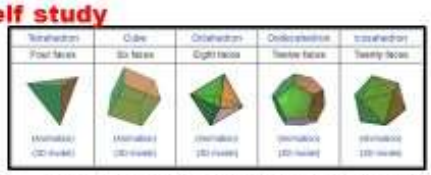
This particular normalizing constant is very easy to calculate, but it *is* an extra step and the relative numbers between the two right columns in Figure 14 are the same except for a multiplicative constant. If we wanted to save a little time we could make excellent decisions without having to calculate the posterior as a percentage.

Example 4 (exploring the likelihoods)

Figure 15 is included for self-study. It is hoped that a reader might reproduce the numbers under the posterior percentage column and record any thoughts about what they see. One thing to consider is why the posterior for the hypothesis that the die had 12 sides is so much larger than the posterior probability that the die had 20 sides.

Bayes law modification – start to be a Bayesian **Example 4- self study**

Problem:
We have several dice in a cup (4, 6, 8, 12 & 20 sides).
Covertly, rolls twice produces the numbers 11 & 12
After hearing the number, what is your belief about the number of faces on the die?



https://en.wikipedia.org/wiki/Platonic_solid

Our one data point
Likelihoods of the data if each hypothesis were to be True
How much the data supports the FIVE different hypothesis
Prior belief in the FIVE different hypothesis
Posterior PMF showing belief in each of the FIVE hypothesis after seeing the data
Σ summing "under" distributions

Hypothesis	prior prob. of die being selected = 1/5	Likelihood of that roll given the kind of die in Ho (AKA likelihood of the hypothesis, given the response)	un-normalized posterior (numerator & Posterior likelihood)	posterior percentage
Ho=4 sided	0.2			0.000
Ho=6 sided	0.2			0.000
Ho=8 sided	0.2			0.000
Ho=12 sided	0.2			0.735
Ho=20 sided	0.2			0.265
				1.000

Your thoughts:

Your comment:

Figure 15

The stated goal for this section was to show some examples that were half way between the way a frequentist would use Bayes' Law and the way a Bayesian would use Bayes' Law. In line with that goal we have a few more examples that bring distributions into the calculations.

Example 5 (Bringing formulas into the likelihoods)

Figure 16 changes the story. Assume we are working in a factory that has three machines that all make the same product. One machine is old and is scheduled for repair in the next fiscal quarter. However; right now it is producing 18% bad parts. We also have a well-maintained machine that produces 10% bad parts. We also have a new computerized machine that produces 1% bad parts.

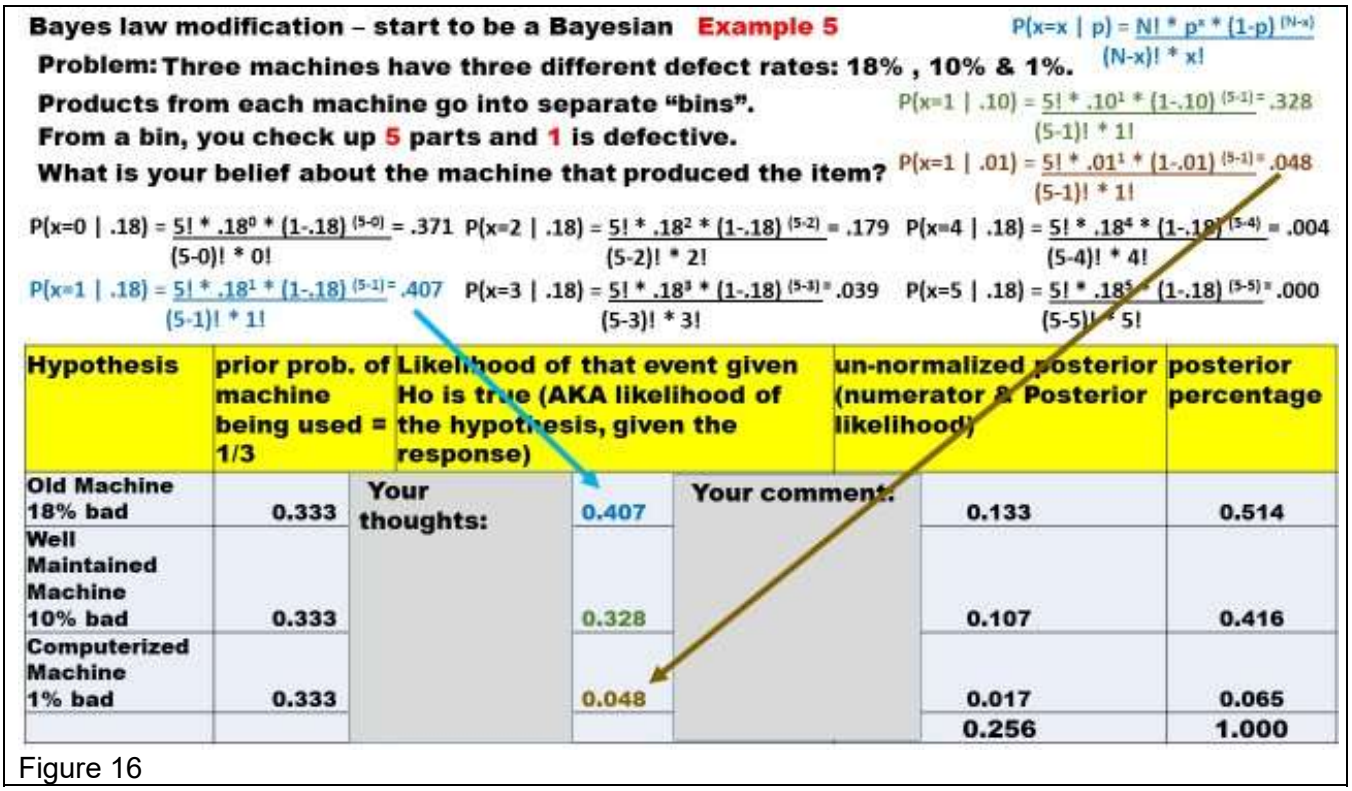


Figure 16

Imagine as parts are produced they get put into a bin so that if you find a bin you can be sure that all the parts and that bin came from one machine. Also assume that machines produce parts at the same rate so each machine should have the same number of bins randomly distributed throughout the factory.

We have three hypotheses. The three hypotheses are that the part came from the 18% machine, the 10% machine or the 1% machine. Assume that you go to a bin, and pick up five parts, and find that one of them is defective.

What is the probability that that bin was produced by each of the three machines?

We're going to say the priors are all .33. We have no idea, from looking at a bin, what machine produced the parts in the bin we just sampled.

We use the binomial formula (see upper right corner of Figure 16) to calculate the likelihoods. Figure 16 shows all the possible calculations for the 18% machine and just the calculations that were used in the table for the 10% bad and 1% bad machines.

Here is the important concept. If the Ho that the parts came from the 18% machine is true, the likelihood of one bad in a sample of five (the likelihood of our data given that Ho is true) is .407.

If the Ho of the parts coming from the 10% machine is true the likelihood of the data we saw (one bad out of five) is .328.

If the Ho of the parts coming from the 1% machine is true the likelihood of the data we saw (one bad out of five) is .048.

It is hoped that the reader might reproduce the numbers in the figure above.

Example 6 (Bringing formulas into the likelihoods)

Figure 17 reuses the same story about 3 machines in a factory and changes the number of defects to three in five. It is hoped that the reader will work through the calculations and make comments about what they notice.

We were able to solve this problem because we had three machines and a simple likelihood formula. When dealing with continuous distributions the calculations will be more complicated.

Bayes law modification – start to be a Bayesian **Example 6**

$$P(x=x \mid p) = \frac{N! * p^x * (1-p)^{(N-x)}}{(N-x)! * x!}$$

Problem: Three machines have three different defect rates: 18% , 10% & 1%.

Products from each machine go into separate "bins".

From a bin, you check up 5 parts and 3 are defective.

What is your belief about the machine that produced the item?

$$P(x=3 \mid .10) = \frac{5! * .10^3 * (1-.10)^{(5-3)}}{(5-3)! * 3!} =$$

$$P(x=3 \mid .01) = \frac{5! * .01^1 * (1-.01)^{(5-3)}}{(5-3)! * 3!} =$$

$$P(x=0 \mid .18) = \frac{5! * .18^0 * (1-.18)^{(5-0)}}{(5-0)! * 0!} = .371 \quad P(x=2 \mid .18) = \frac{5! * .18^2 * (1-.18)^{(5-2)}}{(5-2)! * 2!} = .179 \quad P(x=4 \mid .18) = \frac{5! * .18^4 * (1-.18)^{(5-4)}}{(5-4)! * 4!} = .004$$

$$P(x=1 \mid .18) = \frac{5! * .18^1 * (1-.18)^{(5-1)}}{(5-1)! * 1!} = .407 \quad P(x=3 \mid .18) = \frac{5! * .18^3 * (1-.18)^{(5-3)}}{(5-3)! * 3!} = .039 \quad P(x=5 \mid .18) = \frac{5! * .18^5 * (1-.18)^{(5-5)}}{(5-5)! * 5!} = .000$$

Hypothesis	prior prob. of machine being used = 1/3	Likelihood of that event given Ho is true (AKA likelihood of the hypothesis, given the response)		un-normalized posterior (numerator & Posterior likelihood)	posterior percentage
Old Machine 18% bad	0.333	Please fill in the missing numbers:		0.133	
Well Maintained Machine 10% bad	0.333				.170
Computerized Machine 1% bad	0.333		0.000		
				0.16	1.000

Your comment:

Figure 17

Example 7 (Priors and formulas into the calculations)

Figure 18 is a slight modification of Figure 17. In this example we say that the machines do not produce at the same rate. The computerized machine produces 50% of the parts found in the factory. The well-maintained machine produces 35% of the parts. The old machine produces 15% of the parts. It is hoped that the reader will work the calculations and pencil in their insights from the process.

Bayes law modification – start to be a Bayesian Example 7

$$P(x=x \mid p) = \frac{N! * p^x * (1-p)^{(N-x)}}{(N-x)! * x!}$$

Problem: Three machines have three different defect rates: 18% , 10% & 1%.

Products from each machine go into separate "bins".

From a bin, you check up 5 parts and 3 are defective.

What is your belief about the machine that produced the item?

$$P(x=3 \mid .10) = \frac{5! * .10^3 * (1-.10)^{(5-3)}}{(5-3)! * 3!} = .$$

$$P(x=3 \mid .01) = \frac{5! * .01^1 * (1-.01)^{(5-3)}}{(5-3)! * 3!} = .$$

$$P(x=0 \mid .18) = \frac{5! * .18^0 * (1-.18)^{(5-0)}}{(5-0)! * 0!} = .371 \quad P(x=2 \mid .18) = \frac{5! * .18^2 * (1-.18)^{(5-2)}}{(5-2)! * 2!} = .179 \quad P(x=4 \mid .18) = \frac{5! * .18^4 * (1-.18)^{(5-4)}}{(5-4)! * 4!} = .004$$

$$P(x=1 \mid .18) = \frac{5! * .18^1 * (1-.18)^{(5-1)}}{(5-1)! * 1!} = .407 \quad P(x=3 \mid .18) = \frac{5! * .18^3 * (1-.18)^{(5-3)}}{(5-3)! * 3!} = .039 \quad P(x=5 \mid .18) = \frac{5! * .18^5 * (1-.18)^{(5-5)}}{(5-5)! * 5!} = .000$$

Hypothesis Machines now produce different numbers of parts	prior prob. of machine being used = 1/3	Likelihood of that event given Ho is true (AKA likelihood of the hypothesis, given the response)		un-normalized posterior (numerator & Posterior likelihood)	posterior percentage
Old Machine 18% bad	0.15	Please fill in the missing numbers:	Your comment:	.00585	
Well Maintained Machine 10% bad	0.35				.324
Computerized Machine 1% bad	.500		0.000		
				0.085	1.000

Figure 18

Example 8 (the Monty Hall problem):
This example is intended to illustrate some of the difficulties in calculating likelihoods.

Monty Hall ran this game on TV for years. The problem is as follows. There are 3 doors (A, B and C) behind which are 2 goats and a car. You want to pick the car and you pick a random door (you pick A).

Monty opens another door, door B, and shows you a goat. He then offers you the chance to switch to door C or stay with door A. Which door should you select, A or C.

Bayes law modification – start to be a Bayesian Example 8 The Monty Hall Problem

The formula is not useful unless you know how to interpret the words: *Likelihoods*

Problem: There are 3 doors, behind which are two goats and a car.
You want the car and pick a random door (call it door A).
Monty Hall opens Door B and shows a goat.

He offers you the chance to switch to door C or continue with door A.

Your Decision: Which door should you select, A or C?

Rules:

You Pick a door. Game manager always offers the chance to switch doors.
Game manager always opens a door you did not pick.
Game manager always opens a door hiding a goat and never the car.

the non-random choice provides information,

Cognitive psychologist Massimo Piattelli-Palmarini said "no other statistical puzzle comes so close to fooling all the people all the time", and "even Nobel physicists systematically give the wrong answer, and that they *insist* on it,
and they are ready to berate in print those who propose the right answer".

In 1990 by Marilyn vos Savant (I.Q. measured at 220 +) published the answer.
Prompted 10,000 letters (many from Ph.D.s) saying she was wrong.

Figure 19

Our goal is to fit this problem into the Bayesian structure we've been using for the last several examples. As a warning, precise statements will be very helpful in getting the correct answer.

We need priors and those are easy and all 1/3. We also need the likelihood of the "new data" given that hypothesized prior is true. It is worth spending a minute to try and be very clear on what is the new data. It's the fact that you have selected door A and door B has been opened.

If the Ho that A hides the car is true: If the car is behind A, then Monty can open either B or C with a 50% chance. If the car is behind A, and you select A, the likelihood of Monty opening door B is 50%. If the car is behind the door A and you select A, the likelihood of the data we saw (Monty's action) is 50%.

If the Ho that B hides the car is true: If the car were behind door B, Monty could never open door B – that would end the game. The likelihood of Monty opening door B (the data we saw), if the car is behind door B, is 0.

If the Ho that C hides the car is true: If you select A and the car is behind C, Monty cannot open either door A or door C. If you select A and the car is behind C, Monty must open B 100% of the time. If you select A and the car is behind C, there is a 100% likelihood of seeing the data we saw.

Bayes law modification – start to be a Bayesian Example 8 The Monty Hall Problem

(Ho: C is true) **How likely is "You select A & Monty selects B" if the car is behind C?**
If you select A and the car is behind C(Ho: C is true), Monty cannot open doors A or C.
If you select A and the car is behind C(Ho: C is true), Monty must open B (100%).
If you select A and the car is behind C (Ho: C is true), there is a 100% likelihood of seeing the data we saw (Monty must open door B).
The car could be behind Door A or Door C. We can now apply Bayes' formula.
Summary: Just knowing the formula, alone, is of little use (unless the problem is trivial).
We need the insights, and English interpretations we developed, to give us any chance of recognizing formula components and solving a problem.

Hypothesis	prior prob. of car behind door = 1/3	Likelihood of that event given Ho is true (AKA likelihood of the hypothesis, given the response)	un-normalized posterior (numerator & Posterior likelihood)	posterior percentage
Door A hides car	1/3	1/2	1/6	1/3
Door B hides car	1/3	0.000	0.000	0.000
Door C hides car	1/3	1.000	1/3 = 2/6	2/3
We needed different logic for the different likelihoods.			3/6	1.000

(1/6) / (3/6)
(2/6) / (3/6)

The problems will get worse (math and logic) when the distributions are continuous.

Figure 20

Example 9 (the mammogram problem)

This example is a repeat of an example done very early in this booklet.

It is only included so that we can take that problem and put it into the structure we've been using to solve the previous problems. I think the structure is powerful.

The Hypotheses are cancer or not. The new data is a test result: positive or negative.

Bayes law modification – start to be a Bayesian Example 9 The Mammogram Problem,

I just want to reproduce this problem in the structure we have been using.
A mammogram has an 87 % chance of detecting cancer if it is present
A mammogram signals a false positive 12% of the time.
About 1% of women between 40 and 60 have breast cancer.
What is the chance that a woman, who has a positive test, has cancer.

$$P(C \mid \text{Positive Test}) = \frac{P(\text{Positive Test} \mid \text{Cancer}) * P(\text{Cancer})}{P(\text{Positive Test})}$$

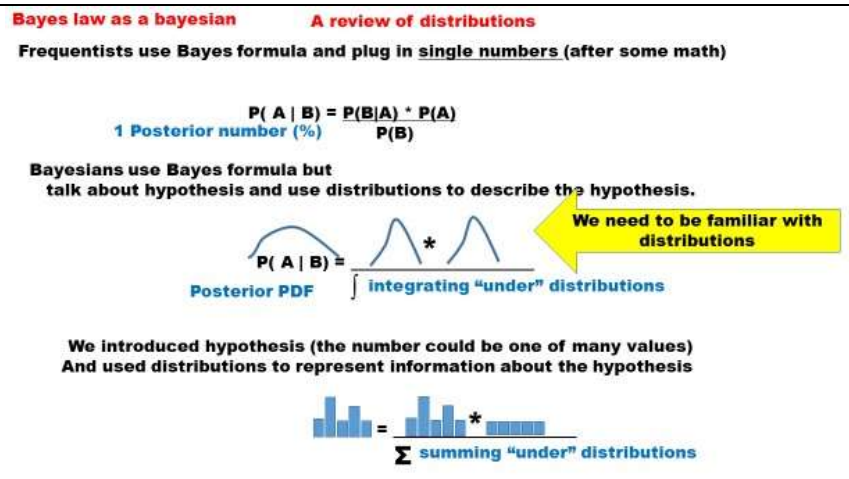
$$P(C \mid \text{Positive Test}) = \frac{.87 * .01}{(.01*.87) + (.99*.12)} = \frac{.0087}{0.0087 + 0.1188} = \frac{.0087}{.1275} = .0682 \text{ or } \sim 7\%$$

Hypothesis	prior prob. of cancer in population	Likelihood of that event given Ho is true (AKA likelihood of the hypothesis, given the response)	un-normalized posterior (numerator & Posterior likelihood)	posterior percentage
Has Cancer	.01	.87	.0087	.0682
No Cancer	.99	0.12	.1188	.9317
			.1275	1.0000

Figure 21

The likelihoods are the probabilities of the data (the test result) given the hypotheses are true. The un-normalized posteriors are the result of multiplications and are numerators in the Bayes' Law equation. The numerator, in the worked out example, is the chance of a positive test and cancer. This example makes it easy to see that the denominator, the normalizing constant, is all the ways we could get a positive test. I leave it to the reader to study this if they choose.

BAYES' LAW AS A BAYESIAN: A REVIEW OF DISTRIBUTIONS

Figure 22 is a transition slide to help us move from one topic to another. It simply reminds us that Bayesian's, when they use Bayes' Law, use distributions as the components of the formula. With that in mind this next section will be a review of the simpler distributions commonly encountered in Bayesian analysis	 Figure 22

There is also a subtle change in the problem being analyzed. In the previous section we were trying to update our relative beliefs in some small number of discrete hypotheses. If we are doing real Bayesian analysis we are going to be analyzing a hypothesis about some characteristic of a distribution – perhaps about the mean.

The result of the Bayesian analysis is a posterior distribution that describes our belief **about the parameters (mean, variance etc.) that describe** a distribution of data. We will be doing a multi-step process and sentences get to be long and complex (and a bit confusing). This is just a warning of what's to come and hopefully things will be clearer when we do some examples.

The multiplication, shown in the graphic in Figure 22, can be difficult. The Exponential family of distributions is important in Bayesian analysis because the only closed form (think simple math) solutions to using Bayes' Law to update our prior beliefs are found using distributions from the Exponential family.

To be a little bit formal: If we assume a parameter η, a distribution is a member of the exponential family

if the density (relative to η) is of the form $\mathbf{p(x|\eta) = h(x)\ exp\ \{\ \eta^T T(x) - A(\eta)\}}$

The members of the exponential family of distributions are:

Normal	Exponential	Gamma	Chi-squared	Beta
Dirichlet	Bernoulli	Categorical	Inverse Wishart	Wishart
Geometric				

the binomial with a fixed number of trials
the multinomial with a fixed number of trials
the negative binomial with a fixed number of failures.

Think again of the graphic in Figure 22 that illustrates the formula for Bayes' Law. There is a term that arises quite often in a discussion of Bayesian analysis and that term is "conjugate prior".

The books say a conjugate prior is a distribution, which when multiplied by the likelihood and then normalized, produces a closed form posterior distribution function which is of the same family as the prior.

The idea for the conjugate prior is that a mathematician started with a formula for a prior, and a type of data that can be used for the updating, and derived a formula that produces a posterior distribution that:
is of the same form as the prior (no one cares about this except the math-ies)
and has a simple formula (we all care about this because it means the answer is easy to calculate).

This means, in simple terms, if your prior is normal and you update it with data you get a normal posterior with a simple formula. Or; if your prior is a beta and you update it with data you get a beta distribution with a simple formula. An example of a conjugate prior and posterior is below.

Coin flips generate a percentage of heads so we will show the conjugate priors for that type of problem. The formulas look similar and are simple to update. Conjugate formulas look like the formulas below.
If the prior formula is $= p^{\alpha} * (1-p)^{\beta}$ and the posterior formula is $= p^{\alpha+k} * (1-p)^{\beta+n-k}$
K is the number of heads in your new data and n-k is the number of tails in your new data.

There are several conjugate priors that have been worked out. A good high–level overview can be found at the wiki page for conjugate priors (https://en.wikipedia.org/wiki/Conjugate_prior. There is a table at the bottom of that webpage that gives a very high–level overview of the known conjugate priors and how you would use these to do a Bayesian update. Deriving formulas can be tough.

Since all conjugate priors are from the Exponential family, let's take a minute to explore links between 3\three members of the exponential family: Poisson, Exponential and Gamma

Poisson: A Poisson distribution can model the number of events occurring in a fixed unit (time period, area, volume). The mean number of events per "unit" is called lambda (λ). For simplicity, assume λ is events per unit time.

Exponential: If the number of events per unit time can be described with a Poisson distribution, the time between any two events can be described using an Exponential distribution. The average time between any two events, for the Exponential, is $1/\lambda_{poisson}$. $1/\lambda_{poisson}$. is often described as the time to the next event (sometimes called Θ.

Gamma: The Gamma distribution is used to model the time between a starting time point and the n^{th} event. If the event you are searching for is the 1^{st} event, the Gamma collapses to the Exponential. The Gamma distribution has two parameters (α and β) and the intuition is that α is the number of events we are waiting for and β is the average time between events or $1/\lambda_{poisson}$.

The formulas for these three distributions are shown to the right. They are very similar. You can see how these formulas are related and imagine, if you're really good at algebra, making relationships between these three formulas might be easier than making relationships between unrelated formulas.	$\text{Poisson} = P(X=k) = \frac{\lambda^k * e^{-\lambda}}{k!}$ $\text{Exponential} = P(X=k) = \frac{\lambda^k * e^{-\lambda k}}{k!}$ $\text{Gamma} = P(X=k) = \frac{\lambda^k * e^{-\lambda k} * x^{k-1}}{\Gamma}$ Figure 23

A CONCEPTUAL REVIEW OF THE BERNOULI DISTRIBUTION

The Bernoulli distribution is a description of the outcome of a random variable which can only take the values of one (with probability equals P) and is 0 (with probability equals 1 minus P).

An example of this is one coin toss, or one at-bat in baseball, or one free throw in basketball. The Bernoulli distribution is the basis for the binomial distribution.

A CONCEPTUAL REVIEW OF THE BINOMIAL DISTRIBUTION

The binomial distribution is often used to model the number of successes, in a sample of size n, drawn with replacement from a group much larger than n.
The formula can best be understood using a decision tree like the one on the right-hand side of Figure 24.

$P^k * (1-P)^{n-k}$ is the probability associated with any one path through the decision tree with K successes and (n - k) failures

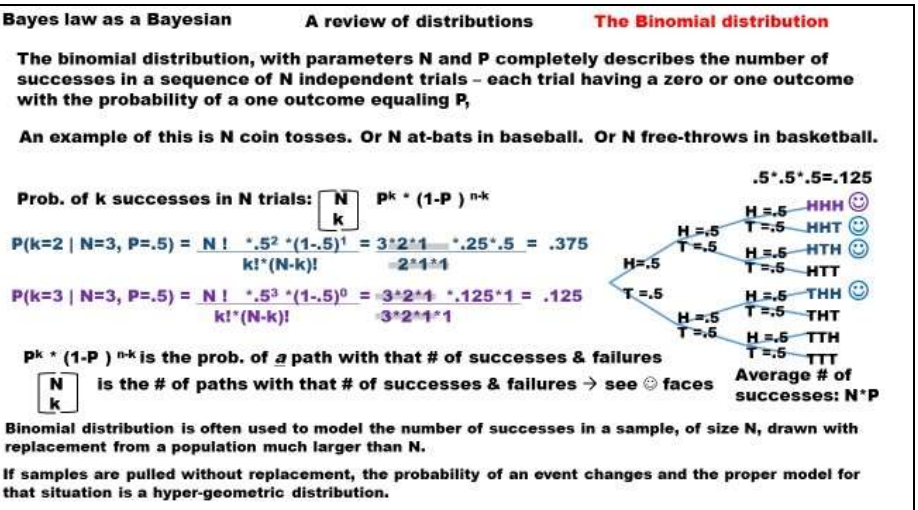


Figure 24

The factorial part of the calculation is used to calculate the number of different paths through the decision tree that have k successes and (n-k) failures.

The calculation of the factorials, for small problems, is simple because cancellations can make the formula simple. However, when the decision tree gets to be large, and one is doing the calculation by hand, the calculation the factorial can be quite a problem.

Because a lot of work using the binomial distribution was done in the days before computers, and even calculators, there are several formulas for approximations to the binomial that are simpler to calculate by hand. These approximations have become less interesting to practitioners since the advent of powerful and cheap computers. However; there is still are of interest to mathematicians because they allow simpler formulas to be substituted when doing of long and complicated derivation.

A CONCEPTUAL REVIEW OF THE NORMAL DISTRIBUTION

The Normal curve is completely described by the mean and standard deviation. The Normal is an integral part of every stat 101 class.

An interesting thing, in the context of this seminar, is that the formula for the PDF has a normalizing constant. This concept has been mentioned before.

The numerator in the formula (the e to the power) gives the curve its shape.

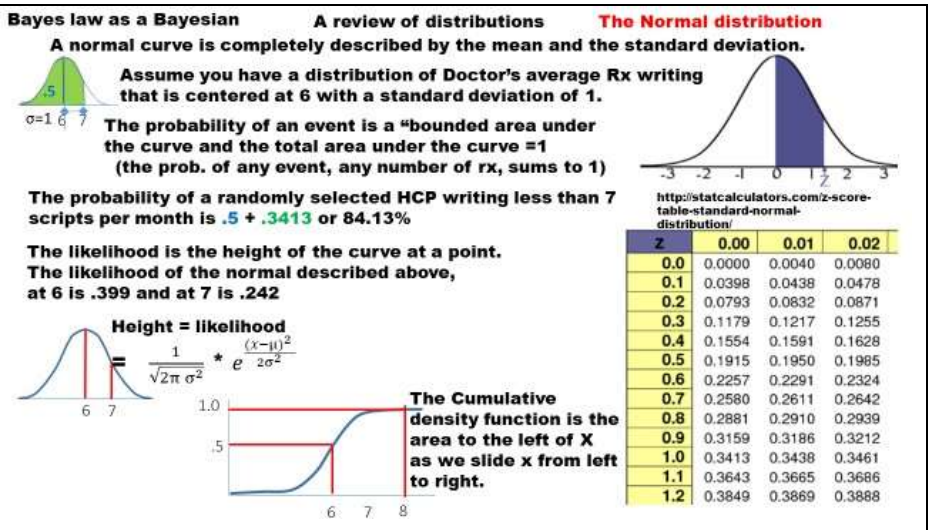


Figure 25

The denominator is a normalizing constant and makes the area under the curve integrate to one. If the area under the curve integrates to one, we can think of calculated areas as probabilities – because probabilities must sum to one.

A CONCEPTUAL REVIEW OF THE BETA DISTRIBUTION

Figure 26 shows graphics about the beta distribution.

The x-axis in the plots is a percentage.

Beta PDF =

$$\text{Prob}(X=x|\alpha\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha\beta)}$$

Where B in the denominator is the beta **function**

$$B(\alpha\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt$$

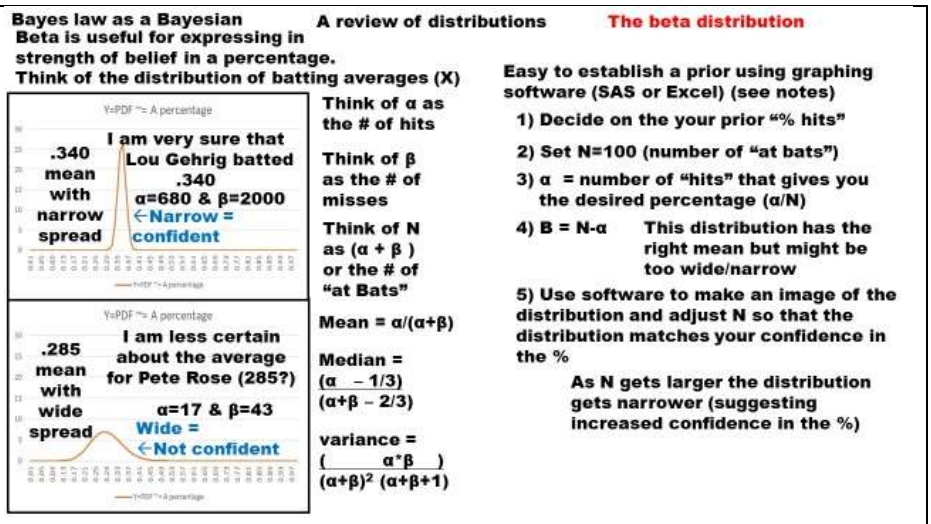


Figure 26

It's very confusing that the word "beta" has three common, very different, uses in this one formula alone..

1st: Beta" is the commonly used name of the PDF & CDF probability ***distributions***
2nd:B(α,β) "Beta" is the name of the ***function*** used in the denominator of the density BETA (PDF) formula
3rd: Beta" is the name of the second ***parameter*** in the density function

People sometimes get lazy and just say beta. If you are familiar with the problem it's not confusing. If you're new to the problem it can be very confusing.

The beta distribution is used to explain to a client, or get from a client, a prior belief about a percentage. Please look at Figure 26 and note that the x-axis is a percentage.

The top chart illustrates a prior belief about Lou Gehrig's batting percentage. I'm pretty sure that the average is .340. I can use a chart of the beta distribution to express my belief about the average and how certain I am in my belief. I think the average is .340 and I'm pretty sure of that. I make my distribution centered at .340 and narrow. I have an xls sheet (and also a SAS program) to make plots of the Beta PDF and allow you to adjust the α and β parameters until the curve has the proper mean and spread – a mean and spread that matches your prior belief about a percentage.

I'm not very sure about Pete Rose's batting average. He was a good player and I'm thinking he's close to .300 lifetime but I'm not very sure at all. It could be higher or lower and I can use a beta distribution to draw a picture that helps me explain, to others and myself, my belief in the average and my confidence about that average. The picture above, with its wide spread, conveys that I'm not very confident about my estimate.

Now that I'm creating this booklet, and really looking at the X axis, I realized that I am more confident than the picture shows. The distribution touches y=0 at about .130 and .450 and I should change that.

I'm pretty sure that Pete Rose was not a .130 hitter because he was famous for his ability to hit the ball. I'm also pretty sure he wasn't a .450 hitter lifetime because no one was a .450 hitter.

The right-hand side of Figure 26 explains how to make the adjustments to a beta distribution to reflect your belief in a percentage.

The 1st step is to decide on the percentage that you think is the center of the distribution.

The 2nd step is to set N= 100 and α to be the percentage you want. α / N will be the center of the beta distribution. We now have the center of the distribution and we need to adjust the width.

Keeping with the idea of a baseball player and using software to create pictures of the distribution:
α (alpha) is the number of hits the player got.
N is the number of at-bats and equals alpha plus beta.
β (beta) is the number of misses or "at bats without a hit" and is N-α.
As N gets larger the distribution gets narrower. To adjust your confidence in your expected percentage you adjust N, up and down, keeping the ratio of α / α+β (or α / N) a constant.

Since I think my spread for the Pete Rose distribution is too wide I will increase the value of N and watch the distribution get "skinnier".

Figure 27 shows output from a SAS program (that will be included with the seminar materials).

It loops over a couple values of alpha and beta and makes plots.

The beta distribution is very flexible.

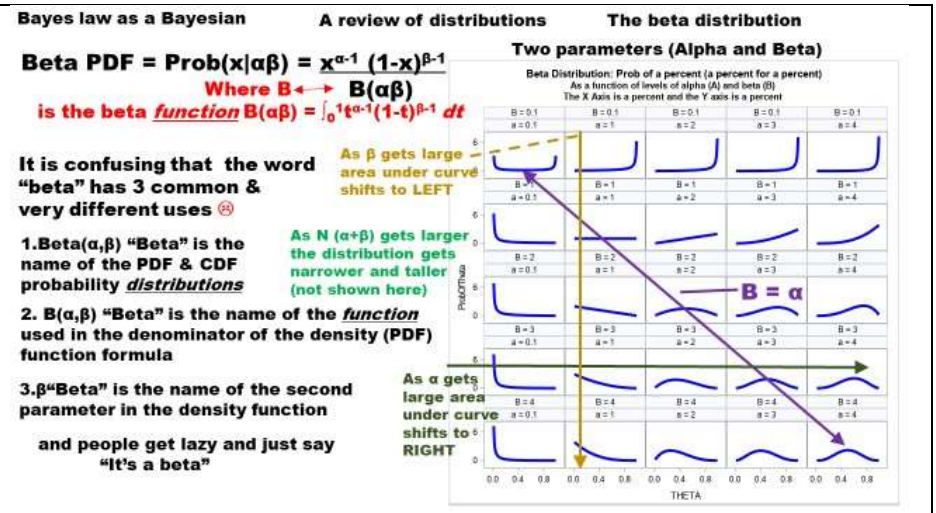


Figure 27

In general, as the beta parameter gets large the area under the curve shifts to the left (missing the ball lowers your %). As the alpha parameter gets large the area under the curve shifts to the right (hitting the ball increases your batting percentage). When alpha and beta have the same value the distribution tends to be symmetric.

There's another thing to learn if we want to use the beta distribution to describe our prior belief (mean and spread) in a percentage. There are times when we have no idea what that percentage might be. We saw this phenomenon, no real idea what the prior parameter should be, in the dice example. We had no idea what the probability of picking any particular dice would be so we assigned them all the same probability.

Figure 28 provides information on how to set parameters for a beta distribution when we have no idea what the proper percentage should be.

One common setting is for a Jeffrey's prior, where alpha and beta both equal .5.

Another common setting is for what is called an uninformative prior and has alpha and beta both equal to 1.

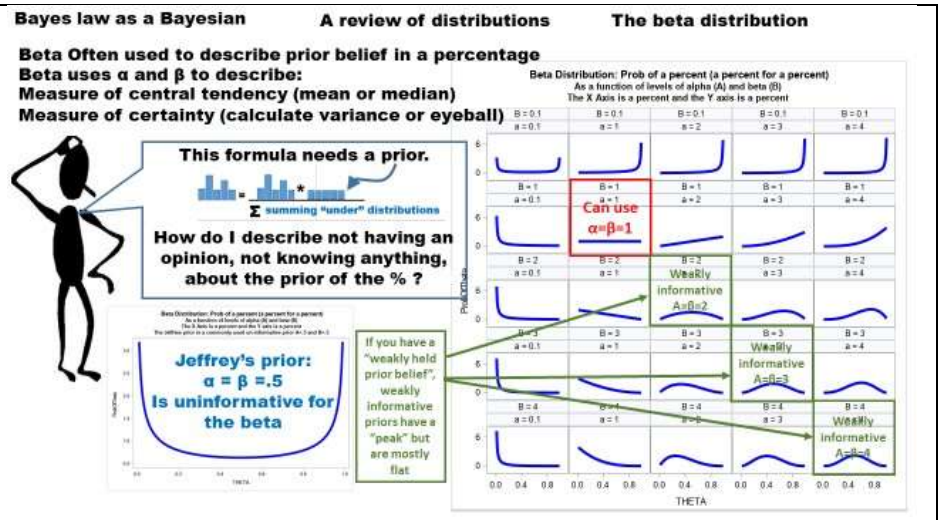


Figure 28

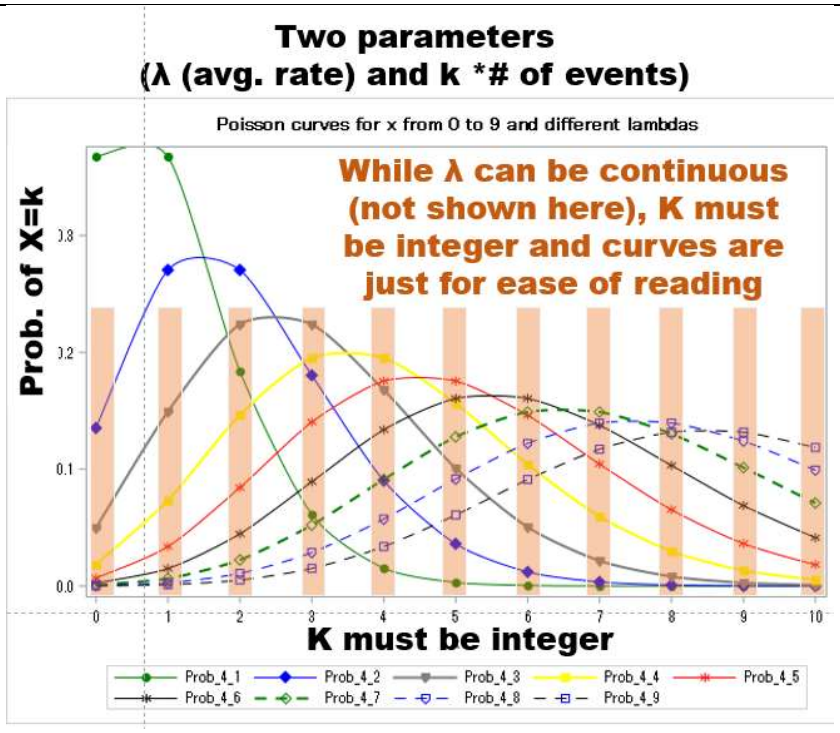
There is one more concept about concerning priors – the weakly informative prior. The weakly informative prior has its center at the percentage we think is most likely but has a wide spread. The distribution in the lower right hand corner of Figure 28 shows a weakly informative prior centered at .5.

A CONCEPTUAL REVIEW OF THE POISSON DISTRIBUTION

The Poisson distribution describes the probability of a number of events, k, occurring in a fixed "interval" of time or area or volume. λ is the average rate and also the variance of the Poisson distribution.

The PDF is:

$P(X=k) = \frac{\lambda^k * e^{-\lambda}}{k!}$ and can be read as the probability of the random variable X having the value k

Since k counts events it must be an integer. λ is the average number of events per unit interval. λ might be the number of cars per hour. If λ is 5 cars per hour it is also 2.5 cars per half-hour. If λ is 18 "metal finish blemishes" in a square yard, λ is also two "metal finish blemishes" in a square foot. λ is often called a rate parameter because it describes the rate at which events occur. While K must be an integer, λ does not have to be an integer. .	 Figure 29

If lambda is small, it's very possible to have 0 events in your interval. If lambda is describing accidents at a particular traffic intersection, and lambda is one per week, it's very likely that it particular week will have 0 accidents. It is not possible to have negative accidents. It is assumed that there is no upper limit to the number of events, though the real world only meets this approximately.

Figure 29 shows continuous curves for all of the different values of λ but this is just to make it easier to read the chart. The fact that k must be integer is emphasized by the dots on the lines showing where k is integer and also by the tan colored vertical bars showing where k is an integer.

Conditions for use of a Poisson are:
Events occur at a known constant rate, λ, which is the average number of events per "interval".
It is assumed that lambda does not change in the "analysis time.
Events are independent of each other and of the time since the last event.
K is an integer and can take on values from 0 to positive infinity. Events are counted in integer numbers. Something to think about, as you remember the binomial distribution, is that in a Poisson problem we can count the number of events but we cannot count, or predict the, the number of "events that did not occur".

Figure 30 shows the Poisson distribution for small, non-integer values.

Note that as λ get smaller the probability of 0 occurrences increases.

Also note that as lambda gets large, and large is only 5 or so, the Poisson approximates the normal distribution.

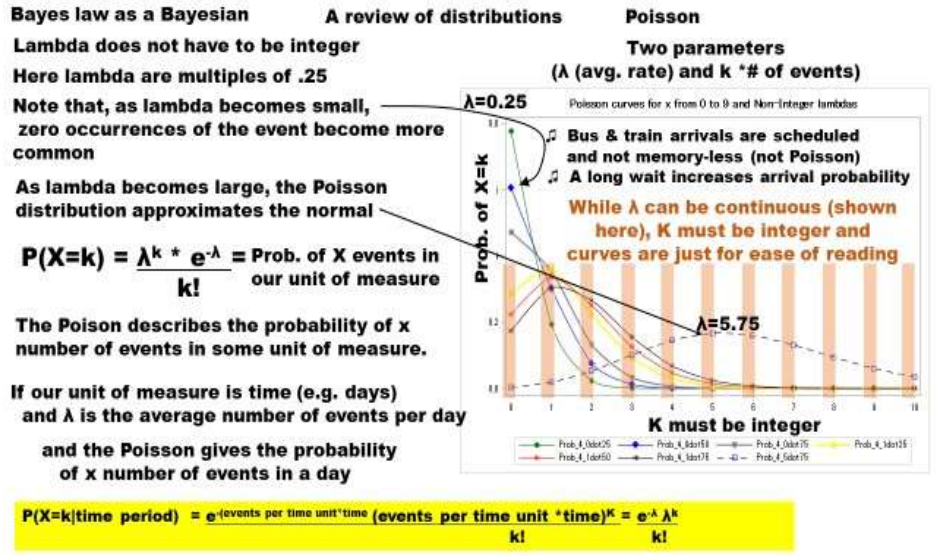


Figure 30

Figure 31 is a bit of a repeat of the information in Figure 30.

It just cleans up the graphic a bit and, hopefully, better shows the effect of changing λ on the distribution.

While the points are connected by smooth curves, the Poisson distribution is only valid for integer values on the x-axis.

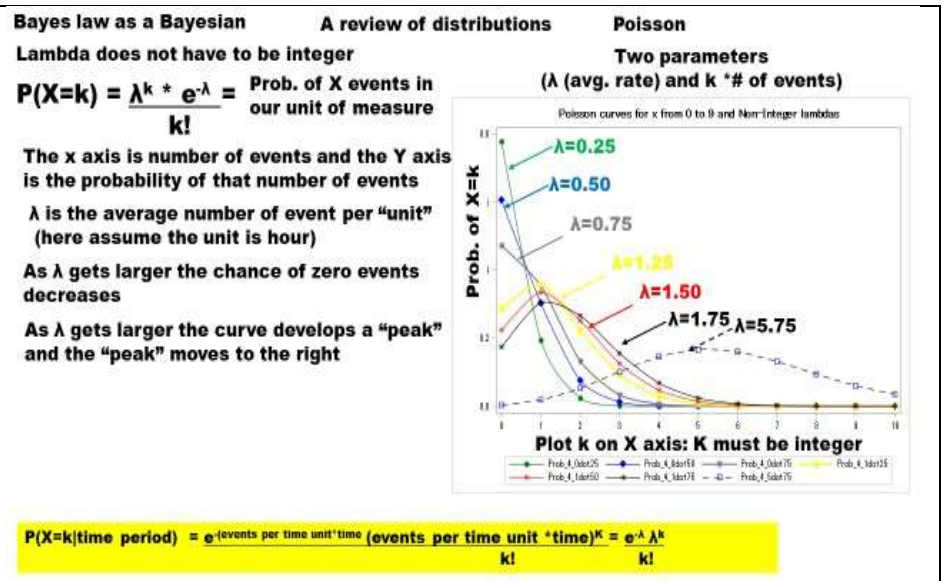


Figure 31

The yellow box, in the bottom of Figure 31, is a small reminder of a complication that can happen with these distributions - unit inconsistency. The parameter, λ, is the number of events in a unit time. It might be the number of calls into your phone center per hour. Occasionally, people try complicate the problem by giving λ in one unit of time and observed data in another unit of time.

An example of this might be you know that you average 8 calls per hour and that is what you consider your λ. When people bring you new data they might say "today, we had 80 calls in an 8 hour day". This is a reasonable way to collect data but not in the same unit of time. You can do Bayesian analysis when people start off not using the same "unit" but you must do a conversion somewhere in your calculations.

CONCEPTUAL REVIEW OF THE EXPONENTIAL DISTRIBUTION

Bayes law as a Bayesian **A review of distributions** **Exponential**

If Poisson gives the probability of x number of events in a day ($\lambda_{poisson}$ is the avg. events/day)
Wait time *until the 1st* event follows an exponential distribution with avg. $\theta_{exponential} = 1/\lambda_{poisson}$
Use of the Poisson distribution often concerns counts ($\lambda_{poisson}$ avg. events / time period)
Exponential distribution often involves time between events ($\theta_{exponential}$ =avg. time period/event)
Bounded by 0 and infinity

Steps to estimate an Exponential
1) Record times
2) Calculate gaps
3) Calculate avg. gap
4) Lambda = 1 / average
5) Exponential PDF = $\lambda * e^{-\lambda * x}$ (when X >= 0, else 0)
6) Exponential Cuml. PDF = $1 - e^{-\lambda * x}$

Events are "Poisson $\lambda_{poisson}$" = 5 events / 10.0 min = . 5 events / min = events/ "unit"
Waits are Exponential $\lambda_{exponential}$=10.0 min / 5 events = 2.0 min wait per event

Sample data set	
Arrival time (Decimal Minutes)	Gap (in Decimal Minutes)
0.50	0.50
0.75	0.25
1.00	0.25
2.00	1.00
10.00	8.00
Exponential λ = Avg. wait/event	2.00
Poisson λ = events / min wait	0.50

use the poisson lambda PDF=$\lambda e^{-\lambda * x}$ CDF=$1 - e^{-\lambda * x}$ 0.50

Time of interest	pdf (hand)	pdf (excel)	Cuml. Prob of an event
0.00	0.4998	0.500	0.000
1.00	0.3033	0.3033	0.393
2.00	0.1839	0.1839	0.632
3.00	0.1116	0.1116	0.777
4.00	0.0677	0.0677	0.865
5.00	0.0410	0.0410	0.918
6.00	0.0249	0.0249	0.950
median = (ln 2)/ λ			1.386294

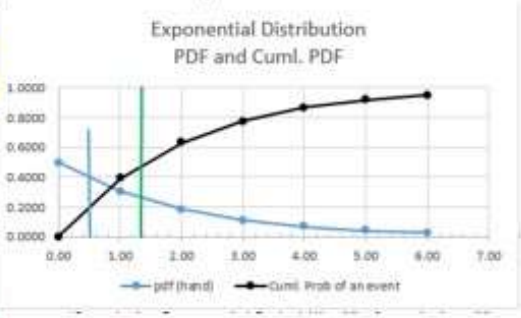


Figure 32

The Exponential distribution also shows up in Bayesian analysis. Members of the Exponential family of distributions differ because of the way the parameters are added and subtracted in the formulas. It is often the case that, if a parameter in one distribution is set to the value one, the parameter is removed from the equation and the equation then describes one of the other members of the Exponential family of distributions.

An interesting bit of logic shows the relationship between the Poisson and the Exponential. The Poisson describes the number of occurrences per unit of time (Unit can be other things as well as time but we will consider time here). Then the Exponential is a description of the length of time between occurrences.

Consider this short derivation. Assume we have a Poisson process and events occur with the rate of λ events per unit of time. Under this assumption there will be λt occurrences in every t units of time.

The Poisson distribution has the formula below.
$P(X=k) = \frac{\lambda^k * e^{-\lambda}}{k!}$ and this can be read as the probability of the random variable X having the value k

If were asking about the probability of zero events in our unit time the formula collapses to $e^{-\lambda t}$. This describes the probability of zero events in t units of time. Another way of thinking of this is that p (x=0) = $e^{-\lambda t}$ is the probability that the time to the first occurrence of an event is greater than t or using a more formal mathematical representation: $P(T > t) = P(X=0 \mid \mu = \lambda t) = e^{-\lambda t}$

We can use the above formula to calculate the probability that an event does **not** occur during t unit of time and that is: $P(T <= t) = 1 - P(x=0 \mid \mu=\lambda t) = 1 - e^{-\lambda t}$

If you take the derivative of this formula with respect to t you get the probability density function of the Exponential distribution which is: PDF(Exponential) = $\lambda e^{-\lambda t}$

This establishes the relationship between the Exponential and the Poisson that is suggested in Figure 23. The Exponential gives the interval of time between any two consecutive arrivals and not just the time to the FIRST event. This is true because the Exponential is memory-less.

The fact that the Poisson and the Exponential distributions are so related, and so commonly used in many different disciplines, can create confusion. Few books seem to bother to relate the two distributions, even though there single parameters are intimately related by being the inverse of each other. Since both of the parameters are something per something they are both often called rate parameters and often formulas show both of these rates as λ (this confused me).

Unfortunately, some disciplines are in the habit of writing the formulas in different ways. Some Exponential formulas require using the λ that you would get from the Poisson. Some formulas require using 1/ $\lambda_{poisson}$ in the Exponential formula. To clarify the issue, this booklet will show a few large scale images showing the relationship, the formulas and the calculations.

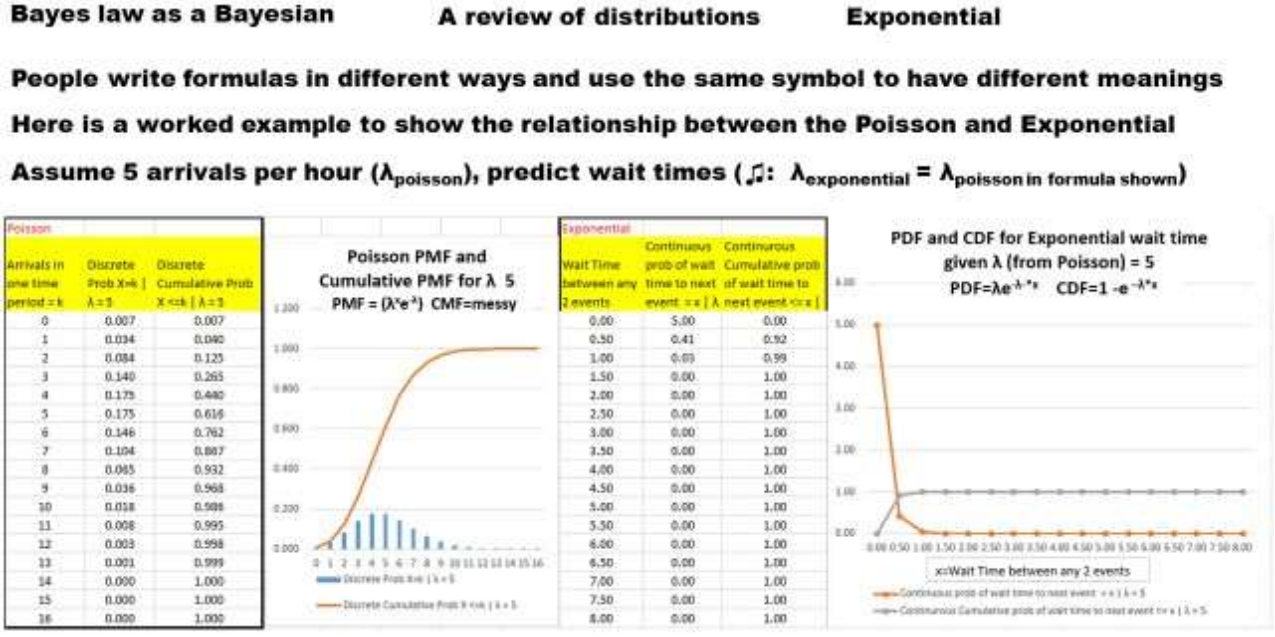


Figure 33

Figure 33 emphasizes the relationship between the Poisson and Exponential. With a Poisson λ = 5, I used Excel to generate the probabilities of counts of events. The left-hand side of the figure shows both the PMF and the CMF (this is a discrete distribution so this is a PMF). For that same λ, 5, we use the formula shown in the right-hand side to calculate the PDF and CDF for an Exponential distribution.

Figure 34, seen below, emphasizes the fact that in the Exponential formula shown in the title of that chart, we use the time gap between events as X and the lambda that is calculated from the Poisson.

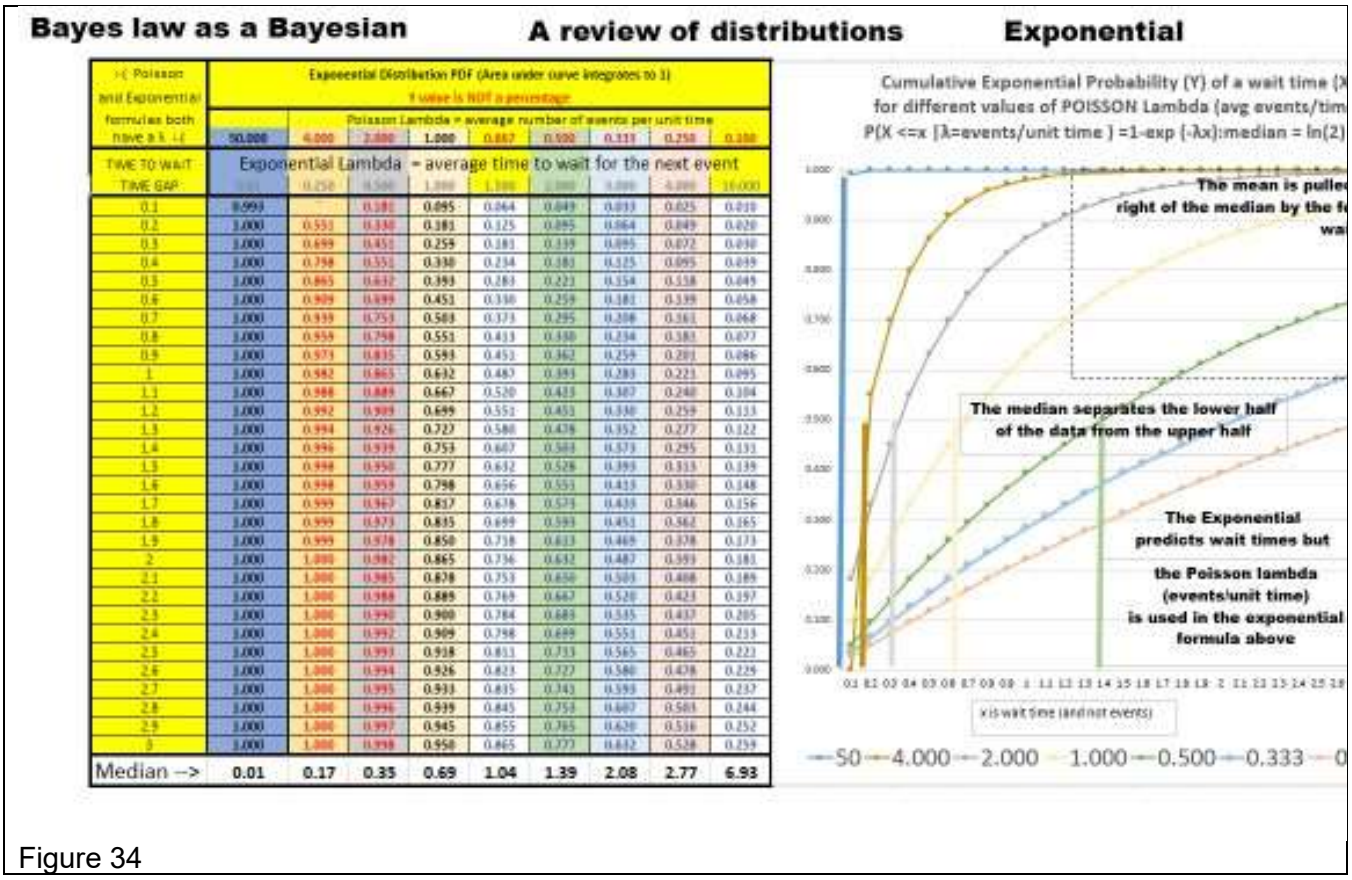


Figure 34

A CONCEPTUAL REVIEW OF THE GAMMA DISTRIBUTION

The Gamma distribution is also important in Bayesian analysis. If the Exponential distribution describes the waiting time to the 1st event the Gamma distribution can be used to describe the waiting time to the nth event.

Using the most common Bayesian formulation, the insight into the Gamma is that the α parameter is the number of events (though alpha can be fractional) we are waiting for.

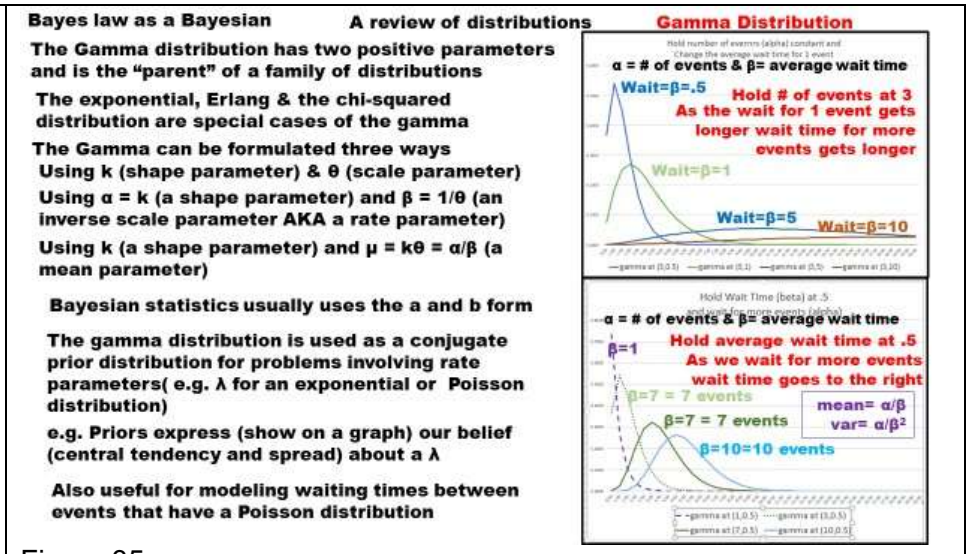


Figure 35

The β parameter is the average wait time for one event.

With that as an understanding (and formulation), it's easy to see that if either α or β gets larger the distribution will shift to the right. If either α (# of events to wait for) or β (average wait time for 1 event) gets larger, an observer must wait longer.

Unfortunately, the Gamma is a bit confusing because it can be parameterized, or formulated, in at least three different ways:

using K as a shape parameter and Θ as a scale parameter
using alpha = K as a shape parameter and β equals 1/Θ as an inverse scale/rate parameter.
using K as a shape parameter and μ=kΘ = α/β as a mean parameter.

…and other ways as well. Thankfully, most discussions in Bayesian statistics use the α and β parameterization.. This is what Excel uses

Some examples of using the Gamma shift from one parameterization to another in the same problem.

Even one Gamma formula is a bit confusing because the word Gamma is reused. The formula for the Gamma distribution has the Gamma function as a component.

$$\text{Gamma} = P(X=k) = \frac{\lambda^k * e^{-\lambda k} * x^{k-1}}{\Gamma}$$

The mean of the Gamma distribution is α/β and the variance is α/β^2 and these can be used to help create a graphic illustrating the distribution that describes a client's belief in some parameter being modeled using the Gamma. The mode is $(\alpha-1)/\beta$ for $\alpha \geq 1$. As β increases the PDF become steeper, since the variance has β^2.in the denominator.

There is quite a tangled relationship between the Gamma and several other distributions The Exponential, the Erlang and the Chi-squared distribution are special cases of the Gamma.

Additionally the Gamma function is a generalization of the factorial process. The factorial process only works for integers but the Gamma function is considered to be a parallel for positive real numbers. If you use integers to create a series of points in X, Y, for the formula y=(x-1)!, you can fit a Gamma function smoothly through all those points. Since the Gamma is valued for non-integer values of X, people consider the Gamma to be a generalization of the factorial.

EXCEL EXAMPLE: A CLOSED FORM SOLUTION FOR UPDATING A PRIOR PERCENTAGE
Figure 36 shows an image of an Excel spreadsheet that is given to seminar attendees. Is interactive and it is intended to mimic the discussion one might have with a client.

It is assumed that the client is a person with a prior knowledge of the percentage and this Excel spreadsheet is intended to help the client express that prior belief in some percentage and then to examine the issue from a few different viewpoints. It is hoped that the interactive nature of this tab on the workbook will allow the reader to become familiar with different parts of this process.

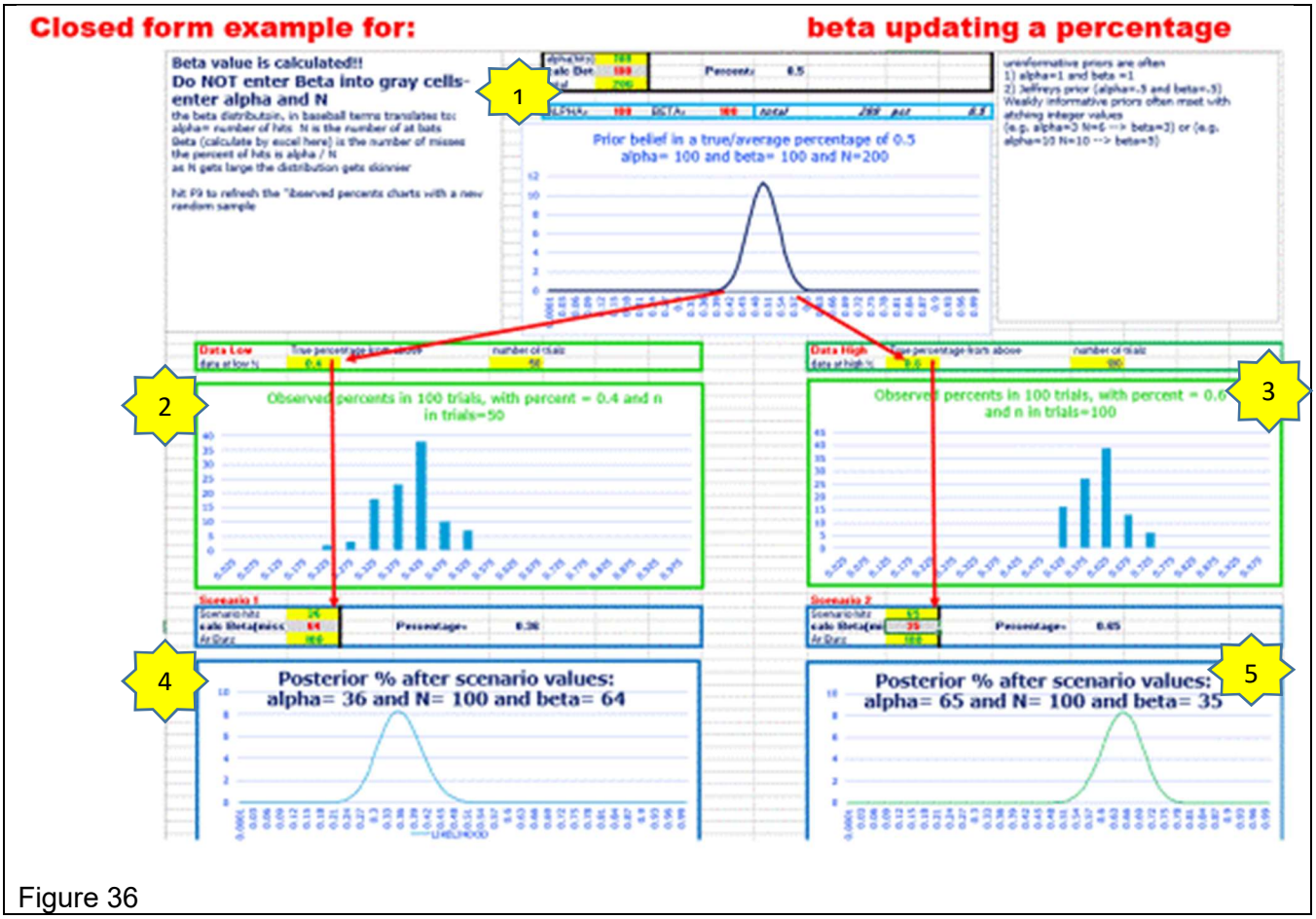


Figure 36

In step 1 we create a picture of a prior distribution of the percentage. We know that the client has some percentage in mind and the easiest way to center this percentage distribution on that number is to enter the percentage number (as integer) as alpha and enter 100 as N. The spreadsheet will automatically calculate beta as N-α.

Remember that, if were thinking in terms of baseball, α is the number of hits. N is the number of at-bats. Therefore; β will be the number of misses and is easily calculated as N-α. That centers the distribution at the proper point.

To make the distribution wider or narrower, mimicking the client's degree of belief in that percentage, one just has to increase/decrease N while keeping the ratio of α/N a constant. Increasing N will make the distribution narrower and decreasing N will make the distribution wider.

At this point we have modeled the clients' belief about a percentage. It is thought that this process would be new to many clients who might be more accustomed to seeing the data rather than some abstracted "distribution about the true percentage that generates the data".

Steps 2 and 3 are very much linked to the prior distribution and are intended to show the client the implications of his idea of the true percentage. It is thought that a client will likely have little experience with mentally imagining data from a distribution of possible percentages and can benefit from adding some concreteness to the problem. It is thought that a manager must often plots his percentage of "conversions per day" as a KPI and might recognize the distributions in step 2 or step 3

In step 2 we take the lowest percentage that the client thinks is possible, the place where the prior distribution touches 0 on the left-hand side, and use that to show the client what the data might look like at that end of his belief system. Step 2 requires that you input the percentage, where the curve in step one touches 0 on the left, into a cell in the Excel SS and Excel will generate random data for that percentage. The thought is that the analyst can show this to the client and say something like "so on your worst month of sales you expect the daily sales percentages to look like this".

In step 3 we take the highest percentage that the client thinks possible and reproduce the steps in step 2. The thought is that the analyst can show this to the client and say something like "so on your best month of sales you expect the daily sales percentages to look like this".

In steps 4 and 5 we allow the client to "see a bit into the future". You can input the clients expected percentage from step one and then say if our "to be collected" data percentage looks like this (might be 20% might be 50%) this is how it will update your belief about the true percentage (of customers who convert on your website). There are two charts here so that the client has the ability to play a little bit of "what if" to see how sensitive the update is to things that he could imagine might happen.

A CONCEPTUAL REVIEW OF THE MCMC ALGORITHM

As has been said before, the calculations required to do a Bayesian analysis are so complicated that they held back progress in the field until computers became powerful enough, and algorithms became good enough, so that the problems could be solved and produce practical results.

The only closed form solutions for Bayesian problems come from the Exponential family and require that someone derive a conjugate prior for a certain distribution. This is extremely difficult to do and has not been done many times.

The solution for this difficulty is to stimulate the Bayesian process and this can be done with many different software packages and many different algorithms. The MCMC algorithm is a classic and worth studying to get the concepts behind using computers and simulation to solve a Bayesian problem.

Figure 37 Is a link between what we've done and what we will do.

The denominator in the examples we did with discrete tables had numbers on the margins of the table. The marginal numbers represented our prior belief in something without having looked at any additional data.

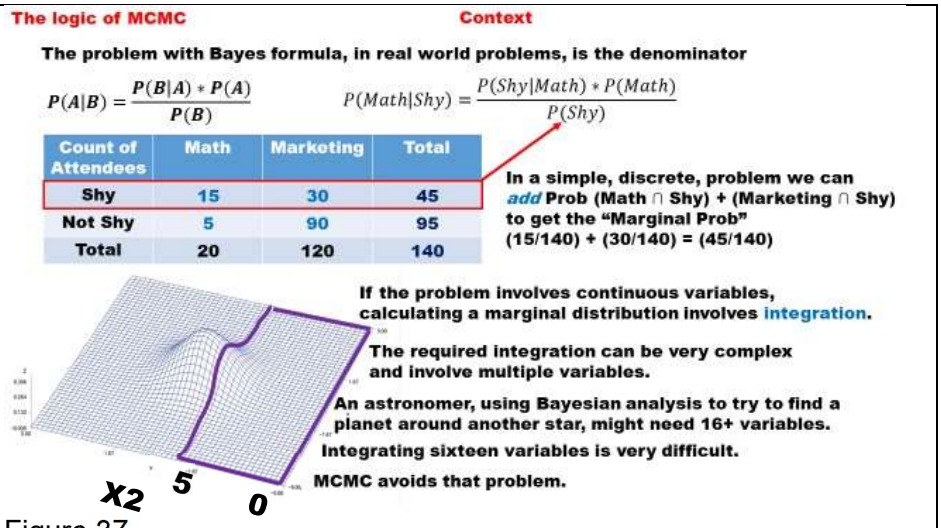


Figure 37

When working with a two dimensional table calculating the marginals is easy – even when they're not on the margin as they were with the smoking example. When one moves to a continuous distribution the summing up of cells has to be replaced by an integration process as is shown in the figure above. In the 2x2 table examples we summed the cells to get a denominator (the total number of people who were shy.) and threw away the rest of the table. Figure 37 shows an analogous process for a continuous variable. While I made up the need to be interested in between 0 and 5, the picture shows us throwing away "information and concentrating on only part of the data". We see a picture of a continuous problem where we might be saying "given X2 is <= 5".

Astronomers do use Bayesian analysis to try and find planets rotating around distant stars in the denominator in that calculation can have 16 to 20 different variables. That integration is difficult if not impossible. Fortunately there is a solution.

Figure 38 links us into something we've seen before. We've all use the normal distribution in one of our stat classes the formula is easy and is something divided by a constant.

That constant is called a normalizing constant and makes the area under the curve integrate to one and use area under the curve's as probabilities of an event happening because probabilities must sum to one.

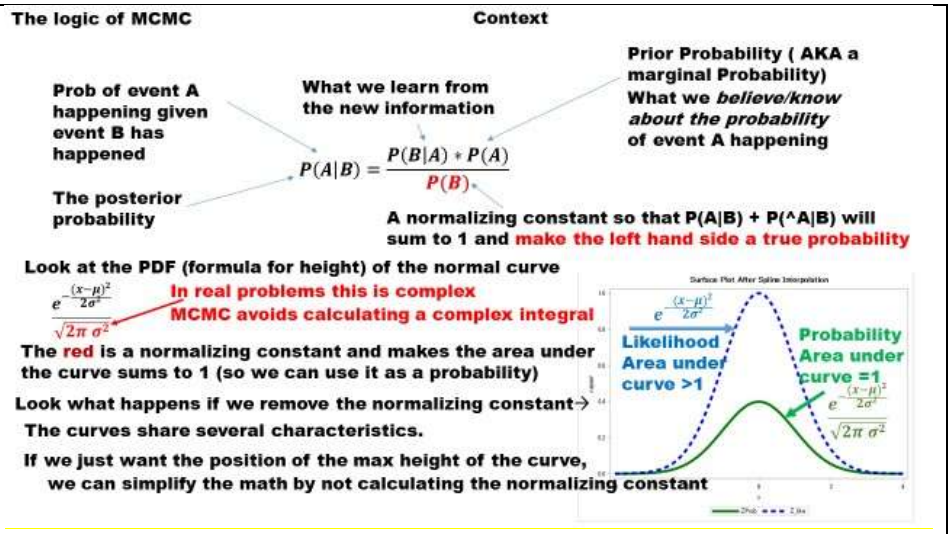


Figure 38

However; if we are just interested in relative probabilities, we can often get answers from a curve whose area does not integrate to one. If were only interested in relative probabilities, and not absolute probabilities, we can dispense with the calculation of the normalizing constant and this can make the calculations (and our life) a bit easier. Many real-world Bayesian calculations dispense with the calculation of the normalizing constant.

Figure 39 lists the steps in the MCMC algorithm for the Metropolis Hastings algorithm.

The goal is to find the highest point in some "accuracy/likelihood" space. The space is defined by the parameters that describe your distribution.
If one were trying to do a Bayesian simple regression the space defined by the two parameters for that regression are βo and β1 as is seen to the right.

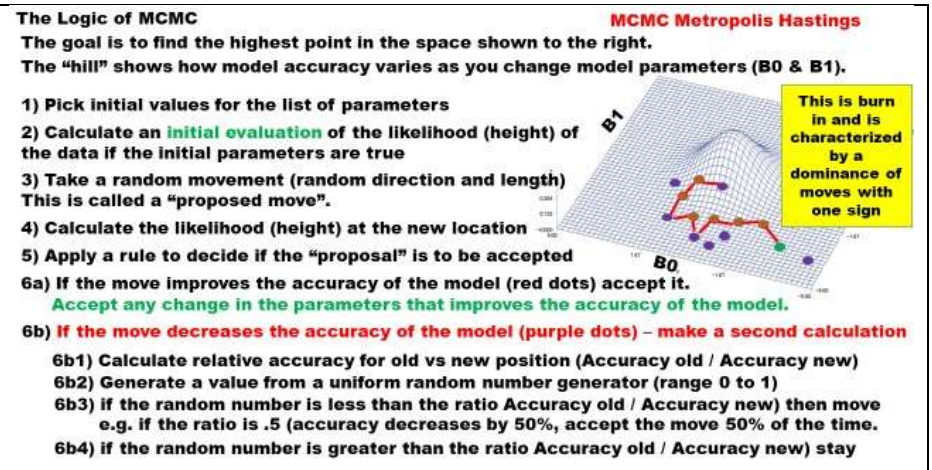


Figure 39

If we can create a distribution that is "highest" at the point where βo and β1 have maximum accuracy we can climb that mountain to find a solution to our problem using the algorithm below.

The Metropolis Hastings algorithm has the following steps:

1) Pick random initial values for your list of parameters (obviously, a good starting point helps a lot).
2) Calculate some initial evaluation of the likelihood of the data if these parameters are true.
3) Propose taking a random jump in βo and β1 (use a bivariate normal to generate length and direction)
4) Calculate the likelihood (likelihood is something like accuracy) of the data at this new point
5) If the model is better at the new point, then make that move.
6) If the model is not better at the new point perform additional calculations

6a) Calculate the ratio of $likelihood_{new}$/ by $likelihood_{old}$.
6b) We are willing to go in the "wrong direction" so that we can explore the space.
6c) Exploring the space keeps us from getting trapped in local minimum.
6d) Generate a random number.
6e) If the random number is smaller than the ratio, accept the jump and move; else stay where you are.

If the jump would put us in a position that's a little bit worse than where we are we should be willing to make that jump with great frequency – just so that we can explore the parameter space. If the move would put us in a position that's much worse than where we are we should be reluctant to make that jump but should occasionally make those jumps to explore the parameter space.

Rejecting jumps keeps us in one place and keeps us from exploring the space quickly. Accepting jumps can move us far in the correct, or in the incorrect, direction.

The picture in the figure above shows a pattern of jumps through parameter space.
It should be noted that, at the beginning, the algorithm will largely be climbing up the hill and the direction will largely be upward.

SAS automatically does some "tuning" of the process of accepting/rejecting "a jump that puts us in a less advantageous place" and does this tuning as a way to reduce run time. Simulations have shown that the algorithm is relatively insensitive to the percentage of jumps that are accepted and rejected but SAs still adjusts. SAS is not just doing the MCMC- Metropolis Hastings algorithm for us, it is removing some of the onerous and time-consuming sub-tasks that used to be performed by hand.

As the algorithm is climbing up the hill the preponderance of the jumps will be in one direction - upwards.

If we plotted the size of the jumps as the algorithm is searching for the hill we would see it moving in one direction – or tending to move in one direction.

The normal curve will tend to make the jumps small. This means that many jumps will be accepted but the hill climb will be slow.

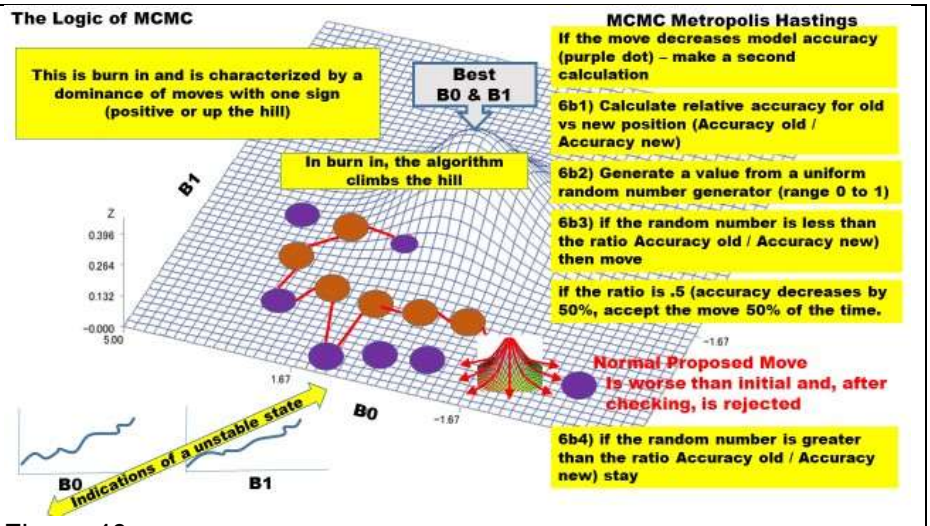


Figure 40

Large jumps explore the space quickly but can actually jump over a high point.

Figure 41 Shows what happens when the algorithm reaches the high point. SAS will tend to jump over and around the maximum point.

Metropolis Hastings, unlike some of the gradient descent algorithms, does not converge to an ideal point.

It converges to an area in parameter space.

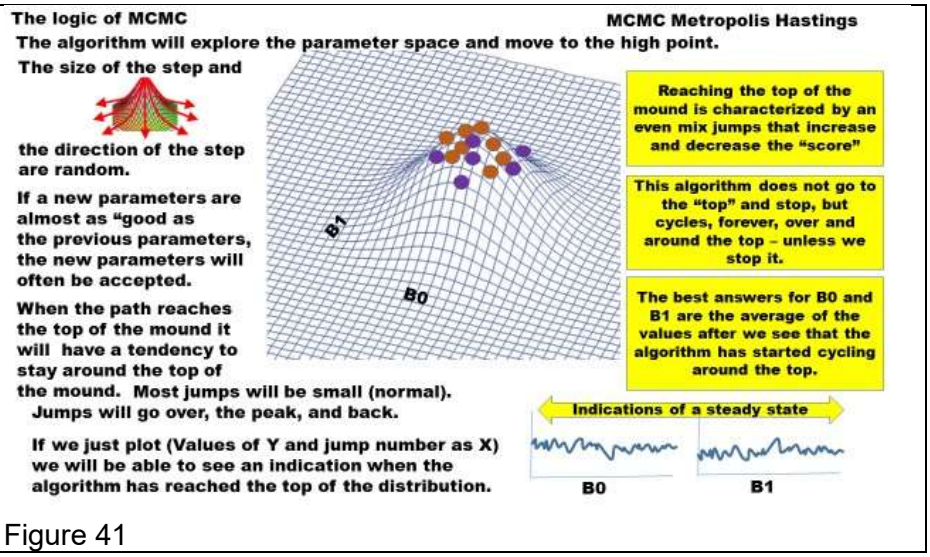


Figure 41

When it has reached a high point a plot of the jump direction and size will look like what can be seen in the bottom right-hand corner of this figure. Plots of the jump direction and size are the best indication of convergence.

Both β0 and β1 must converge, and both must show trace plots similar to what you can see in the picture above. If the parameter space is more than two dimensions, all trace plots must indicate convergence.

The Metropolis Hastings is only one of many algorithms that can be used in Bayesian Analysis. To take advantage of the increased computing power of multi-core chips, another algorithm is the Metropolis Coupled Markov Chain Monte Carlo or MCMCMC. It achieves the goal of exploring parameter space by having multiple "searcher chains" searching in parallel. At the end of every jump, an evaluation is made as to which searcher is in the "best position" and the searcher chain with the best score is allowed to write to the output file.

Selecting the search algorithm is quite a complicated subject and beyond the scope of this seminar. SAS, because it is trying to remove onerous tasks from a user, has very good defaults and will shift from one algorithm/set of settings to another without any input from the user.

Effective Sample Size

Effective sample size is often discussed as a way of measuring how well the Markov chain is "mixing". $ESS=n1 + 2\sum(n-1)k = 1\rho k(\theta)$ where N is the total sample size (number of "jumps") after burning is finished and $\rho k(\theta)$ is the autocorrelation of leg K for θ.

The closer ESS is to N the better the mixing, though people think that in ESS of around 1,000 is sufficient for estimating a posterior density. If you're looking to estimate the tails of the posterior density plots with some precision, you might want to have more than that 1000. ESS is not a significance test but it's more of a criteria that you can use to figure out whether it's reasonable to use the estimate on your printout. .

A CONCEPTUAL REVIEW OF GIBBS SAMPLING

While it's never too wrong to just take the defaults that SAS has set, it is worthwhile learning about another algorithm – because SAS uses it and because many people discuss it. That algorithm is Gibbs sampling and it is implemented in SAS.

Figure 42 starts an explanation of Gibbs sampling.

MCMC is very efficient if the distribution you're selecting from is the posterior distribution – but it's hard to tell what that posterior distribution is until the algorithm converges. Gibbs, like the Metropolis Hastings is trying to jump through parameter space to find a high point of a distribution.

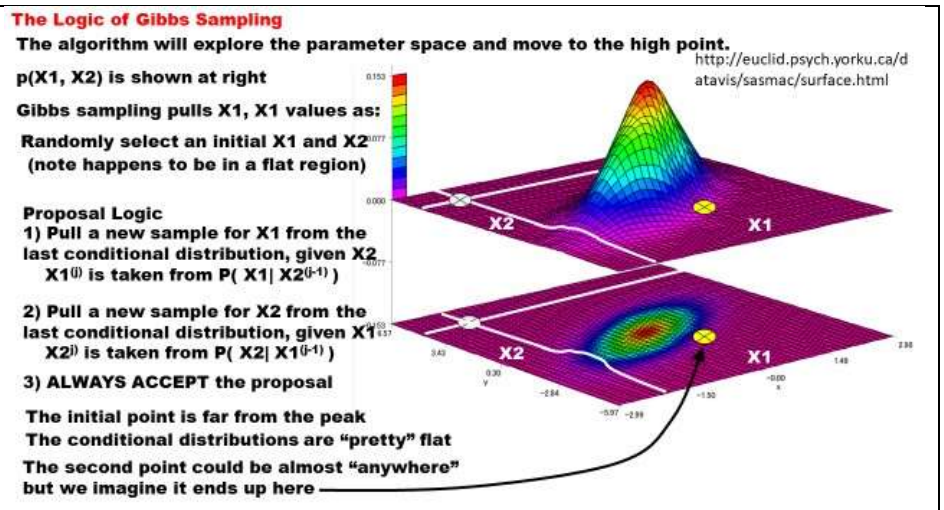


Figure 42

In this example parameters are called X1 and X2. As a 1st step, Gibbs will randomly select values of X1 and X2 (see white circle) and calculate the likelihood. Gibbs will now want to jump to a better position. Metropolis Hastings generated proposals using a normal distribution. Gibbs wants to sample from the posterior distribution but does not know what the posterior is.

As an approximation, Gibbs pulls a new sample for X1 from the last distribution, given that X2 is the value that it had in the last distribution (see white lines). It then repeats the process for X2. Gibbs does not know the posterior distribution but samples from "what it already knows". The idea is that whatever it already knows about the shape of the posterior is likely better than a normal distribution. These distribution are pretty flat and the first jump could be to anywhere. The first jump is to the yellow circle.

Figure 43 shows the next step for the Gibbs sampler. Gibbs pulls a new sample for X1 from the last conditional distribution holding X2 constant (see gold lines). It then repeats the process for X2.

The proposed conditional distributions are shown by the gold lines.

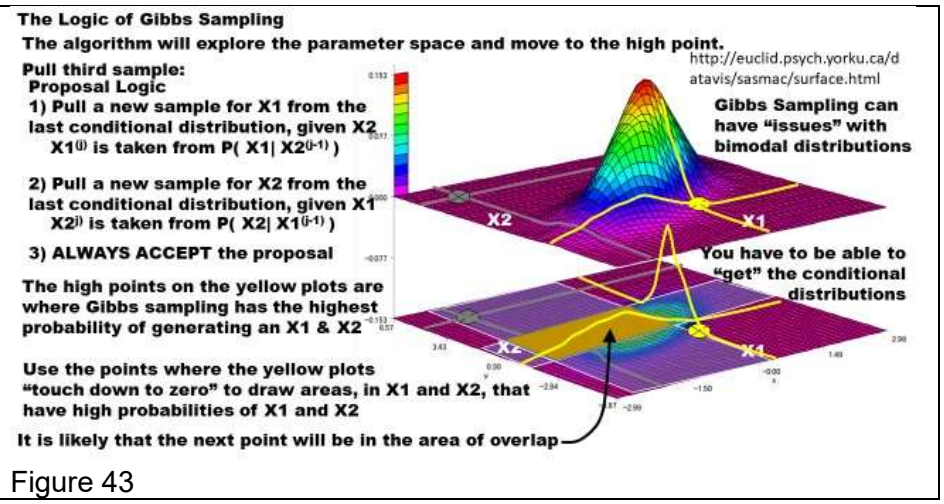


Figure 43

One of the limitations, in using the Gibbs, is that the algorithm must be able to calculate the conditional distributions, and this is not always possible.

Please look at Figure 43. If the next sample is going to be pulled from the place where the distribution for X1 and X2 are high it will likely be in the gold area and that is a significant improvement. Gibbs always accepts its proposed jumps. The process repeats

SAS EXAMPLE 1: MULTIPLE REGRESSION: GLM GENMOD AND MCMC

With a sufficient sample size, when you update the priors with “the data”, the data can overwhelm the influence of your prior belief, and Bayesian and frequentist analysis get to the same answer. In this section were going to do a compare and contrast between frequentist and Bayesian analysis in SAS and we expect similar answers.

PROC MCMC is the most powerful, customizable and complicated Bayesian procedure in SAS but there are other procedures that also give good Bayesian answers and are easier to use. PROC GENMOD is much easier to use than PROC MCMC. Let’s compare outputs from GLM, GENMOD and MCMC.

Figure 44 shows code and output for PROC GLM and PROC GENMOD. The actual code is below the figure.

All that is needed to get Bayesian out of PROC GENMOD.is a bayes statement. .

The answers produced by the two procedures are very similar.

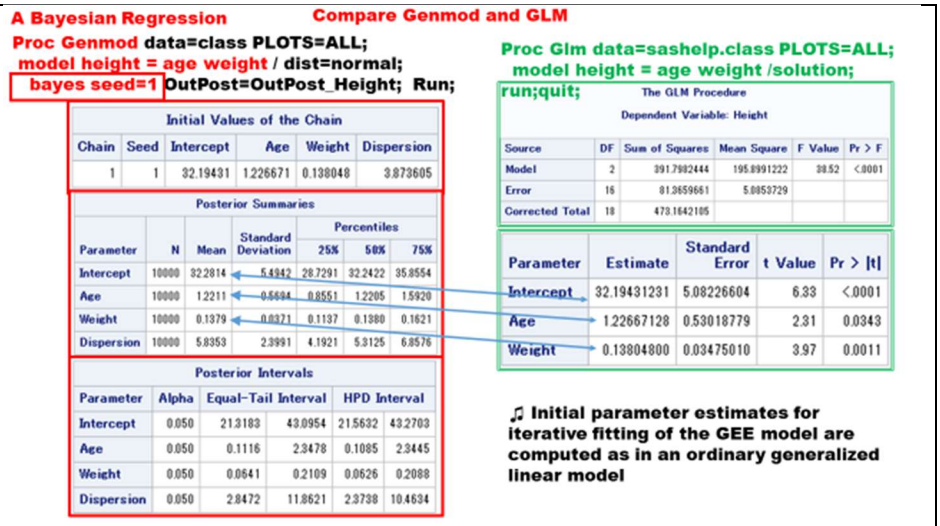


Figure 44

In place of confidence intervals, the Bayesian analysis will produce posterior credible intervals, that are kind of like confidence intervals. The code for the figure above is immediately below.

Proc Glm data=sashelp.class PLOTS=ALL; model height = age weight /solution; Figure 45	**Proc Genmod data=class PLOTS=ALL;** **model height = age weight / dist=normal;** **bayes seed=1 OutPost=OutPost_Height;** **Run;**

PROC GENMOD is smart enough to know when a conjugate exists for a particular model and will, because SAS wants you to get quick run times, use the conjugate formula rather than a simulation if it is possible.

The GENMOD Bayesian analysis also produces some diagnostics that are shown at right.

Most importantly are the plots on the right-hand side. They're describing the changes in the betas, caused by jumps, after "burn in" is finished and the jumps have reached the "top of the hill".

.

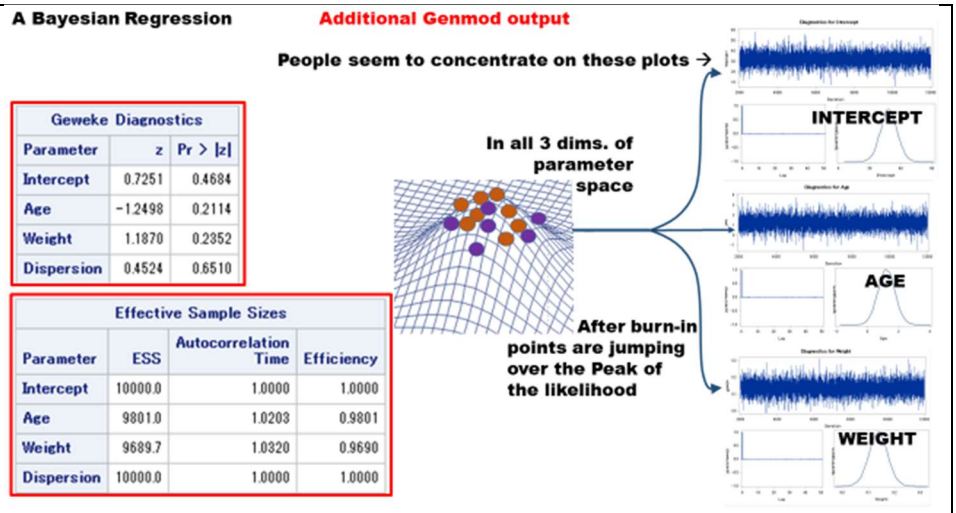


Figure 46

The plots you see at above show a mixture of small, and large, positive, and negative, jumps and are characteristic of the algorithm jumping, over and around, the top of the hill

The PROC MCMC code, and output, are shown below.
Data specifies the input data set.
Outpost specifies where the output should go.
NBI tells the number of jumps to be used as burn in. Burn-in jumps are not shown in the trace plots. We hope to reach the top of the hill before this number of jumps.
NMC is the number of jumps after burn in. The show up in the trace plots that can be seen in the figure above.
NThreads will allow you to take advantage of a multiple CPU computer.
Thin n only shows one n^{th} of the observations on the trace plot. This reduces correlation but I'm not sure if it helps much.
Seed starts a random number generator so that results can be reproduced.

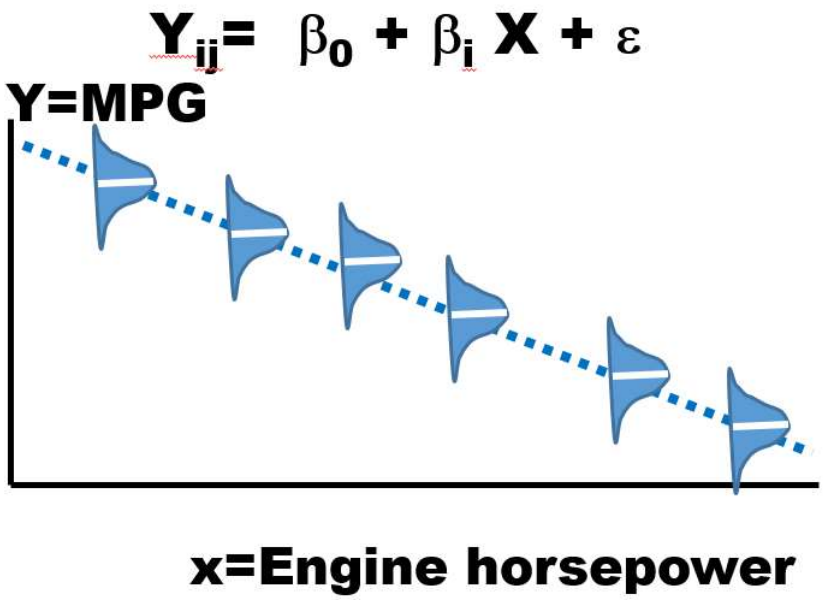
```
Proc MCMC data=SASHelp.class outpost=A03_MCMC_MltReg
  NBI=5000  /*# of burn-in jumps*/
  NMC=10000 /*# of MCMC jumps, excluding burn-in*/
  NThreads =3
  thin=5
  seed=246810;
  parms beta_Int 0 beta_Age 0 beta_Weight 0;
  parms sigma2 1;
  prior
     beta_Int beta_Age beta_Weight ~ normal(mean=0, var = 1e6);
  prior sigma2 ~ igamma(shape = 3/10, scale = 10/3);

  mu = beta_Int + beta_Age*Age + Beta_Weight * weight;

  model Height ~ n(mu, var = sigma2);     run;   ods graphics off;
```

Figure 47

Parm specifies the parameters and sets their initial values. You can have multiple parm statements.

The 1st parm statement specifies initial values for the X variables in our problem. We don't have much of an idea of what the beta values should end up being so were going to start them off at 0. If we have enough observations in our data set, the initial values will have very little impact on the final result.

The 2nd parm statement says we think the value of the standard deviation should start out at 1 (see Figure 47).

The prior statements provide more information about the prior distributions of the parameters. The betas for the X variables are expected to have a mean of 0 and the variance that's very large, making this a pretty uninformative prior. The beta for the error around the fitted regression line will be modeled using an inverse gamma with a shape and scale that makes this an uninformative prior. The mu statement and the model statement go together.	 Figure 48

Remember that regression predicts the conditional mean Y value for a set of Xs. A regression line predicts the conditional means of Y, given that the independent variables have some value. Regression also assumes that the errors around those means are normal.

The mu statement specifies the model relationship. The model statement says that height is a function of the relationship specified in the mu statement. It says that, if the hypothesized relationship, defined by the mu formula, is true, then the data should be normally distributed for a particular value of Y. Please go back a look at the dice example for a parallel.

The MCMC output, seen to right, matches the output from the previous two procedures.

The three PROCs gave the same answer.

Note: This model might benefit from a longer burn-in. The ESS numbers could be improved.

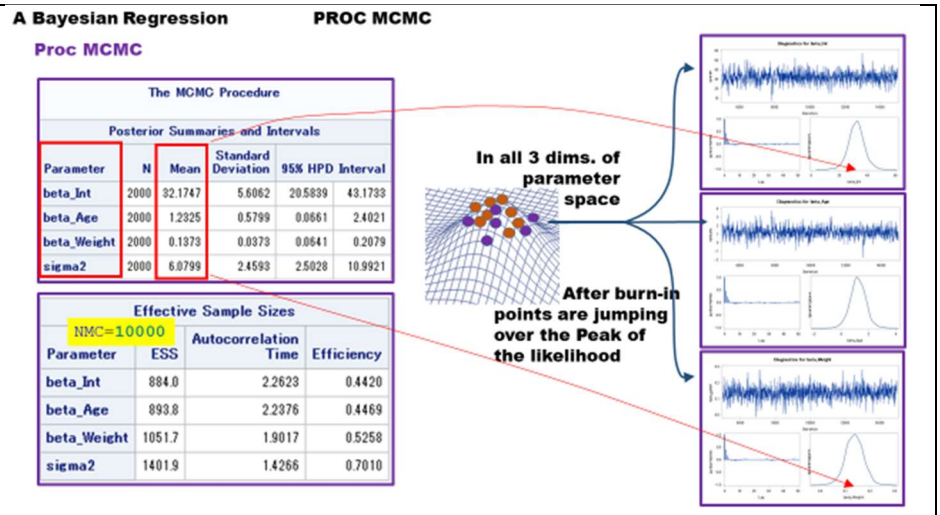


Figure 49

SAS EXAMPLE 2: LOGISTIC REGRESSION: GLM GENMOD AND MCMC

Figure 49, to right, shows the code and output from a logistic regression using PROC Logistic.

The code is simple and easy to understand.

PROC Logistic *data=vaso descending plots=ALL;*
model resp(Event="1") = lvol lrate / link=logit;

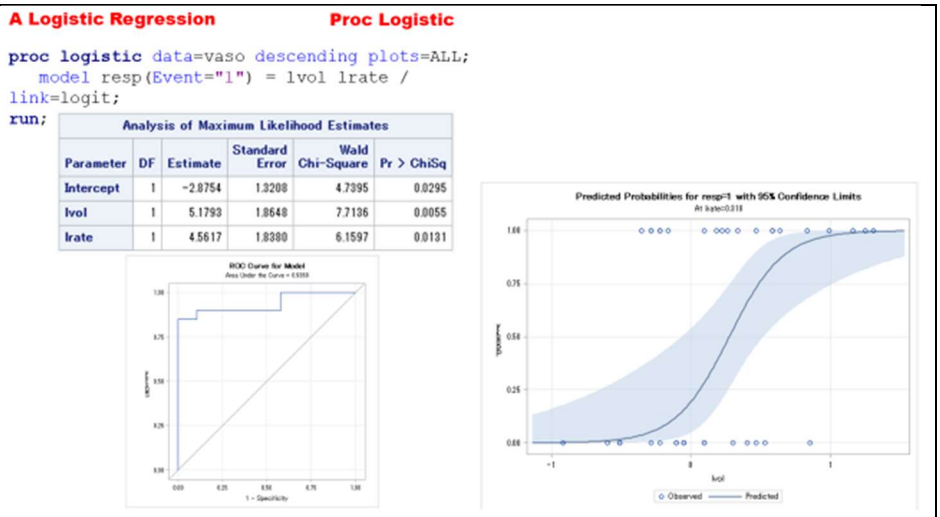


Figure 50

I think SAS ODS output is not only helpful but very professional and ready for inclusion in any sort of report.

The PROC GENMOD code is also easy to understand and options are similar to those in PROC MCMC.

PROC GENMOD
data=vaso
descending;
ods select
PostSummaries
PostIntervals;
model resp = lvol lrate
/ d=bin link=logit;
bayes seed=**17**
coeffprior=jeffreys
nmc=**20000** thin=**2**;
run;

A Bayesian Logistic **Proc Genmod**

```
proc genmod data=vaso descending;
   ods select PostSummaries PostIntervals;
   model resp = lvol lrate / d=bin link=logit;
   bayes seed=17 coeffprior=jeffreys
        nmc=20000 thin=2;
run;
```

```
NOTE: The PLOTS= option is ignored for a Bayesian analysis.
NOTE: The default sampling algorithm is the Gamerman algorithm,
which is different from the default in SAS/STAT 9.3 and
    earlier releases. To revert to the previous behavior, specify the
SAMPLING=ARMS option in the BAYES statement.
NOTE: PROC GENMOD is modeling the probability that resp='1'.
NOTE: Algorithm converged.
NOTE: The scale parameter was held fixed.
```

The SAS System

The GENMOD Procedure

Bayesian Analysis

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
Intercept	10000	-2.8773	1.3213	-3.8821	-2.7326	-1.9097
lvol	10000	5.2059	1.8707	3.8535	4.9574	6.3337
lrate	10000	4.5525	1.8140	3.2281	4.3722	5.6643

Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
Intercept	0.050	-5.7447	-0.6877	-5.4593	-0.5488
lvol	0.050	2.2066	9.4415	2.0729	9.2043
lrate	0.050	1.5906	8.5272	1.3351	8.1152

Figure 51

The seminar ends with the next, and last, example for PROC MCMC. **PROC MCMC is its own programming language, like the data step**. A detailed explanation of this language is beyond the scope of this seminar. The manual for PROC MCMC is a bit over 300 pages long and makes assumptions that the reader has a pretty good understanding of Bayesian statistics and of SAS programming.

With that caveat, it would probably be helpful to take a high-level walk-through of the code below.

```
%let n = 39;
PROC MCMC data=vaso nmc=10000 outpost=MCMCLogistic seed=17;
  array beta[3] beta0 beta1 beta2;    array m[&n, &n];          array x[1] / nosymbols;
  array xt[3, &n];                    array xtm[3, &n];         array xmx[3, 3];
  array p[&n];

  parms beta0 1 beta1 1 beta2 1;

  begincnst;
    if (ind eq 1) then do;
      rc = read_array("vaso", x, "cnst", "lvol", "lrate");
      call transpose(x, xt);
      call zeromatrix(m);
    end;
  endcnst;

  beginnodata;
  call mult(x, beta, p);            /* p = x * beta */
  do i = 1 to &n;
    p[i] = 1 / (1 + exp(-p[i]));    /* p[i] = 1/(1+exp(-x*beta)) */
    m[i,i] = p[i] * (1-p[i]);
  end;
  call mult (xt, m, xtm);           /* xtm = xt * m       */
  call mult (xtm, x, xmx);          /* xmx = xtm * x      */
  call det (xmx, lp);               /* lp = det(xmx)      */
  lp = 0.5 * log(lp);               /* lp = -0.5 * log(lp) */
  prior beta: ~ general(lp);
  endnodata;

  model resp ~ bern(p[ind]);  run;
```

Figure 52

In the first section of code we see array statements. Array statements will associate a name to a list of variables or constants. Note that, in MCMC, the array statement is similar to, but not identical with, the array statement in a typical SAS data step. The array statement in MCMC is the same as the array statements in NLIN, NLP, NLMIXED and model procedures. Note that the array statement in PROC MCMC is more limited than the array statement in the data step. The statement SCSUG(5) creates variables with the names SCSUG1 to SCSUG5.

The parm statement lists the names of the parameters to be used in the model and can also specify initial values for these parameters. You can specify multiple parm statements and this is useful when you want to apply different options to different parameters. Each parm statement defines a “block of parameters” and the Metropolis Hastings algorithm will attempt to jump all the parameters in a block simultaneously.

MCMC is faster if we put correlated parameters in the same “parameter block”. However; if many parameters are updated at one time, it is likely that one of the parameters in the block will end up being a large jump in one (and bad) direction – thereby lowering the ability of the “group of the parameters” to predict and causing PROC MCMC to reject all of the moves for parameters in that block. There’s a trade-off involved in blocking parameters.

Putting five parameters in five different blocks increases the run time by a factor of five, though it does increase the chance of jumps being accepted.

An important issue, in using PROC MCMC, is having a good mixture of the sequence chains. The jump points should quickly explore the parameter space and the blocking of parameters can be very important in this. The common practice is to block, on the same parm statement, small groups of correlated parameters.

The BEGINCNST-ENDCNST block causes PROC MCMCM to skip unnecessary evaluation in order to reduce the run time. Statements inside this block execute only during the set up stage of the simulation and are very useful to define constants or to load variables into arrays. If you have program statements, remember PROC MCMC is itself a programming language, statements that do not need to be evaluated for each loop through the simulation should be put in the BEGINCNST-ENDCNST. PROC MCMC will evaluate the programming statements inside this block once for each observation in the data set and ignore the statements during the rest of the run.

PROC MCMC can use multiple data sets as input and sometimes users need to store variables in arrays and use those arrays to specify the model. The read_array function, seen in this block of code is an easy way to read a data set into an array. PROC MCMC provide users with over 10 call routines that allow the user to perform matrix operations on arrays inside PROC MCMC.

PROC MCMC does not have a class statement and so design matrices must be created using PROC TransReg and imported into PROC MCMC.

The BEGINNODATA-ENDNODATA block defines statements that get executed without stepping through the entire data set. These statements are executed only two times: at the first and at the last observations in the entire data set. Any calculations that would be identical for every observation in the data set should be enclosed in this block. This block is run for the 1st observation to ensure that values have been calculated correctly. It’s run for the last observation because PROC MCMC executes the statements and then adds results to the output dataset.

Each parameter must have a prior. Inside the BEGINNODATA-ENDNODATA we see that this program uses the colon operator to initialize all of the beta values. This syntax is: PRIOR; one or more parameters; a tilde (~) and then the distribution to be applied with all of the distributions parameters. The name of the distribution to be applied is “general”.

The model statement specifies the conditional distribution of the data given the parameters. This is also the likelihood function given the hypothesized model. The model statement assumes observations are independent of each other and conditional on model parameters.

If you do not have data that’s independent of each other, SAS provides other procedures.

The model statement must come after any SAS programming statements that define, or modify, arguments used in the construction of the log likelihood in the model. This model statement uses bern, for Bernoulli indicating a binary distribution with having P as the probability of success. Ind evaluates to the probability of success.

The output from the PROC MCMC is below.

The PROC MCMC output starts with the information in Figure 53.

All of the observation that were read word used and it tells that the three parameters in the model were in one block and had an initial value of one.

The code that requests this can be seen in Figure 52.

Logistic Regression Model with Jeffreys Prior

The MCMC Procedure

Number of Observations Read	39
Number of Observations Used	39

Parameters

Block	Parameter	Sampling Method	Initial Value	Prior Distribution
1	beta0	N-Metropolis	1.0000	general(lp)
	beta1		1.0000	general(lp)
	beta2		1.0000	general(lp)

Figure 53

Figure 54 shows the estimated coefficients. The column titled mean, to the right, are the beta values. The estimated value for the intercept is -2.85.

The 95% HPD interval is a Bayesian version of a confidence interval. HPD stands for highest posterior density. If you think of a normal curve, there are many different cutpoints one could select such that 95% of the area under the normal curve is between the two cutpoints.

That same phenomenon happens in the Bayesian world. The 95% HPD interval is the narrowest interval that can be created that contains 95% of the observations.
Many statisticians suggest using the 95% HPD interval.

The ESS values are greater than 1000 and that gives us some comfort.

Logistic Regression Model with Jeffreys Prior

The MCMC Procedure

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
beta0	100000	-2.8513	1.3031	-5.4622	-0.5151
beta1	100000	5.1554	1.8517	1.7829	8.8176
beta2	100000	4.5236	1.8129	1.1834	8.0820

Logistic Regression Model with Jeffreys Prior

The MCMC Procedure

Effective Sample Sizes

Parameter	ESS	Autocorrelation Time	Efficiency
beta0	3314.5	30.1708	0.0331
beta1	3409.0	29.3341	0.0341
beta2	3254.1	30.7308	0.0325

Figure 54

Here is the report for beta 0.

The curve in the lower right hand corner shows the belief in the value for beta 0. The high point of the curve is not above the -2.85 that we saw in Figure 54. You will notice that this distribution has a long tail to the left and that will pull the average to the left of the mode.

The mixing shown at the top indicates that we have converged on the top of our distribution.

It would be hoped that the correlations were less severe but they do not impact the accuracy of the estimate and can be tolerated.

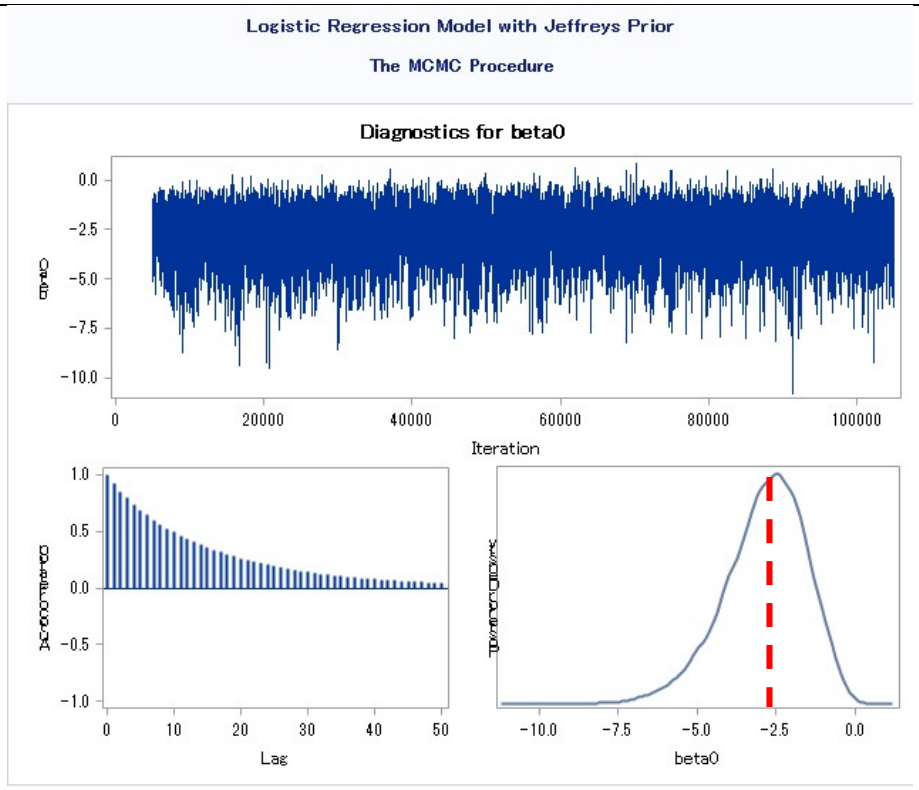


Figure 55

To the right is the report for beta 1.

The interpretation of this is very similar to the interpretation in Figure 55.

The numbers that we can see in Figures 55, 56, and 57 are very close to the numbers that were reported using other techniques on this data.

Those numbers are reported in figures above.

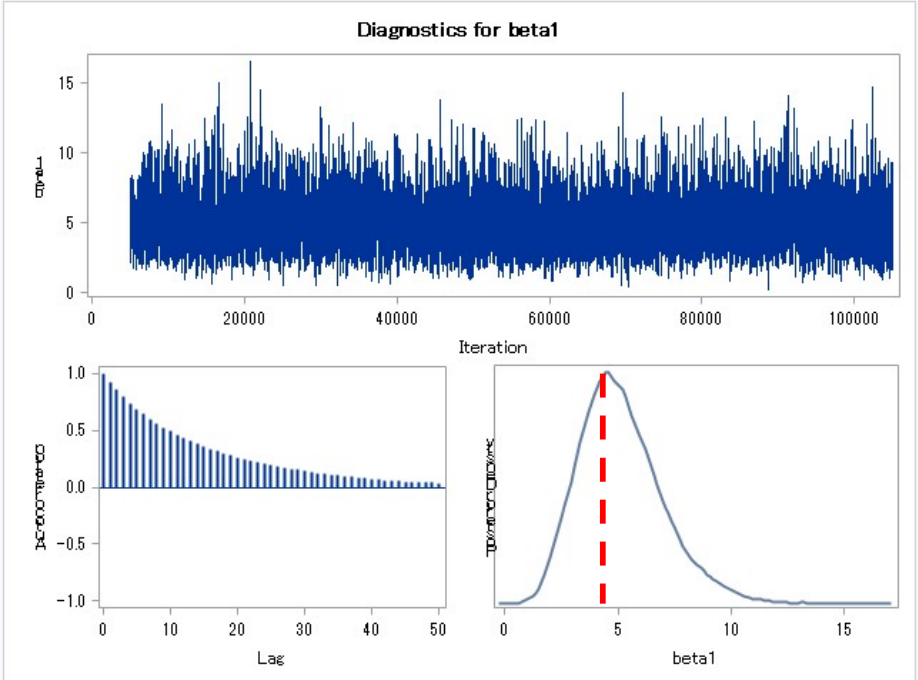


Figure 56

Figure 57 shows the output from PROC MCMC for beta 2.

The interpretation of this figure would be the same as for the previous two figures.

In summary, the three procedures that we used on the same data set all gave very similar answers.

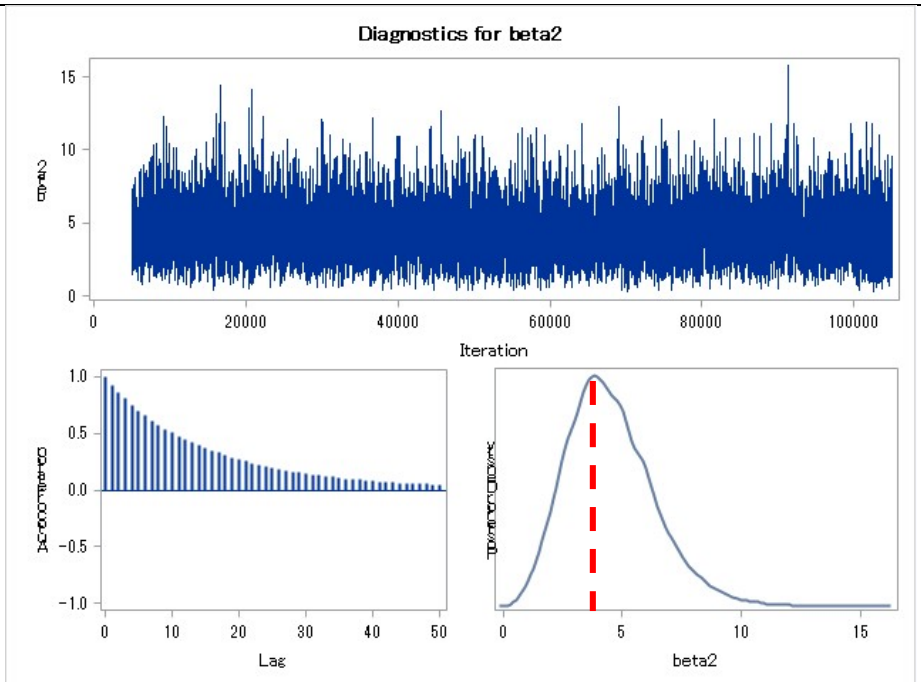


Figure 57

While PROC MCMC is very versatile and very powerful, Bayesian analysis, in a great many situations, can be done with procedures that have a syntax more familiar to people with experience with PROC GLM and would require less study.

CONCLUSION

Thank you for attending this seminar and I hope this has been helpful to you.

Russ Lavery

An Introduction to Bayesian Analysis Russ Lavery

INTRODUCTION:
This booklet is intended to support a seminar on introduction to Bayesian analysis. It is intended to ***not*** be mathematical but, instead, to focus on insight and making the concepts understandable to people who were not majors in mathematics or statistics.

BAYES' LAW PROVIDES USEFUL, AND COUNTER-INTUITIVE, RESULTS

The data in this next example is real and most doctors get this answer wrong. We are not going to do much explanation in this section. This is just one worked example that, hopefully, will get you interested in the subject.

A mammogram has an 87% chance of detecting breast cancer when it is present. A mammogram will falsely signal breast cancer 12% of the time (when it is not present). About 1% of women between 40 and 60 have breast cancer.

If a mammogram signals cancer, what is the chance of the subject having cancer? To get the answer we should use Bayes' law.

$$P(C \mid \text{Positive Test}) = \frac{P(\text{Positive Test} \mid \text{Cancer}) * P(\text{Cancer})}{P(\text{Positive Test})}$$

$$P(C \mid \text{Positive Test}) = \frac{.87 * .01}{(.01*.87)\ (.99*.12)} = \frac{.0087}{.1275} = .0682 \text{ or } \sim 7\% \leftarrow \text{does this surprise you?}$$

Most people guess a number closer to 87%. The false positive percentage I used is the upper limit on the actual percentage (7% to 12%). If a woman were to get 10 yearly mammograms the chance of at least one false positive result is about 50%. Thankfully breast cancer has been declining since 2000.

HISTORY:
The Bayes formula is not new. The original insight was created by Thomas Bayes who lived between 1702 and 1761. A little history might help in understanding how we got to where we are.

In 1662 the English Parliament passed the act of uniformity which required all English subjects to use the Book of Common Prayer. Thomas Bayes father objected to this and was what was called a "non-conformist" minister. Nonconformists were any non-Anglican or non-Christians (Methodists, Baptists, Congregationalists, Puritans, etc.), groups that are pretty mainstream today in America. Nonconformists could not get university degrees in England for over 150 years and so Thomas Bayes got his degree in Scotland.

He never published a math paper in his life but was elected to be a fellow of the Royal Society. The formula we know as Bayes' law was discovered in his papers after his death and was published by a friend. I have heard that Bayes was trying to discover a formula to prove the existence of God. It is possible he never saw the full practical applications of his concept. The formula that we know as Bayes' law was formalized by Pierre-Simon the marquis de Laplace in *Essai philosophique sur les probabilités* (1814). Laplace wanted to apply his new idea, that he called inverse probability, to things like courts of law. I think that you can see some of this influence in our famous statement "a man is assumed to be innocent until proven guilty beyond a reasonable doubt." The reason people think this is Bayesian is because Bayes' law models how we update our beliefs in the presence of new information/evidence.

For a long time Bayes' Law was not used very often because calculations are difficult and because of opposition to priors.

R. A. Fisher (1890-1962) has been described as "a genius who almost single-handedly created the foundations for modern statistical science" and "the single most important figure in 20th century statistics". Oddly, Fisher was primarily an agricultural researcher and was concerned with improving food production. He was the first to use diffusion equations to attempt to determine the distributions of different forms of genes (allele) by analyzing genetic links applying MLE to populations. He has been called the greatest of Darwin's successors.

Fisher was a very powerful statistician and Fisher hated Bayesian Analysis. He wrote "The theory of inverse probability is founded upon an error and must be wholly rejected." He attacked people who supported Bayesian analysis and he was a powerful deterrent.

Neyman and Pearson helped put Bayesian Analysis on the back burner when they helped formalize/systematize the methodology of research. They developed the methods of hypothesis testing and confidence intervals that revolutionized applied and theoretical statistics. It was not victory of competing theories but of practical application. Bayesian calculations were difficult and Fisher, Neyman and Pearson had good practical methods for doing analyses.

A recent book, "The Theory that Would Not Die" asserts that there was lots of Bayesian analysis done during World War II but that the projects were all classified – during the war and for a long time afterwards. The book asserts that Bayesian analysis was used to crack the Enigma code. Bayesian analysis is asserted to have been used to determine the number of German tanks.

The book also cites a story about using Bayesian analysis to discover a lost nuclear bomb in the ocean. I have read several accounts of this and would like to fill in some of the information that's sometimes is mentioned and sometimes not. Bayesian analysis is supposed to use prior information. A local fisherman reported seeing something fall into the ocean right next to where the bomb was found. That information seems to have been ignored until late in the search.

The ocean bottom had been divided into fairly large squares and each square was assigned a calculated probability of containing the bomb. As each square was searched, and found to be not containing the bomb, the probabilities for all of the squares had to be recalculated. There was relatively little computing power on the search vessel and so the updates of probabilities took quite some time. The great amount of time required to update the probability for a particular square had forced the squares to be fairly large.

The bomb was found, very close to where the fisherman said it had landed, in a low-probability square but near the border between that low probability square and a higher probability square. After reading several accounts of this process I have an odd feeling that if they had listen to the fisherman at the beginning they would've found the bomb quite quickly.

However; Bayesians claim this as a success. I have listened to many Bayesian statements and I think the claim that we are all Bayesian's is founded on the idea that we all use prior information in making decisions. We learn from our mistakes and, hopefully, don't repeat them. When a Bayesian sees a mouse learning to find cheese at the end of the maze the Bayesian say "mice are Bayesian". I leave this discussion to other people.

I would like to take a bit of this paper to explain what I have gleaned from readings about cracking the German code in World War II. Alan Turing was once asked, of a technique he had invented, “Isn't this really Bayes' theorem?" His reply was “I suppose so.” and I find that a lukewarm confirmation.

The code cracking problem was that the Germans had a mechanical coding machine with several “swappable” rotors”. The number of possible combinations of rotors in the German code machine was astronomical. Applying brute force methods, just trying a particular combination and seeing if it worked, would have taken too long.

However; several German radio operators violated common secrecy practices allowing the British to have “cribs”. British universities used to focus on translating Latin and Greek into English. A “crib” was used by students to help them in their translation homework. It was basically an English version of the Latin text. Students would pass down completed translations, “cribs”, to younger students and these “cribs” greatly helped struggling students.

Weather ships often sent the same information every day in the same way (longitude, latitude, rain, wind direction, wind speed, cloud cover etc.). One German radio operator was in love with a girl named Cecilia and, every day, he would set the code wheels to the settings for that day and then broadcast to the world his girlfriend’s name. The code-breakers knew the contents of these messages and this was a great help. Some books suggest that the codebreakers didn’t do much work on a day until the “cribs” arrived. One book has a code breaker asking “have we heard the silly yet” as a reference to the coded version of Cecilia.

Having the cribs greatly reduced the number of “rotor setting” combinations that had to be considered but that number was still a very large number. I’ve heard that the cracking method was to take all the possible combinations of “rotor settings” and apply each setting to a large number of messages. If the combination was wrong, the message would be nonsensical. If the combination was right then the message that came out would make sense to a human. But there were too many combinations for a human to try and read the output.

The ingenious trick I heard about involved letter frequencies. Each language has frequency of use for each of the letters. In English, E is the most common letter and Q is relatively infrequent. German has a letter frequency that is different from English. Letter frequency was key to cracking Enigma.

After the British computer produced a translation for a hopeful setting of the coding machine wheels a little program calculated the relative frequencies of the individual letters in the message. The observed relative frequencies were compared to the relative frequencies to be expected from German military messages. Messages where the sum of observed minus expected was small were sent to human to read.

If this sounds like the chi-squared formula to you, you are correct. So maybe chi-squared cracked the Enigma. Maybe it’s unfair to say that Bayes’ Law cracked the code produced by the Enigma machine simply because we used prior knowledge contained in the cribs.

Stories being heard now are that the military continued to apply Bayes’ law after the war but kept all the success stories confidential. In summary: the difficulty of computing answers, the opposition by giant figures in the field of statistics and the lack of publically available success stories about how Bayes’ law had been applied contributed to Bayesian analysis being a minor activity.

Most writers suggest that the increase in Bayesian use came about when the MCMC algorithm was matched up with the increased computing power that started becoming available in the 1960's. This made it practical to solve Bayesian problems – removing one of the main factors in the dominance of frequentists.

The rise of Bayesian Statistics, while benefitting from MCMC and computing power, was a complex process and many people contributed. A fine history is "When Did Bayesian Inference Become Bayesian"? by Stephen E. Fienberg https://projecteuclid.org/download/pdf_1/euclid.ba/1340371071

YOU ARE A BAYESIAN.

People who are proselytizing Bayesian analysis will say we all are Bayesians. I've even read that dogs are Bayesian. The support for this statement is that people who make statements like this believe that anyone who uses prior information is using a Bayesian process. I do think that we all update our plans when we hear new information but am not sure that we all apply, in the neurons of our brain, a formula that's like the Bayes' Law. I don't think our minds are that simple and would suggest that reading a book called "Thinking Fast and Slow" is well worthwhile.

BAYES LAW: THE FORMULA AS IT IS USED BY A FREQUENTIST

The Bayes' law formula has many looks, and this can make it a bit confusing. We will endeavor to use worked examples to give concrete illustrations of the words that are used in Bayesian statistics and the parts of the formula.

P(A|B) * P(B) = P(B|A) * P(A)
is the basic formula and, to the author, offers no insight. We ask for the reader's indulgence while we state some theory.

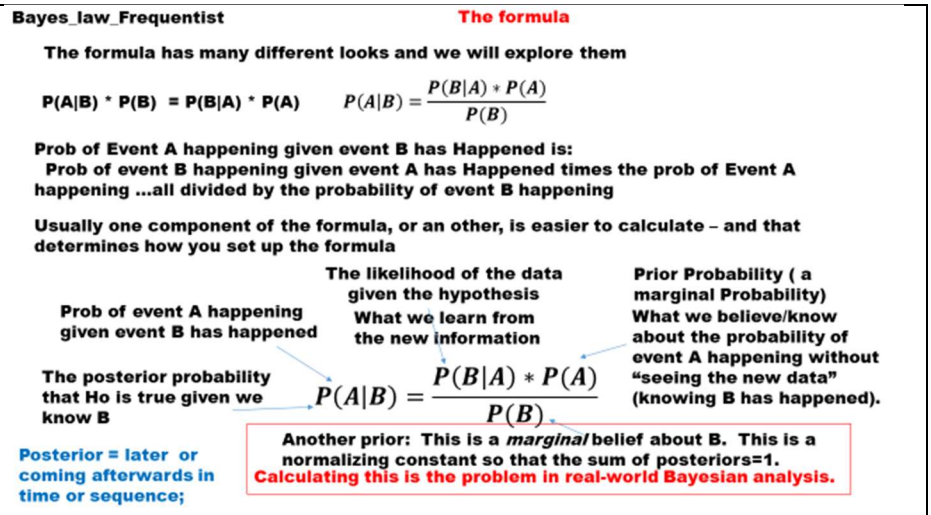


Figure 1

I suggest a quick read of the theory – then a study of the examples – and then looping back to the theory again.

Bayesian analysis is the updating of our beliefs in the presence of new information. We represent our prior belief by P(A) and the new information by P(B). The most commonly seen version of the formula is shown below.

$P(A \mid B) * P(B) = P(B \mid A) * P(A)$ → very simple math → $P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$

$(A \mid B)$ is the posterior probability- our belief after we have seen the new information. It says that the probability of event A happening, given that we know event B has happened, is equal to the right-hand side of the equation.
$(B \mid A)$ is the likelihood of event B happening given that event A is happened. It is the likelihood of the data given that our hypothesis is true (more on this later). This is an inverse probability, at least inverse to the posterior probability.
(A) is the prior probability. It is our belief about the probability of A happening without knowing about B.
(B) is a normalizing constant. This is a prior belief about B occurring. It makes the result of the calculation a true percentage and will need to see some more information before we can explain why that is true.

NOTE: In formulas, I will flip between using A and B in formulas and other letters. A and B show up in books and we should see that notation. However; I find that using S for shy and A for actuary makes things easier to follow.

Let's use an example to make some of these concepts more concrete. Imagine going to a university event that is a mixture of marketing and actuaries (apologies to all). Your +1 says that they met a shy person. You think that actuaries are shy and wonder what was the chance that this person was an actuary.

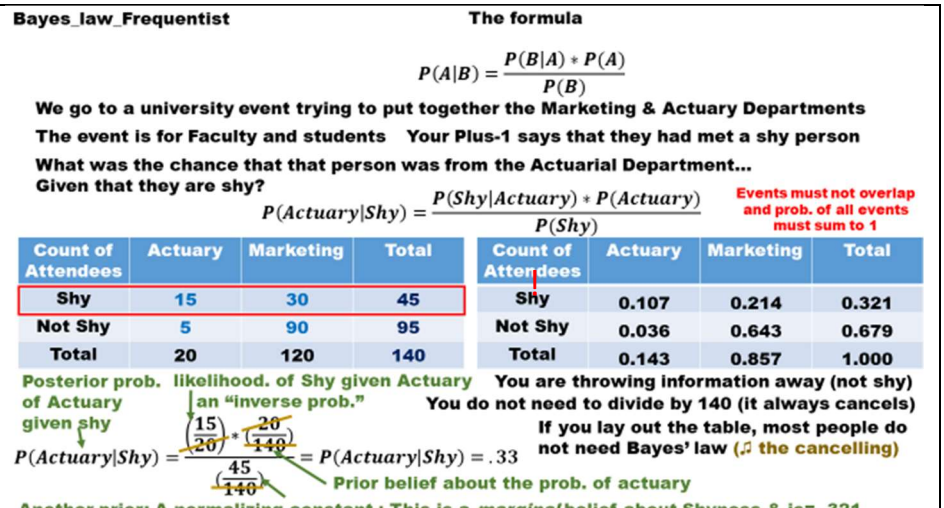


Figure 2

The table, in Figure 2 allows us to work out Bayes' Law. In fact, if the problems are laid out in tables most people do not need Bayes' Law to get the correct answers.

The basic equation of Bayes' law can be seen in Figure 2.

The probability of being an actuary and shy (see exclamation point) is .107 <u>of the total number of people</u>

and, importantly, we can get to that .107 two different ways: P(A|B) * P(B) = .P(B|A) * P(A) = (A ∩ B). This is evident from 2x2 tables as seen in Figure 2

The next two lines illustrate the equivalence of P(A|B) * P(B) = .P(B|A) * P(A)

Prob(S | A) * P (A) ← = → P(M | S) *P(S)
(15/45) *.(45/140) = .1071 ← = → (15 / 20) * (20 / 140) = .1071

Above is the basic logic. Using simple division on P(A|B) * P(B) = .P(B|A) * P(A) gives Bayes' law below.

$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$ ← so we have a proof of this ☺ and can go on and use the formula.

That .107 number is the percentage of shy **and** actuary as a percentage of the total number of people. Our question is "what's the probability of being actuary given that we know the person was shy". Attendees at the meeting were a mixture of shy and "not shy".

Please look at the red box in Figure 2. Given that we know you're shy we throw away part of the information in the table. Given that we know you're shy we only are concerned with the information inside the red box. Bayesian analysis always throws away information and, if we look at the red box, we are throwing away the information about people who are not shy.

What does the denominator do for us? *The answer only can be seen if we do TWO calculations*.

$P(A \mid S) = \frac{P(S \mid A) * P(A)}{P(S)}$ → $(15/45) = \frac{P(15/20) * P(20/140)}{P(45/140)} = \frac{.75 * .1428}{.3214} = .333$

$P(\text{Not } A \mid S) = \frac{P(S \mid \text{Not } A) * P(\text{Not } A)}{P(S)}$ → $(30/45) = \frac{P(30/120) * P(120/140)}{P(45/140)} = \frac{.250 * .857}{.3214} = .666$ 1.00 ☺

The denominator is a normalizing constant that makes the probabilities of the two possible outcomes (actuary or not actuary) sum to one and this means this calculation returns a true percentage and probability. Remember, if you sum the probabilities of occurrence, for all things that can occur, the sum must equal one.

Bayes_law_Frequentist - practice **The formula**

The formula has many different looks and we will explore them

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} = P(F|needs\ help) = \frac{P(B|A) * P(A)}{P(A \cap B) + P(\hat{}A \cap B)} = = \frac{P(needs\ help|F) * P(F)}{P(F \cap Help) + P(M \cap help)}$$

You are always throwing information away – 20 people need help

Count of Students	Needs Reading help =YES	Needs Reading help =NO	Total
Male	15	30	45
Female	5	90	95
Total	20	120	140

P(F | Yes) = $\frac{P(Yes \mid F) * P(F)}{P(Yes)}$ = _____ =

P(M | Yes) = = _____ =

P(Yes | M) = = _____ =

P(No | M) = = _____ =

P(Yes |F) = = _____ =

P(No | F) = = _____ =

Figure 3

I think this formula is easy to calculate but kind of confusing and therefore worth a couple of examples. Figure 3 shows another 2x2 table and gives us some space to do some calculations.
The answers are below. Arrows point to formulas with the same denominator and these should sum to 1.

Answers for Figure 3				
P(M|Y)	= 0.75		P(Y|M)	= 0.333
P(M|N)	= 0.25		P(Y|F)	= 0.053
P(F|Y)	= 0.25		P(N|M)	= 0.667
P(F|N)	= 0.75		P(N|F)	= 0.947

Let's look at how this formula can quickly become confusing.

P(help) = 20/140 and that can be written as P(help ∩ M) + P(help ∩ F) ←these are just cell counts of 5 and 15. We could also write this as P(help ∩ Male) + P(help | NOT Male) and some books do,
The situation is: Many formulas and little insight.

Here is a major hint – it changes Bayes' law into English words and provides insight/

$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$ → P (A|B) = the probability of A and B occurring together divided by the probability of all the ways B can occur.

This can be helpful in word problems. The numerator is the probability **of A and B both occurring**. The denominator is just the **sum of the probabilities of all the different ways that B can occur.**

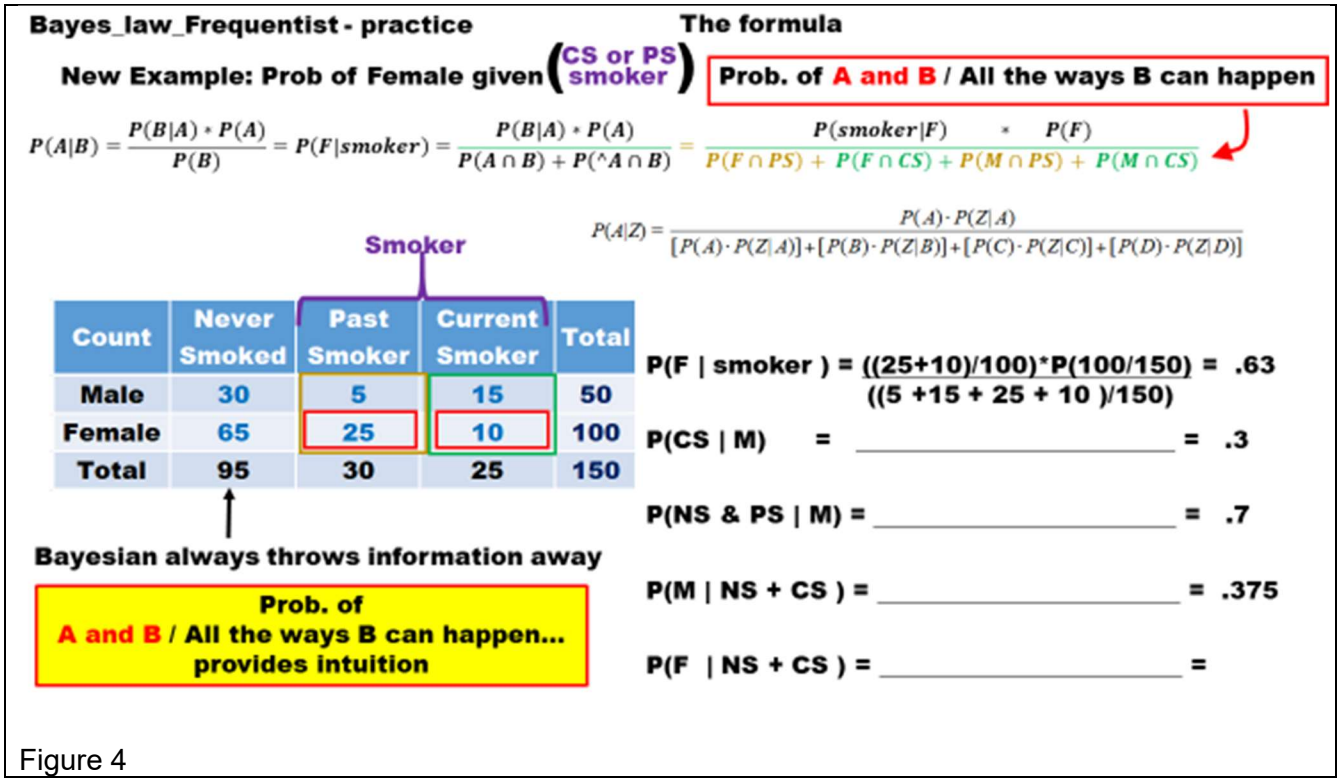


Figure 4

Figure 4 is another chance to do some calculations and emphasizes the helpful hint mentioned above. Here we have three categories and this allows us to emphasize that the denominator is the sum of the probabilities associated with all of the different ways that B can occur.

I think saying that Bayes' law is the probability of A and B occurring together divided by the probability associated with all of the different ways that B can occur allows me to check my logic in more complicated tables.

The red rectangles shows the probability of A and B occurring at the same time – that is the probability of being a female and a smoker (past or current). The orange and the green boxes show all the different ways you can be a smoker (past or current). Some of the calculations requested above are a bit odd but they will allow us to explore the Bayesian formula on something other than a 2 x 2 table.

Some books explain Bayes' law using a Venn diagram and I think that's not the best way. Please look at the Venn diagram in Figure 5 and compare it to the cross tab in Figure 5.

The Venn diagram shows the shy and actuary and the intersection of the two characteristics.

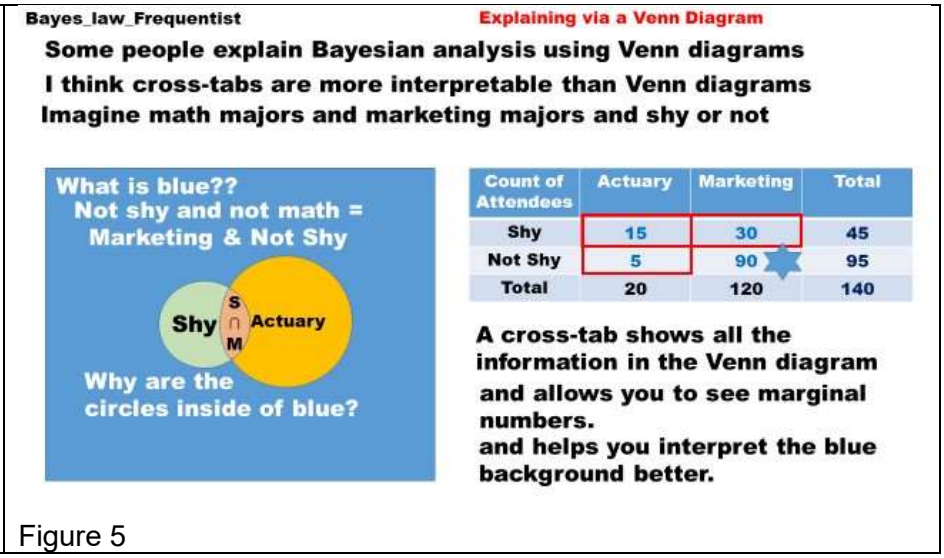


Figure 5

I think a cross tab is superior because it shows the same information as the Venn diagram and more. The crosstab shows the actual numbers that are associated with the characteristics of the people. The crosstab also shows marginal totals and information about the problem that is not shown well in the Venn diagram. Marketing people that are not shy are the blue area in the Venn diagram, but that is not obvious.

I would suggest that, if you have a 2X2 table and you want to make a Venn diagram, draw little boxes on the table itself. That will be more informative than creating a Venn diagram.

A SHORT DIGRESSION ON THE IDEA OF A LIKELIHOOD

Bayesian analysis often uses likelihoods as a way to avoid calculating the denominator in the Bayes' Law formula. It would be good to spend some time thinking about the difference between a likelihood and a percentage.

In the upper right corner of Figure 6 we see the formula for the height (the PDF) for the normal curve.

The red denominator in the formula is a normalizing constant. It makes the area under the normal curve integrate to one.

The fact that the area under the curve integrates to one is a requirement, if areas under the curve are to be interpreted as a percentage and a probability.

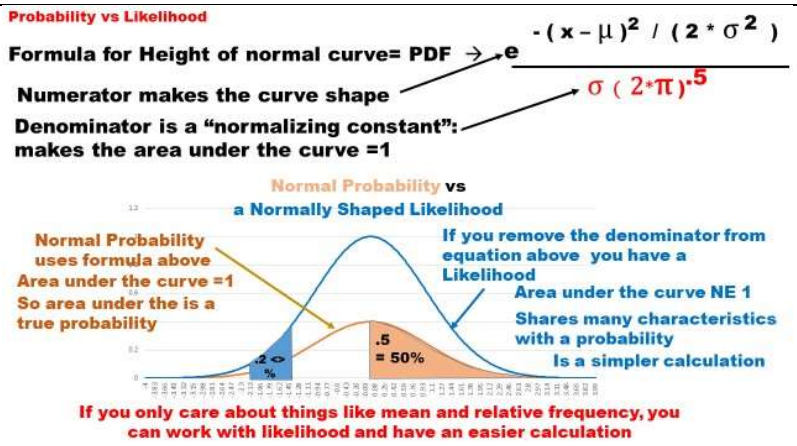


Figure 6

The fact that the area under the tan curve integrates to one allows us to interpret the number we get when we integrate under that curve– the one area number that is the output of the integration - as a percentage and a probability. If we integrated under the tan curve from Z equals 0 to Z equals positive infinity the area would equal .5 and because the area under the curve sums to one we can interpret that area as a percentage and a probability.

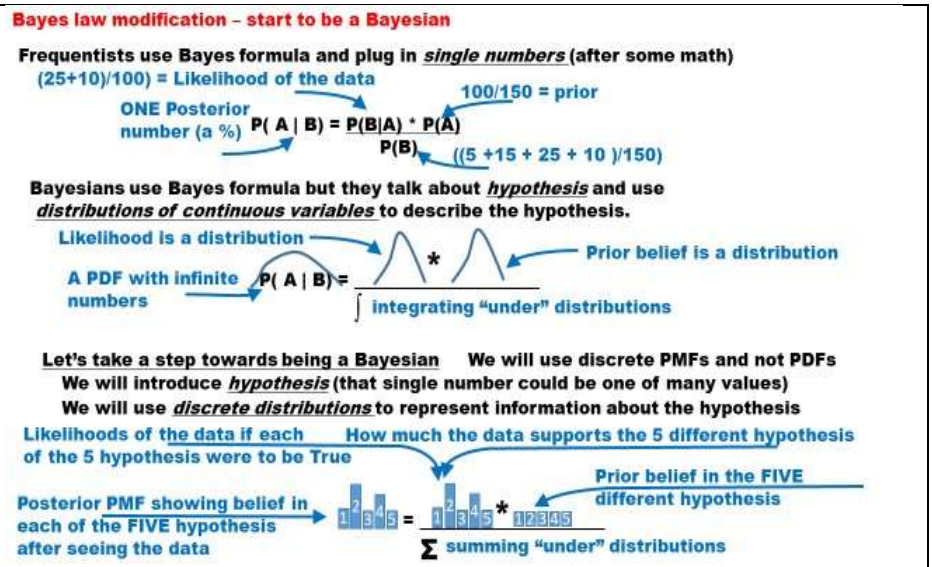
If we remove the denominator from the "normal" formula we get the blue curve in Figure 6. If we integrated under that curve we might get a number - say 10 as the area under the curve. If we integrate to find the blue shaded region we might get a number like .8. We cannot say that the probability of getting a number in that range shown by the blue distribution is 80%. Results of integrating under the blue curve cannot be directly interpreted as percentages or probabilities.

However; the two curves have a lot of similarities. Their centers are the same and they have the same shape. If that denominator is difficult to calculate, and all we need are a few simple characteristics of the curve, it is sometimes easier to work with the likelihood curve (blue) then the normal curve. It just saves us some calculations.

Remember that calculating the denominator in the formula for Bayes' Law is usually the difficult part of the formula. The denominator is a normalizing constant so that the area under the curve sums to one and so that the areas can be directly interpreted as percentages. Will see this in the examples to follow.

BAYES LAW: USING THE FORMULA MORE LIKE A BAYESIAN WOULD USE THE FORMULA

Example 1

Frequentists use Bayes' Law quite often and use it as we have shown above. They tend to use single numbers (or the results of simple calculations) as each of the components of the formula. They are certain of the values that they are entering into the formula and therefore only need to enter one, scaler, number. .	 Figure 7

Bayesian's have a different approach and are less certain about their numbers Because they are less certain about their numbers Bayesian's will describe their prior belief in someone being shy as a distribution. A Bayesian might say the following.

"I think 32% of the students at this university are shy but I want to have some uncertainty. My best guess is 32% but it might be 30% or maybe 35%. I'm pretty sure it's not 50% and very sure it's not 60% or 5%. Rather than providing you one number about my belief I'd like to provide a distribution."

That is what we are working towards but we want to get there in baby steps. There are some mind–stretching concepts to get through before we can apply Bayes' Law as a Bayesian. It's going to be useful to do several examples so that we can see pictorial representations of these new concepts.

The top formula in Figure 7 is Bayes' law as used by a frequentist. Each part of the formula is one number and this implies we are very certain about this number. Frequentists look only at the data that they have and so they can be very certain about the characteristics of the data that they collected.

The middle formula in Figure 7 is Bayes' law as used by a Bayesian. Each of the parts of the formula are continuous distributions. The divisor is actually an integral over part of a distribution and this is where Bayes's Law gets really difficult. In the numerator we are going to be multiplying two distributions. In the denominator were going to be doing an integral over a joint distribution. Astronomers, when they're using Bayesian analysis to try and find a planet around another star, might have 15 to 20 variables describing the joint distribution in the denominator. We mentioned before that the Bayesian calculations were difficult. It's the calculation of the denominator, the integral over many different X variables, which makes Bayesian calculations so difficult. In fact, there are only a few exact, or closed form, solutions to Bayesian problems. Modern Bayesian analysis uses approximation algorithms, not closed form equations.

So frequentists use the formula with only one number for each of the components. Bayesian's use the formula with continuous distributions as each part of the formula.

Our next step is shown in the bottom of Figure 7. We are going to take a “baby step” towards using a continuous distribution by using a discrete distribution. Instead of taking integrals, we will be able to sum numbers. This will simplify calculations but still give concrete examples of the difficult concepts.

Imagine we have several dice in a cup. You can buy four sided dice six sided dice eight sided dice 12 sided dice and 20 sided dice and they look like the picture in the upper right-hand corner of Figure 8.

We put the dice in a cup and allow someone to, without showing anyone else, pick a die and then roll it. That person tells us the number on the dice and says it is a 12.

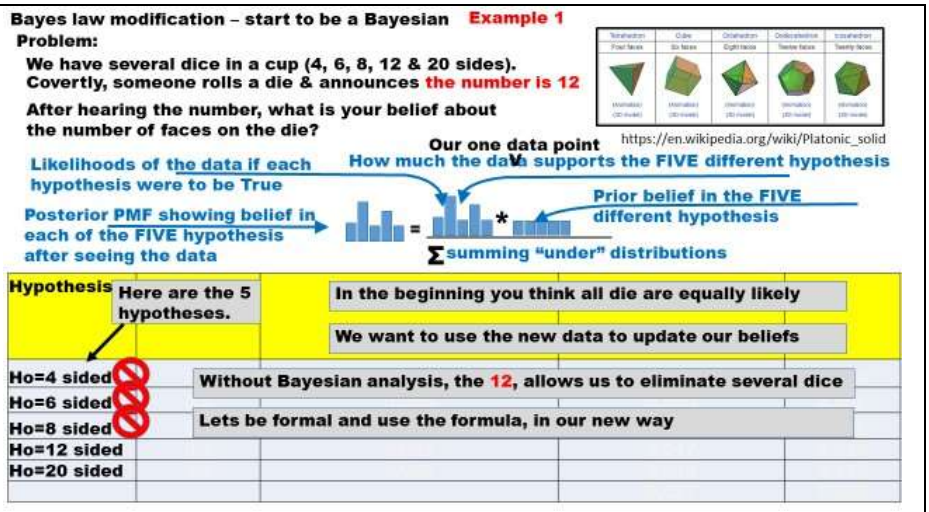


Figure 8

Figure 8 shows a shell of a “calculation aid” table. We’re going to gradually fill in the columns in a way that mimics the formula for Bayes’ Law.

THE HYPOTHESIS: (in example 1)

We would like to use Figure 8 to help understand the concept of a hypothesis. Hypotheses can be true or false and we are setting up five competing hypotheses. Our competing our hypotheses are shown in the leftmost column in Figure 8.

Since we only have five dice we are going to use a discrete distribution. If we could create dice that would roll a fractional number like 3.1725 we could make this into a continuous distribution – and then we would have to do different math. Let’s use discrete distributions, in the next few examples, so that we can develop the concepts.

In Figure 9 we add our prior belief - how likely we think it is that each die was picked.

This example illustrates us having no idea, before hearing about the 12, which die would be picked.

Our idea is that without any information, and five dice in the cup, we would think that each die has a 1/5 chance of being selected.

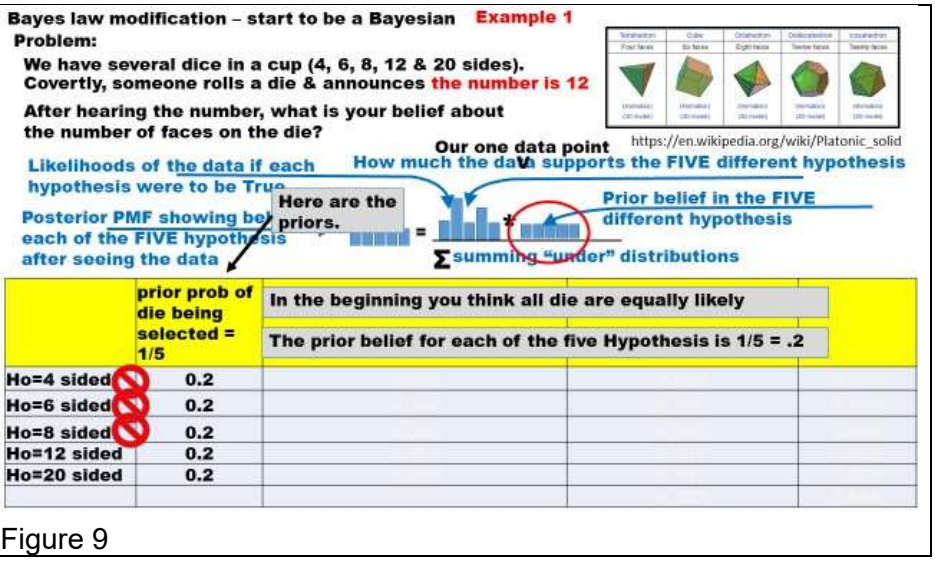


Figure 9

That is our prior belief in the probability of each hypotheses.

Figure 9 shows a prior belief in the different hypotheses that would be called a flat prior or an uninformative prior. That prior describes our belief in how likely each die is to be selected – our belief in the truth of each hypothesis. We don't have any idea which die would be selected. This column corresponds to the prior distribution and that part of the formula is indicated with a red circle. Notice that, in the picture, the prior distribution, is flat.

This is our prior belief in the hypotheses and we want to update our belief in the hypotheses based on new information (the 12). I'm sure you realized, and didn't need Bayes' Law to understand, that if the dice rolled the 12 it could not be one of the dice with four sides or six sides or eight sides. We didn't need Bayes' Law for that. We want to use Bayes' Law, and formalize, the understanding that you already have.

In Figure 10 we add what are commonly called "the likelihoods" and I'd like to add some more words to that much too short phrase. The column we added shows the likelihoods of getting the data we saw if the Ho on that row is true.

The likelihoods part of the formula is circled in red on the slide.

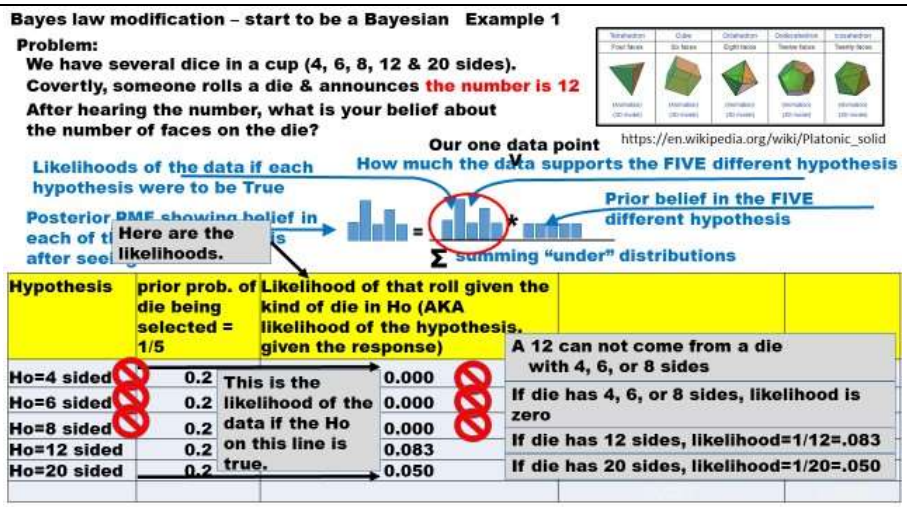


Figure 10

The heights of the bars are only illustrative of the concept and are not proportional to the likelihoods. I will use this blue graphic in many places and updating the sizes tiny boxes too tedious. However; if we were to update this graphic to match the data in this slide the first three bars would have zero height.

In the Bayes' formula, this part of the formula is P (B|A) and this stands for the likelihood of the data occurring if the hypothesis is true. If you have a 20 sided die, any particular number has a 1/20 chance of occurring. If you have a 12 sided die, any particular number has a 1/12 chance of occurring. These are the likelihoods of getting a 12 given that you have a 12 sided die or a 20 sided die.

Figure 11 adds a column that does the multiplications for the numerator of the Bayes' Law formula. This action is indicated by the red oval on the graphic. These numbers do not sum to one and so they are not percentages or probabilities.

They do show our relative belief in the hypotheses but they cannot be interpreted as percentages or probabilities.

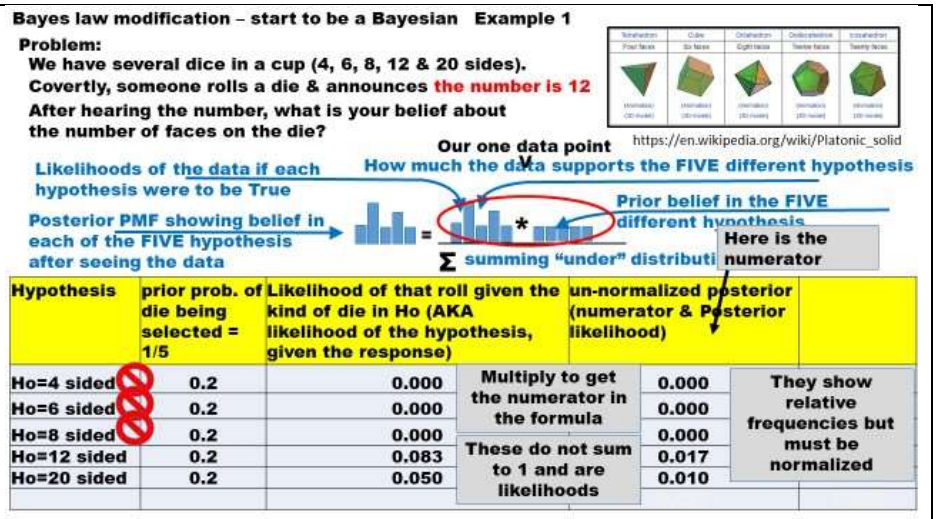


Figure 11

To do that we must bring in the denominator – the normalizing constant.

Figure 12 shows the completed calculations. If we add up the column called un-normalized posterior we get .27 and that is the normalizing constant. We use that to create the posterior percentage. As examples:

.017÷.027 equals .625. and .01÷.027 equals .375.

These numbers sum to 1 so we can talk about the ***percentage chance*** of these hypotheses being true.

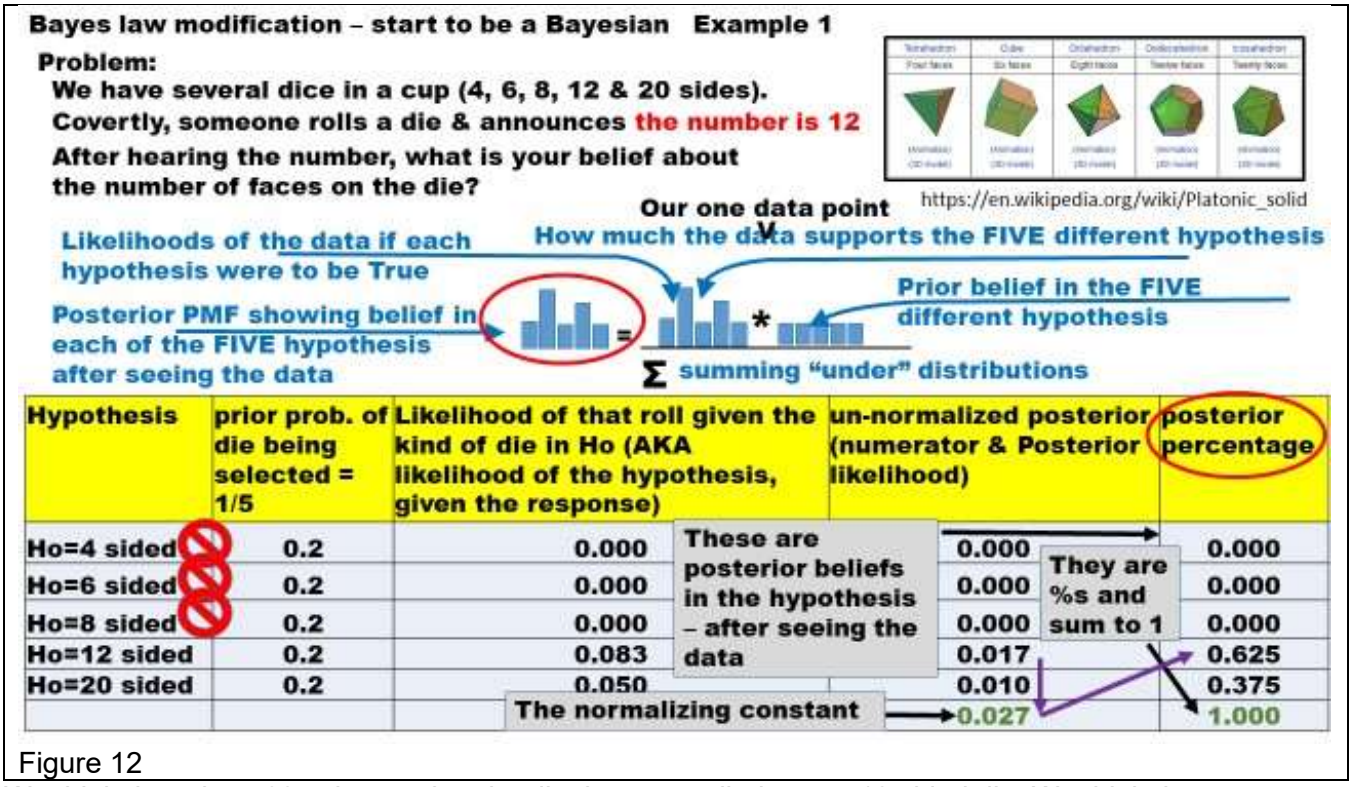


Figure 12

We think there is a .625 chance that the die that was rolled was a 12 sided die. We think there was a .375 probability that the die that was rolled was a 20 sided die.

The 12 sided die only has 12 possible outcomes. The 20 sided die has 20 possible outcomes. If we roll a number that can be produced by either of the two die it's more likely to have come from the die with a fewer number of possible outcomes. That's the end of this example and will do a few more.

Example 2 (exploring the likelihoods)

Example 2, shown in Figure 13, uses the same "set up" but the difference is that the number rolled is a 4.

A 4 sided die has a 25% chance of rolling a 1, 2, 3 or 4. A 6 sided die has 6 possible outcomes and has a lower likelihood for each of those possible outcomes.

The same logic applies to 8, 12 and 20 sided die.

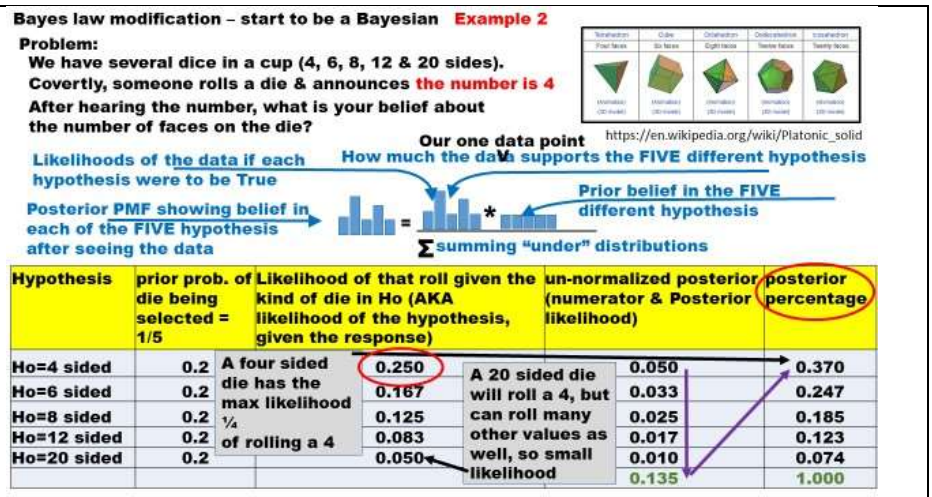


Figure 13

It is suggested that a reader might work through the calculations on this page. The steps in the table can be mapped directly to the graphic that illustrates the Bayes' Law formula.

Example 3 (exploring the likelihoods)

In Figure 14, we add a little complexity to the likelihoods.

We roll two die and see that the numbers are a 4 and a 7. If these are independent rolls then we can multiply the probability of each roll to get the probability of the two different rolls.

If the 4 sided die is picked the probability of a 4 is .25. If the 4 sided die is picked the probability of a 7 is 0.

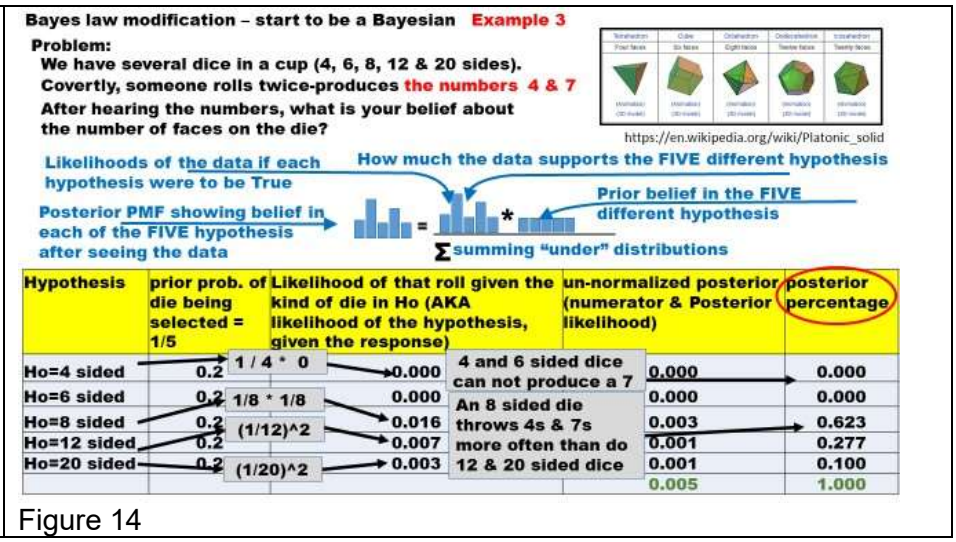


Figure 14

The probability of a 4 and a 7 is the product of those two probabilities and is 0. This logic can be applied to calculate the likelihoods for all of the other die. After calculating the normalizing constant the un-normalized posterior numbers can be changed into posterior probabilities by dividing by the normalizing constant and we can then talk about the chance of each of the hypotheses being true.

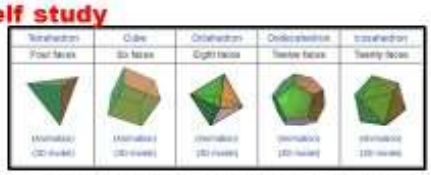
This particular normalizing constant is very easy to calculate, but it *is* an extra step and the relative numbers between the two right columns in Figure 14 are the same except for a multiplicative constant. If we wanted to save a little time we could make excellent decisions without having to calculate the posterior as a percentage.

Example 4 (exploring the likelihoods)

Figure 15 is included for self-study. It is hoped that a reader might reproduce the numbers under the posterior percentage column and record any thoughts about what they see. One thing to consider is why the posterior for the hypothesis that the die had 12 sides is so much larger than the posterior probability that the die had 20 sides.

Bayes law modification – start to be a Bayesian **Example 4- self study**

Problem:
We have several dice in a cup (4, 6, 8, 12 & 20 sides).
Covertly, rolls twice produces the numbers 11 & 12
After hearing the number, what is your belief about the number of faces on the die?



https://en.wikipedia.org/wiki/Platonic_solid

Our one data point
Likelihoods of the data if each hypothesis were to be True
How much the data supports the FIVE different hypothesis
Prior belief in the FIVE different hypothesis
Posterior PMF showing belief in each of the FIVE hypothesis after seeing the data
Σ summing "under" distributions

Hypothesis	prior prob. of die being selected = 1/5	Likelihood of that roll given the kind of die in Ho (AKA likelihood of the hypothesis, given the response)	un-normalized posterior (numerator & Posterior likelihood)	posterior percentage
Ho=4 sided	0.2			0.000
Ho=6 sided	0.2			0.000
Ho=8 sided	0.2			0.000
Ho=12 sided	0.2			0.735
Ho=20 sided	0.2			0.265
				1.000

Your thoughts:

Your comment:

Figure 15

The stated goal for this section was to show some examples that were half way between the way a frequentist would use Bayes' Law and the way a Bayesian would use Bayes' Law. In line with that goal we have a few more examples that bring distributions into the calculations.

Example 5 (Bringing formulas into the likelihoods)

Figure 16 changes the story. Assume we are working in a factory that has three machines that all make the same product. One machine is old and is scheduled for repair in the next fiscal quarter. However; right now it is producing 18% bad parts. We also have a well-maintained machine that produces 10% bad parts. We also have a new computerized machine that produces 1% bad parts.

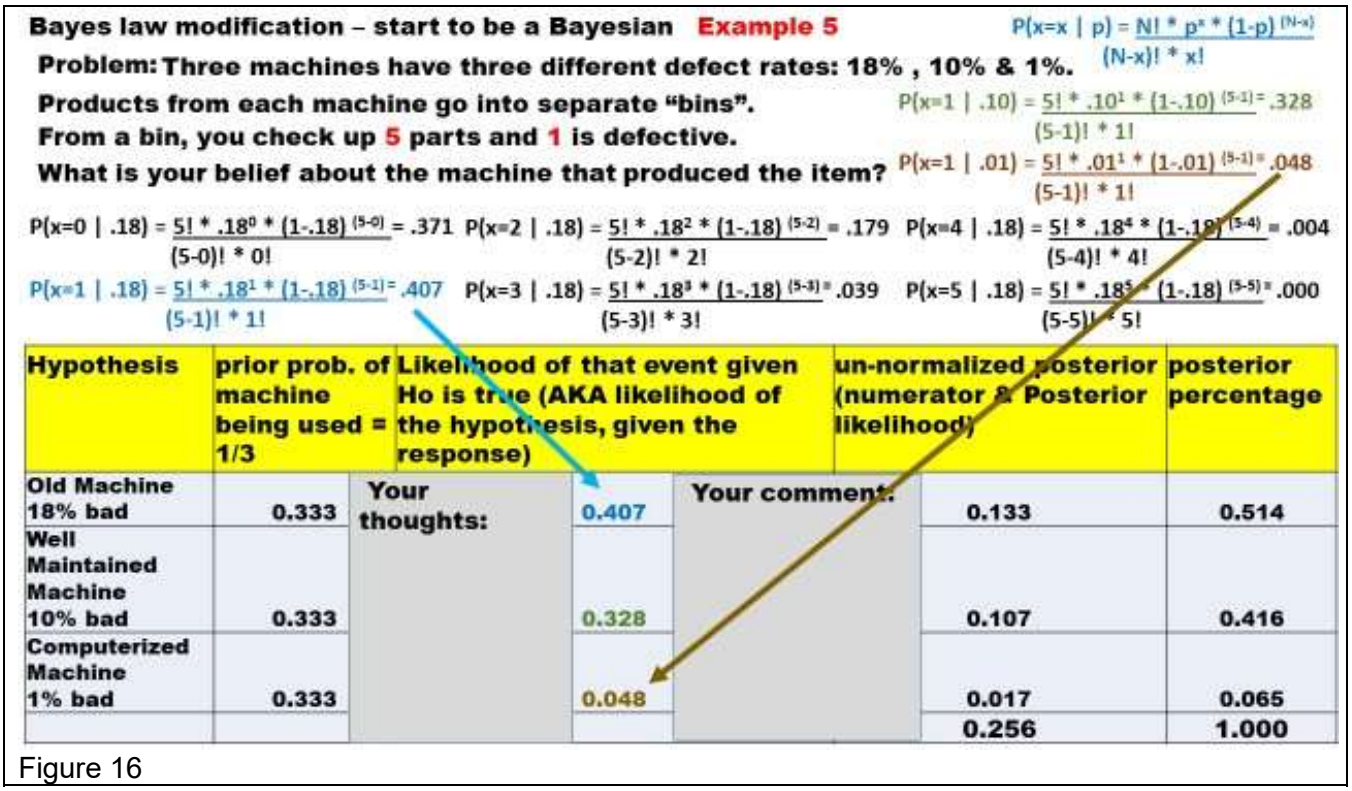


Figure 16

Imagine as parts are produced they get put into a bin so that if you find a bin you can be sure that all the parts and that bin came from one machine. Also assume that machines produce parts at the same rate so each machine should have the same number of bins randomly distributed throughout the factory.

We have three hypotheses. The three hypotheses are that the part came from the 18% machine, the 10% machine or the 1% machine. Assume that you go to a bin, and pick up five parts, and find that one of them is defective.

What is the probability that that bin was produced by each of the three machines?

We're going to say the priors are all .33. We have no idea, from looking at a bin, what machine produced the parts in the bin we just sampled.

We use the binomial formula (see upper right corner of Figure 16) to calculate the likelihoods. Figure 16 shows all the possible calculations for the 18% machine and just the calculations that were used in the table for the 10% bad and 1% bad machines.

Here is the important concept. If the Ho that the parts came from the 18% machine is true, the likelihood of one bad in a sample of five (the likelihood of our data given that Ho is true) is .407.

If the Ho of the parts coming from the 10% machine is true the likelihood of the data we saw (one bad out of five) is .328.

If the Ho of the parts coming from the 1% machine is true the likelihood of the data we saw (one bad out of five) is .048.

It is hoped that the reader might reproduce the numbers in the figure above.

Example 6 (Bringing formulas into the likelihoods)

Figure 17 reuses the same story about 3 machines in a factory and changes the number of defects to three in five. It is hoped that the reader will work through the calculations and make comments about what they notice.

We were able to solve this problem because we had three machines and a simple likelihood formula. When dealing with continuous distributions the calculations will be more complicated.

Bayes law modification – start to be a Bayesian **Example 6**

$$P(x=x \mid p) = \frac{N! * p^x * (1-p)^{(N-x)}}{(N-x)! * x!}$$

Problem: Three machines have three different defect rates: 18% , 10% & 1%.

Products from each machine go into separate "bins".

From a bin, you check up 5 parts and 3 are defective.

What is your belief about the machine that produced the item?

$$P(x=3 \mid .10) = \frac{5! * .10^3 * (1-.10)^{(5-3)}}{(5-3)! * 3!} =$$

$$P(x=3 \mid .01) = \frac{5! * .01^1 * (1-.01)^{(5-3)}}{(5-3)! * 3!} =$$

$$P(x=0 \mid .18) = \frac{5! * .18^0 * (1-.18)^{(5-0)}}{(5-0)! * 0!} = .371 \quad P(x=2 \mid .18) = \frac{5! * .18^2 * (1-.18)^{(5-2)}}{(5-2)! * 2!} = .179 \quad P(x=4 \mid .18) = \frac{5! * .18^4 * (1-.18)^{(5-4)}}{(5-4)! * 4!} = .004$$

$$P(x=1 \mid .18) = \frac{5! * .18^1 * (1-.18)^{(5-1)}}{(5-1)! * 1!} = .407 \quad P(x=3 \mid .18) = \frac{5! * .18^3 * (1-.18)^{(5-3)}}{(5-3)! * 3!} = .039 \quad P(x=5 \mid .18) = \frac{5! * .18^5 * (1-.18)^{(5-5)}}{(5-5)! * 5!} = .000$$

Hypothesis	prior prob. of machine being used = 1/3	Likelihood of that event given Ho is true (AKA likelihood of the hypothesis, given the response)		un-normalized posterior (numerator & Posterior likelihood)	posterior percentage
Old Machine 18% bad	0.333	Please fill in the missing numbers:		0.133	
Well Maintained Machine 10% bad	0.333				.170
Computerized Machine 1% bad	0.333		0.000		
				0.16	1.000

Your comment:

Figure 17

Example 7 (Priors and formulas into the calculations)

Figure 18 is a slight modification of Figure 17. In this example we say that the machines do not produce at the same rate. The computerized machine produces 50% of the parts found in the factory. The well-maintained machine produces 35% of the parts. The old machine produces 15% of the parts. It is hoped that the reader will work the calculations and pencil in their insights from the process.

Bayes law modification – start to be a Bayesian Example 7

$$P(x=x \mid p) = \frac{N! * p^x * (1-p)^{(N-x)}}{(N-x)! * x!}$$

Problem: Three machines have three different defect rates: 18% , 10% & 1%.

Products from each machine go into separate "bins".

From a bin, you check up 5 parts and 3 are defective.

What is your belief about the machine that produced the item?

$$P(x=3 \mid .10) = \frac{5! * .10^3 * (1-.10)^{(5-3)}}{(5-3)! * 3!} = .$$

$$P(x=3 \mid .01) = \frac{5! * .01^1 * (1-.01)^{(5-3)}}{(5-3)! * 3!} = .$$

$$P(x=0 \mid .18) = \frac{5! * .18^0 * (1-.18)^{(5-0)}}{(5-0)! * 0!} = .371 \quad P(x=2 \mid .18) = \frac{5! * .18^2 * (1-.18)^{(5-2)}}{(5-2)! * 2!} = .179 \quad P(x=4 \mid .18) = \frac{5! * .18^4 * (1-.18)^{(5-4)}}{(5-4)! * 4!} = .004$$

$$P(x=1 \mid .18) = \frac{5! * .18^1 * (1-.18)^{(5-1)}}{(5-1)! * 1!} = .407 \quad P(x=3 \mid .18) = \frac{5! * .18^3 * (1-.18)^{(5-3)}}{(5-3)! * 3!} = .039 \quad P(x=5 \mid .18) = \frac{5! * .18^5 * (1-.18)^{(5-5)}}{(5-5)! * 5!} = .000$$

Hypothesis Machines now produce different numbers of parts	prior prob. of machine being used = 1/3	Likelihood of that event given Ho is true (AKA likelihood of the hypothesis, given the response)		un-normalized posterior (numerator & Posterior likelihood)	posterior percentage
Old Machine 18% bad	0.15	Please fill in the missing numbers:	Your comment:	.00585	
Well Maintained Machine 10% bad	0.35				.324
Computerized Machine 1% bad	.500		0.000		
				0.085	1.000

Figure 18

Example 8 (the Monty Hall problem):
This example is intended to illustrate some of the difficulties in calculating likelihoods.

Monty Hall ran this game on TV for years. The problem is as follows. There are 3 doors (A, B and C) behind which are 2 goats and a car. You want to pick the car and you pick a random door (you pick A).

Monty opens another door, door B, and shows you a goat. He then offers you the chance to switch to door C or stay with door A. Which door should you select, A or C.

Bayes law modification – start to be a Bayesian Example 8 The Monty Hall Problem

The formula is not useful unless you know how to interpret the words: *Likelihoods*

Problem: There are 3 doors, behind which are two goats and a car.
You want the car and pick a random door (call it door A).
Monty Hall opens Door B and shows a goat.

He offers you the chance to switch to door C or continue with door A.

Your Decision: Which door should you select, A or C?

Rules:

You Pick a door. Game manager always offers the chance to switch doors.
Game manager always opens a door you did not pick.
Game manager always opens a door hiding a goat and never the car.

the non-random choice provides information,

Cognitive psychologist Massimo Piattelli-Palmarini said "no other statistical puzzle comes so close to fooling all the people all the time", and "even Nobel physicists systematically give the wrong answer, and that they *insist* on it,
and they are ready to berate in print those who propose the right answer".

In 1990 by Marilyn vos Savant (I.Q. measured at 220 +) published the answer.
Prompted 10,000 letters (many from Ph.D.s) saying she was wrong.

Figure 19

Our goal is to fit this problem into the Bayesian structure we've been using for the last several examples. As a warning, precise statements will be very helpful in getting the correct answer.

We need priors and those are easy and all 1/3. We also need the likelihood of the "new data" given that hypothesized prior is true. It is worth spending a minute to try and be very clear on what is the new data. It's the fact that you have selected door A and door B has been opened.

If the Ho that A hides the car is true: If the car is behind A, then Monty can open either B or C with a 50% chance. If the car is behind A, and you select A, the likelihood of Monty opening door B is 50%. If the car is behind the door A and you select A, the likelihood of the data we saw (Monty's action) is 50%.

If the Ho that B hides the car is true: If the car were behind door B, Monty could never open door B – that would end the game. The likelihood of Monty opening door B (the data we saw), if the car is behind door B, is 0.

If the Ho that C hides the car is true: If you select A and the car is behind C, Monty cannot open either door A or door C. If you select A and the car is behind C, Monty must open B 100% of the time. If you select A and the car is behind C, there is a 100% likelihood of seeing the data we saw.

Bayes law modification – start to be a Bayesian Example 8 The Monty Hall Problem

(Ho: C is true) **How likely is "You select A & Monty selects B" if the car is behind C?**
If you select A and the car is behind C(Ho: C is true), Monty cannot open doors A or C.
If you select A and the car is behind C(Ho: C is true), Monty must open B (100%).
If you select A and the car is behind C (Ho: C is true), there is a 100% likelihood of seeing the data we saw (Monty must open door B).
The car could be behind Door A or Door C. We can now apply Bayes' formula.
Summary: Just knowing the formula, alone, is of little use (unless the problem is trivial).
We need the insights, and English interpretations we developed, to give us any chance of recognizing formula components and solving a problem.

Hypothesis	prior prob. of car behind door = 1/3	Likelihood of that event given Ho is true (AKA likelihood of the hypothesis, given the response)	un-normalized posterior (numerator & Posterior likelihood)	posterior percentage
Door A hides car	1/3	1/2	1/6	1/3
Door B hides car	1/3	0.000	0.000	0.000
Door C hides car	1/3	1.000	1/3 = 2/6	2/3
We needed different logic for the different likelihoods.			3/6	1.000

(1/6) / (3/6)
(2/6) / (3/6)

The problems will get worse (math and logic) when the distributions are continuous.

Figure 20

Example 9 (the mammogram problem)

This example is a repeat of an example done very early in this booklet.

It is only included so that we can take that problem and put it into the structure we've been using to solve the previous problems. I think the structure is powerful.

The Hypotheses are cancer or not. The new data is a test result: positive or negative.

Bayes law modification – start to be a Bayesian Example 9 The Mammogram Problem,

I just want to reproduce this problem in the structure we have been using.
A mammogram has an 87 % chance of detecting cancer if it is present
A mammogram signals a false positive 12% of the time.
About 1% of women between 40 and 60 have breast cancer.
What is the chance that a woman, who has a positive test, has cancer.

$$P(C \mid \text{Positive Test}) = \frac{P(\text{Positive Test} \mid \text{Cancer}) * P(\text{Cancer})}{P(\text{Positive Test})}$$

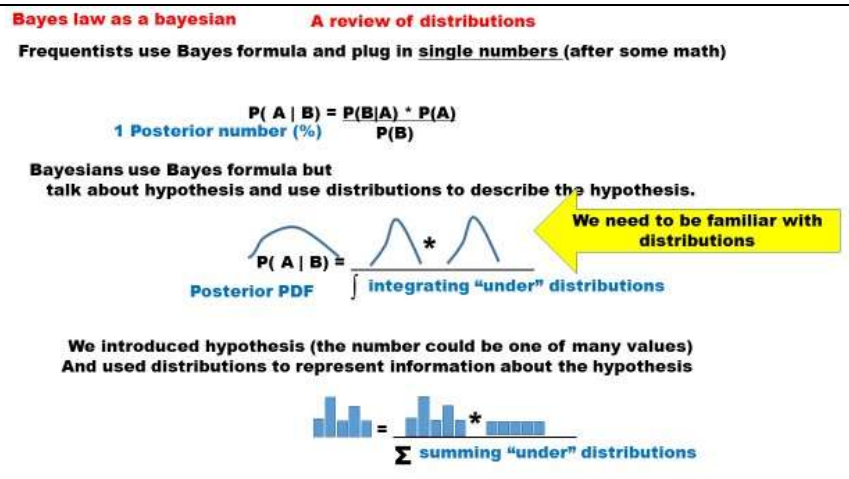
$$P(C \mid \text{Positive Test}) = \frac{.87 * .01}{(.01*.87) + (.99*.12)} = \frac{.0087}{0.0087 + 0.1188} = \frac{.0087}{.1275} = .0682 \text{ or } \sim 7\%$$

Hypothesis	prior prob. of cancer in population	Likelihood of that event given Ho is true (AKA likelihood of the hypothesis, given the response)	un-normalized posterior (numerator & Posterior likelihood)	posterior percentage
Has Cancer	.01	.87	.0087	.0682
No Cancer	.99	0.12	.1188	.9317
			.1275	1.0000

Figure 21

The likelihoods are the probabilities of the data (the test result) given the hypotheses are true. The un-normalized posteriors are the result of multiplications and are numerators in the Bayes' Law equation. The numerator, in the worked out example, is the chance of a positive test and cancer. This example makes it easy to see that the denominator, the normalizing constant, is all the ways we could get a positive test. I leave it to the reader to study this if they choose.

BAYES' LAW AS A BAYESIAN: A REVIEW OF DISTRIBUTIONS

Figure 22 is a transition slide to help us move from one topic to another. It simply reminds us that Bayesian's, when they use Bayes' Law, use distributions as the components of the formula. With that in mind this next section will be a review of the simpler distributions commonly encountered in Bayesian analysis	 Figure 22

There is also a subtle change in the problem being analyzed. In the previous section we were trying to update our relative beliefs in some small number of discrete hypotheses. If we are doing real Bayesian analysis we are going to be analyzing a hypothesis about some characteristic of a distribution – perhaps about the mean.

The result of the Bayesian analysis is a posterior distribution that describes our belief **about the parameters (mean, variance etc.) that describe** a distribution of data. We will be doing a multi-step process and sentences get to be long and complex (and a bit confusing). This is just a warning of what's to come and hopefully things will be clearer when we do some examples.

The multiplication, shown in the graphic in Figure 22, can be difficult. The Exponential family of distributions is important in Bayesian analysis because the only closed form (think simple math) solutions to using Bayes' Law to update our prior beliefs are found using distributions from the Exponential family.

To be a little bit formal: If we assume a parameter η, a distribution is a member of the exponential family

if the density (relative to η) is of the form $\mathbf{p(x|\eta) = h(x)\ exp\ \{\ \eta^T T(x) - A(\eta)\}}$

The members of the exponential family of distributions are:

Normal	Exponential	Gamma	Chi-squared	Beta
Dirichlet	Bernoulli	Categorical	Inverse Wishart	Wishart
Geometric				

the binomial with a fixed number of trials
the multinomial with a fixed number of trials
the negative binomial with a fixed number of failures.

Think again of the graphic in Figure 22 that illustrates the formula for Bayes' Law. There is a term that arises quite often in a discussion of Bayesian analysis and that term is "conjugate prior".

The books say a conjugate prior is a distribution, which when multiplied by the likelihood and then normalized, produces a closed form posterior distribution function which is of the same family as the prior.

The idea for the conjugate prior is that a mathematician started with a formula for a prior, and a type of data that can be used for the updating, and derived a formula that produces a posterior distribution that:
is of the same form as the prior (no one cares about this except the math-ies)
and has a simple formula (we all care about this because it means the answer is easy to calculate).

This means, in simple terms, if your prior is normal and you update it with data you get a normal posterior with a simple formula. Or; if your prior is a beta and you update it with data you get a beta distribution with a simple formula. An example of a conjugate prior and posterior is below.

Coin flips generate a percentage of heads so we will show the conjugate priors for that type of problem. The formulas look similar and are simple to update. Conjugate formulas look like the formulas below.
If the prior formula is $= p^{\alpha} * (1-p)^{\beta}$ and the posterior formula is $= p^{\alpha+k} * (1-p)^{\beta+n-k}$
K is the number of heads in your new data and n-k is the number of tails in your new data.

There are several conjugate priors that have been worked out. A good high–level overview can be found at the wiki page for conjugate priors (https://en.wikipedia.org/wiki/Conjugate_prior. There is a table at the bottom of that webpage that gives a very high–level overview of the known conjugate priors and how you would use these to do a Bayesian update. Deriving formulas can be tough.

Since all conjugate priors are from the Exponential family, let's take a minute to explore links between 3\three members of the exponential family: Poisson, Exponential and Gamma

Poisson: A Poisson distribution can model the number of events occurring in a fixed unit (time period, area, volume). The mean number of events per "unit" is called lambda (λ). For simplicity, assume λ is events per unit time.

Exponential: If the number of events per unit time can be described with a Poisson distribution, the time between any two events can be described using an Exponential distribution. The average time between any two events, for the Exponential, is $1/\lambda_{poisson}$. $1/\lambda_{poisson}$. is often described as the time to the next event (sometimes called Θ.

Gamma: The Gamma distribution is used to model the time between a starting time point and the n^{th} event. If the event you are searching for is the 1^{st} event, the Gamma collapses to the Exponential. The Gamma distribution has two parameters (α and β) and the intuition is that α is the number of events we are waiting for and β is the average time between events or $1/\lambda_{poisson}$.

The formulas for these three distributions are shown to the right. They are very similar. You can see how these formulas are related and imagine, if you're really good at algebra, making relationships between these three formulas might be easier than making relationships between unrelated formulas.	$\text{Poisson} = P(X=k) = \frac{\lambda^k * e^{-\lambda}}{k!}$ $\text{Exponential} = P(X=k) = \frac{\lambda^k * e^{-\lambda k}}{k!}$ $\text{Gamma} = P(X=k) = \frac{\lambda^k * e^{-\lambda k} * x^{k-1}}{\Gamma}$ Figure 23

A CONCEPTUAL REVIEW OF THE BERNOULI DISTRIBUTION

The Bernoulli distribution is a description of the outcome of a random variable which can only take the values of one (with probability equals P) and is 0 (with probability equals 1 minus P).

An example of this is one coin toss, or one at-bat in baseball, or one free throw in basketball. The Bernoulli distribution is the basis for the binomial distribution.

A CONCEPTUAL REVIEW OF THE BINOMIAL DISTRIBUTION

The binomial distribution is often used to model the number of successes, in a sample of size n, drawn with replacement from a group much larger than n.
The formula can best be understood using a decision tree like the one on the right-hand side of Figure 24.

$P^k * (1-P)^{n-k}$ is the probability associated with any one path through the decision tree with K successes and (n - k) failures

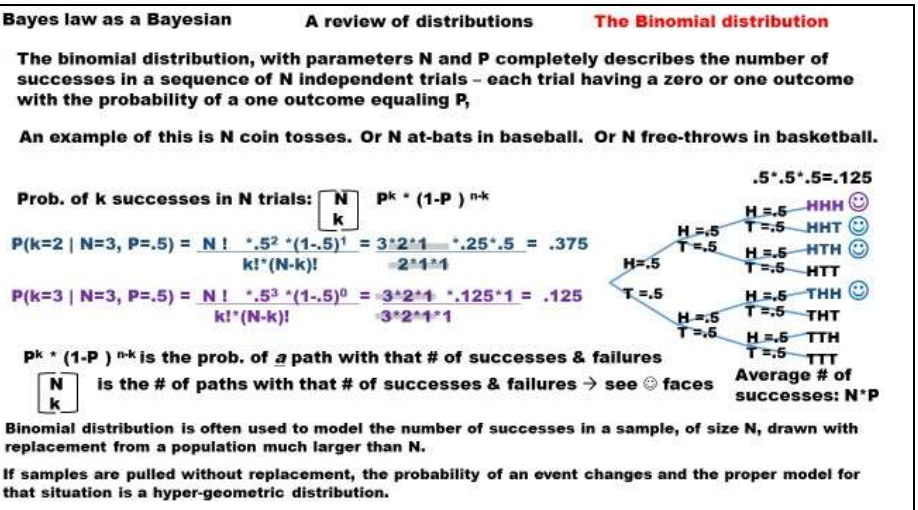


Figure 24

The factorial part of the calculation is used to calculate the number of different paths through the decision tree that have k successes and (n-k) failures.

The calculation of the factorials, for small problems, is simple because cancellations can make the formula simple. However, when the decision tree gets to be large, and one is doing the calculation by hand, the calculation the factorial can be quite a problem.

Because a lot of work using the binomial distribution was done in the days before computers, and even calculators, there are several formulas for approximations to the binomial that are simpler to calculate by hand. These approximations have become less interesting to practitioners since the advent of powerful and cheap computers. However; there is still are of interest to mathematicians because they allow simpler formulas to be substituted when doing of long and complicated derivation.

A CONCEPTUAL REVIEW OF THE NORMAL DISTRIBUTION

The Normal curve is completely described by the mean and standard deviation. The Normal is an integral part of every stat 101 class.

An interesting thing, in the context of this seminar, is that the formula for the PDF has a normalizing constant. This concept has been mentioned before.

The numerator in the formula (the e to the power) gives the curve its shape.

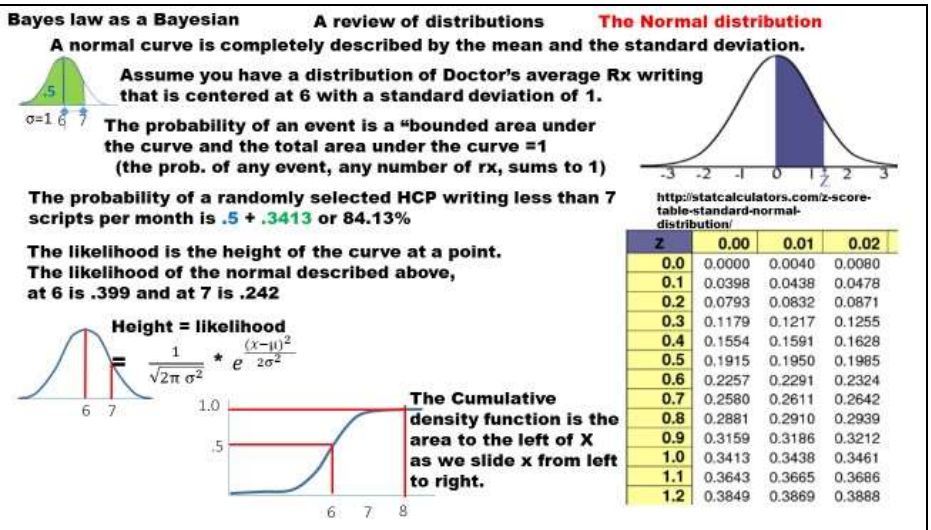


Figure 25

The denominator is a normalizing constant and makes the area under the curve integrate to one. If the area under the curve integrates to one, we can think of calculated areas as probabilities – because probabilities must sum to one.

A CONCEPTUAL REVIEW OF THE BETA DISTRIBUTION

Figure 26 shows graphics about the beta distribution.

The x-axis in the plots is a percentage.

Beta PDF =

$$\text{Prob}(X=x|\alpha\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha\beta)}$$

Where B in the denominator is the beta **function**

$$B(\alpha\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt$$

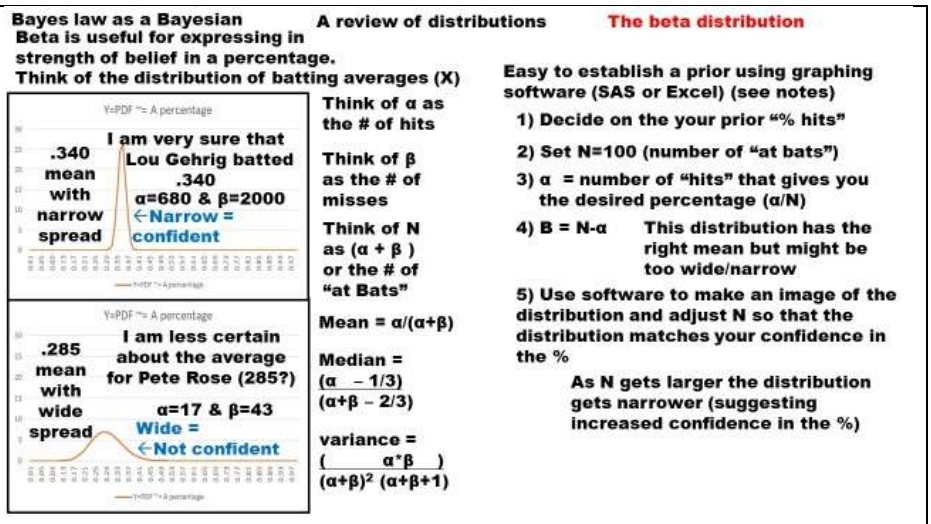


Figure 26

It's very confusing that the word "beta" has three common, very different, uses in this one formula alone..

1st: Beta" is the commonly used name of the PDF & CDF probability ***distributions***
2nd:B(α,β) "Beta" is the name of the ***function*** used in the denominator of the density BETA (PDF) formula
3rd: Beta" is the name of the second ***parameter*** in the density function

People sometimes get lazy and just say beta. If you are familiar with the problem it's not confusing. If you're new to the problem it can be very confusing.

The beta distribution is used to explain to a client, or get from a client, a prior belief about a percentage. Please look at Figure 26 and note that the x-axis is a percentage.

The top chart illustrates a prior belief about Lou Gehrig's batting percentage. I'm pretty sure that the average is .340. I can use a chart of the beta distribution to express my belief about the average and how certain I am in my belief. I think the average is .340 and I'm pretty sure of that. I make my distribution centered at .340 and narrow. I have an xls sheet (and also a SAS program) to make plots of the Beta PDF and allow you to adjust the α and β parameters until the curve has the proper mean and spread – a mean and spread that matches your prior belief about a percentage.

I'm not very sure about Pete Rose's batting average. He was a good player and I'm thinking he's close to .300 lifetime but I'm not very sure at all. It could be higher or lower and I can use a beta distribution to draw a picture that helps me explain, to others and myself, my belief in the average and my confidence about that average. The picture above, with its wide spread, conveys that I'm not very confident about my estimate.

Now that I'm creating this booklet, and really looking at the X axis, I realized that I am more confident than the picture shows. The distribution touches y=0 at about .130 and .450 and I should change that.

I'm pretty sure that Pete Rose was not a .130 hitter because he was famous for his ability to hit the ball. I'm also pretty sure he wasn't a .450 hitter lifetime because no one was a .450 hitter.

The right-hand side of Figure 26 explains how to make the adjustments to a beta distribution to reflect your belief in a percentage.

The 1st step is to decide on the percentage that you think is the center of the distribution.

The 2nd step is to set N= 100 and α to be the percentage you want. α / N will be the center of the beta distribution. We now have the center of the distribution and we need to adjust the width.

Keeping with the idea of a baseball player and using software to create pictures of the distribution:
α (alpha) is the number of hits the player got.
N is the number of at-bats and equals alpha plus beta.
β (beta) is the number of misses or "at bats without a hit" and is N-α.
As N gets larger the distribution gets narrower. To adjust your confidence in your expected percentage you adjust N, up and down, keeping the ratio of α / α+β (or α / N) a constant.

Since I think my spread for the Pete Rose distribution is too wide I will increase the value of N and watch the distribution get "skinnier".

Figure 27 shows output from a SAS program (that will be included with the seminar materials).

It loops over a couple values of alpha and beta and makes plots.

The beta distribution is very flexible.

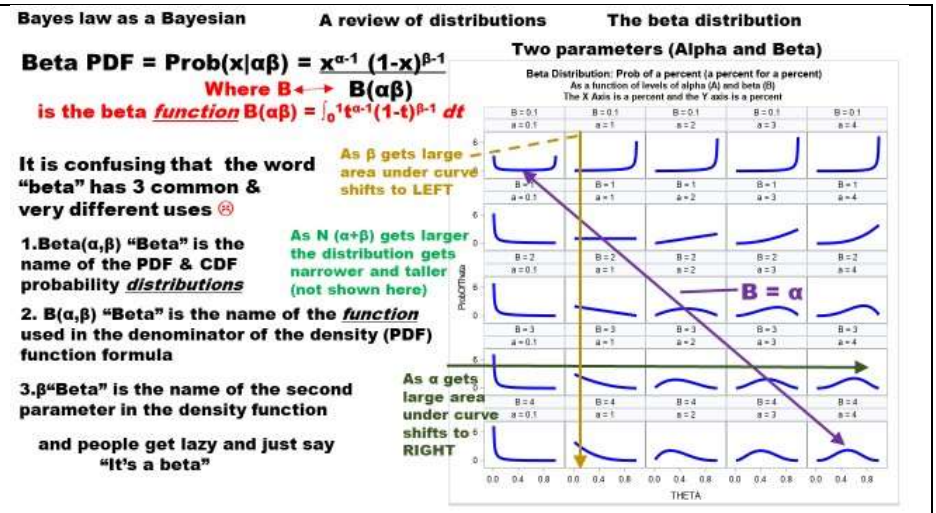


Figure 27

In general, as the beta parameter gets large the area under the curve shifts to the left (missing the ball lowers your %). As the alpha parameter gets large the area under the curve shifts to the right (hitting the ball increases your batting percentage). When alpha and beta have the same value the distribution tends to be symmetric.

There's another thing to learn if we want to use the beta distribution to describe our prior belief (mean and spread) in a percentage. There are times when we have no idea what that percentage might be. We saw this phenomenon, no real idea what the prior parameter should be, in the dice example. We had no idea what the probability of picking any particular dice would be so we assigned them all the same probability.

Figure 28 provides information on how to set parameters for a beta distribution when we have no idea what the proper percentage should be.

One common setting is for a Jeffrey's prior, where alpha and beta both equal .5.

Another common setting is for what is called an uninformative prior and has alpha and beta both equal to 1.

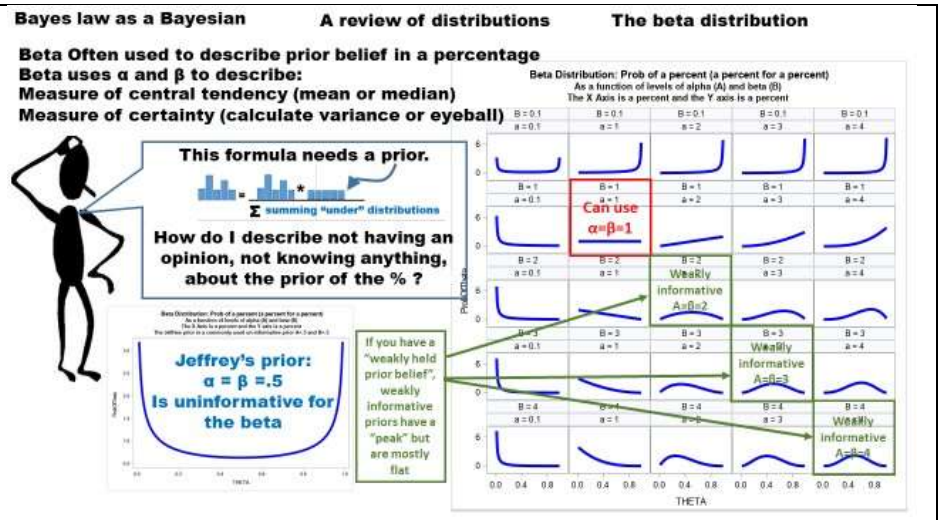


Figure 28

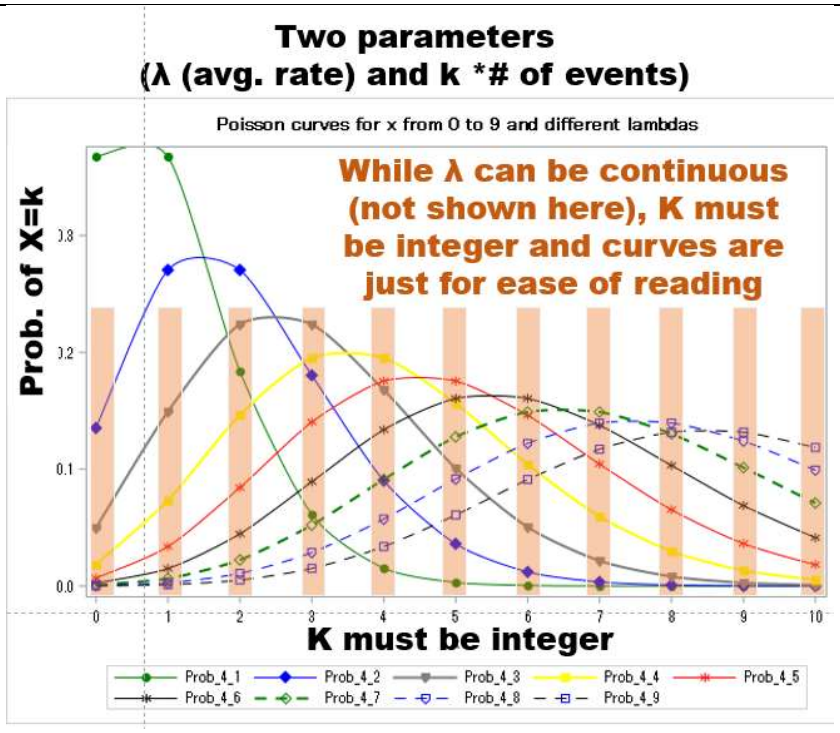
There is one more concept about concerning priors – the weakly informative prior. The weakly informative prior has its center at the percentage we think is most likely but has a wide spread. The distribution in the lower right hand corner of Figure 28 shows a weakly informative prior centered at .5.

A CONCEPTUAL REVIEW OF THE POISSON DISTRIBUTION

The Poisson distribution describes the probability of a number of events, k, occurring in a fixed "interval" of time or area or volume. λ is the average rate and also the variance of the Poisson distribution.

The PDF is:

$P(X=k) = \frac{\lambda^k * e^{-\lambda}}{k!}$ and can be read as the probability of the random variable X having the value k

Since k counts events it must be an integer. λ is the average number of events per unit interval. λ might be the number of cars per hour. If λ is 5 cars per hour it is also 2.5 cars per half-hour. If λ is 18 "metal finish blemishes" in a square yard, λ is also two "metal finish blemishes" in a square foot. λ is often called a rate parameter because it describes the rate at which events occur. While K must be an integer, λ does not have to be an integer. .	 Figure 29

If lambda is small, it's very possible to have 0 events in your interval. If lambda is describing accidents at a particular traffic intersection, and lambda is one per week, it's very likely that it particular week will have 0 accidents. It is not possible to have negative accidents. It is assumed that there is no upper limit to the number of events, though the real world only meets this approximately.

Figure 29 shows continuous curves for all of the different values of λ but this is just to make it easier to read the chart. The fact that k must be integer is emphasized by the dots on the lines showing where k is integer and also by the tan colored vertical bars showing where k is an integer.

Conditions for use of a Poisson are:
Events occur at a known constant rate, λ, which is the average number of events per "interval".
It is assumed that lambda does not change in the "analysis time.
Events are independent of each other and of the time since the last event.
K is an integer and can take on values from 0 to positive infinity. Events are counted in integer numbers. Something to think about, as you remember the binomial distribution, is that in a Poisson problem we can count the number of events but we cannot count, or predict the, the number of "events that did not occur".

Figure 30 shows the Poisson distribution for small, non-integer values.

Note that as λ get smaller the probability of 0 occurrences increases.

Also note that as lambda gets large, and large is only 5 or so, the Poisson approximates the normal distribution.

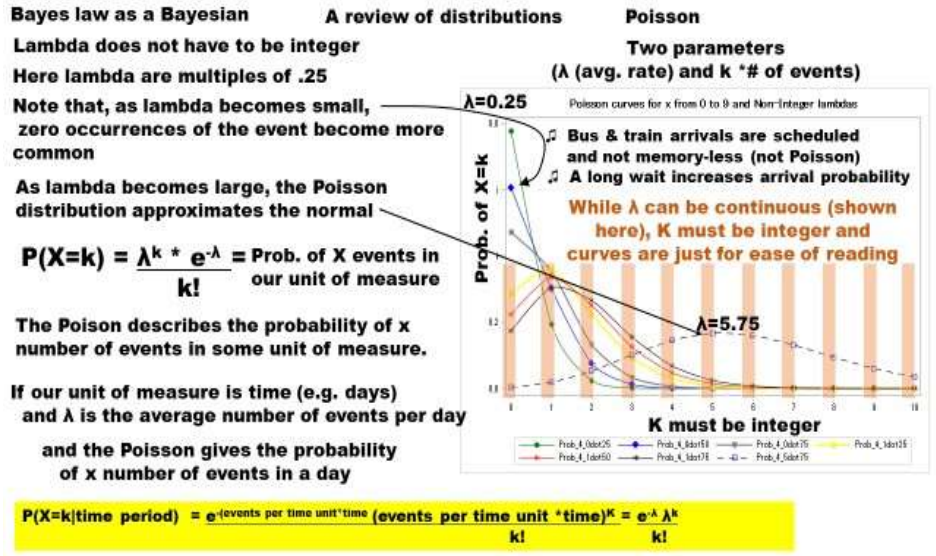


Figure 30

Figure 31 is a bit of a repeat of the information in Figure 30.

It just cleans up the graphic a bit and, hopefully, better shows the effect of changing λ on the distribution.

While the points are connected by smooth curves, the Poisson distribution is only valid for integer values on the x-axis.

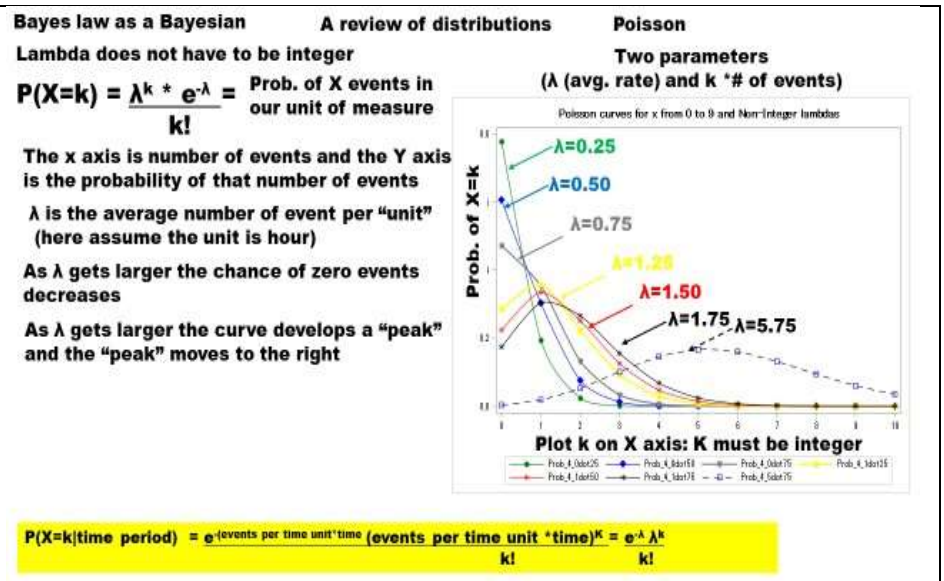


Figure 31

The yellow box, in the bottom of Figure 31, is a small reminder of a complication that can happen with these distributions - unit inconsistency. The parameter, λ, is the number of events in a unit time. It might be the number of calls into your phone center per hour. Occasionally, people try complicate the problem by giving λ in one unit of time and observed data in another unit of time.

An example of this might be you know that you average 8 calls per hour and that is what you consider your λ. When people bring you new data they might say "today, we had 80 calls in an 8 hour day". This is a reasonable way to collect data but not in the same unit of time. You can do Bayesian analysis when people start off not using the same "unit" but you must do a conversion somewhere in your calculations.

CONCEPTUAL REVIEW OF THE EXPONENTIAL DISTRIBUTION

Bayes law as a Bayesian **A review of distributions** **Exponential**

If Poisson gives the probability of x number of events in a day ($\lambda_{poisson}$ is the avg. events/day)
Wait time *until the 1st* event follows an exponential distribution with avg. $\theta_{exponential} = 1/\lambda_{poisson}$
Use of the Poisson distribution often concerns counts ($\lambda_{poisson}$ avg. events / time period)
Exponential distribution often involves time between events ($\theta_{exponential}$ =avg. time period/event)
Bounded by 0 and infinity

Steps to estimate an Exponential
1) Record times
2) Calculate gaps
3) Calculate avg. gap
4) Lambda = 1 / average
5) Exponential PDF = $\lambda * e^{-\lambda * x}$ (when X >= 0, else 0)
6) Exponential Cuml. PDF = $1 - e^{-\lambda * x}$

Events are "Poisson $\lambda_{poisson}$" = 5 events / 10.0 min = . 5 events / min = events/ "unit"
Waits are Exponential $\lambda_{exponential}$=10.0 min / 5 events = 2.0 min wait per event

Sample data set	
Arrival time (Decimal Minutes)	Gap (in Decimal Minutes)
0.50	0.50
0.75	0.25
1.00	0.25
2.00	1.00
10.00	8.00
Exponential λ = Avg. wait/event	2.00
Poisson λ = events / min wait	0.50

use the poisson lambda PDF=$\lambda e^{-\lambda * x}$ CDF=$1 - e^{-\lambda * x}$ 0.50

Time of interest	pdf (hand)	pdf (excel)	Cuml. Prob of an event
0.00	0.4998	0.500	0.000
1.00	0.3033	0.3033	0.393
2.00	0.1839	0.1839	0.632
3.00	0.1116	0.1116	0.777
4.00	0.0677	0.0677	0.865
5.00	0.0410	0.0410	0.918
6.00	0.0249	0.0249	0.950
median = (ln 2)/ λ			1.386294

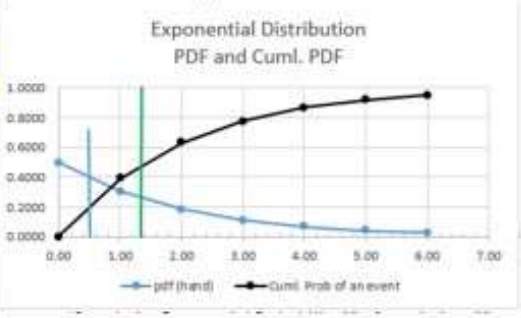


Figure 32

The Exponential distribution also shows up in Bayesian analysis. Members of the Exponential family of distributions differ because of the way the parameters are added and subtracted in the formulas. It is often the case that, if a parameter in one distribution is set to the value one, the parameter is removed from the equation and the equation then describes one of the other members of the Exponential family of distributions.

An interesting bit of logic shows the relationship between the Poisson and the Exponential. The Poisson describes the number of occurrences per unit of time (Unit can be other things as well as time but we will consider time here). Then the Exponential is a description of the length of time between occurrences.

Consider this short derivation. Assume we have a Poisson process and events occur with the rate of λ events per unit of time. Under this assumption there will be λt occurrences in every t units of time.

The Poisson distribution has the formula below.
$P(X=k) = \frac{\lambda^k * e^{-\lambda}}{k!}$ and this can be read as the probability of the random variable X having the value k

If were asking about the probability of zero events in our unit time the formula collapses to $e^{-\lambda t}$. This describes the probability of zero events in t units of time. Another way of thinking of this is that p (x=0) = $e^{-\lambda t}$ is the probability that the time to the first occurrence of an event is greater than t or using a more formal mathematical representation: $P(T > t) = P(X=0 \mid \mu = \lambda t) = e^{-\lambda t}$

We can use the above formula to calculate the probability that an event does **not** occur during t unit of time and that is: $P(T <= t) = 1 - P(x=0 \mid \mu=\lambda t) = 1 - e^{-\lambda t}$

If you take the derivative of this formula with respect to t you get the probability density function of the Exponential distribution which is: PDF(Exponential) = $\lambda e^{-\lambda t}$

This establishes the relationship between the Exponential and the Poisson that is suggested in Figure 23. The Exponential gives the interval of time between any two consecutive arrivals and not just the time to the FIRST event. This is true because the Exponential is memory-less.

The fact that the Poisson and the Exponential distributions are so related, and so commonly used in many different disciplines, can create confusion. Few books seem to bother to relate the two distributions, even though there single parameters are intimately related by being the inverse of each other. Since both of the parameters are something per something they are both often called rate parameters and often formulas show both of these rates as λ (this confused me).

Unfortunately, some disciplines are in the habit of writing the formulas in different ways. Some Exponential formulas require using the λ that you would get from the Poisson. Some formulas require using 1/ $\lambda_{poisson}$ in the Exponential formula. To clarify the issue, this booklet will show a few large scale images showing the relationship, the formulas and the calculations.

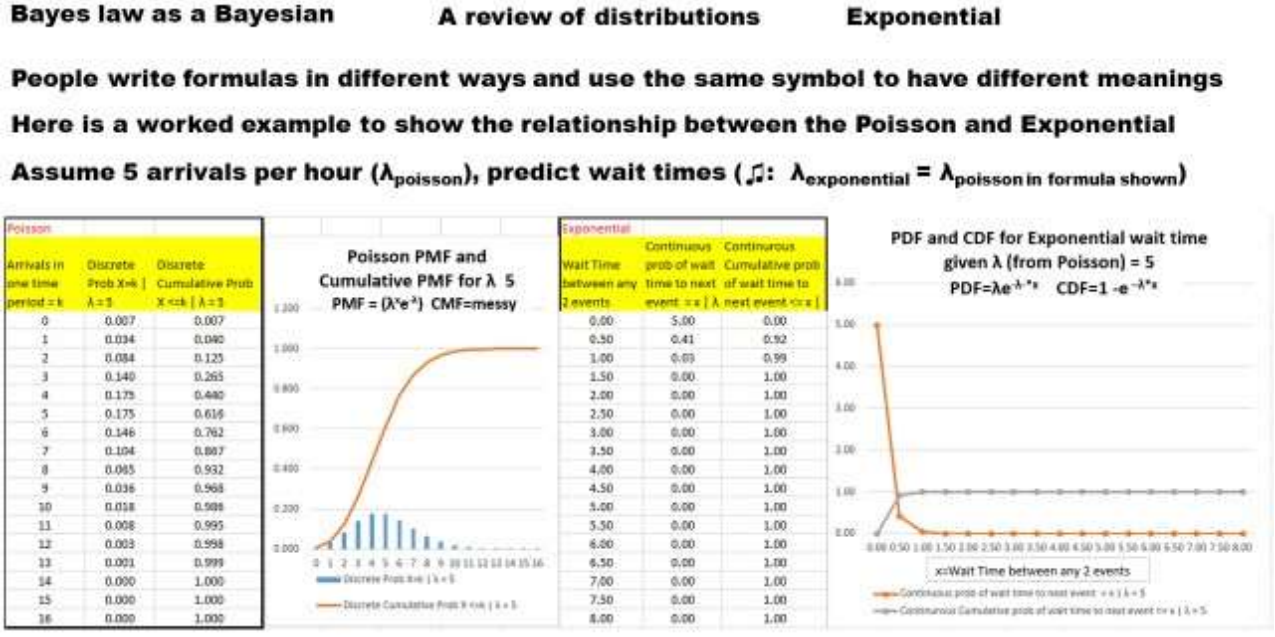


Figure 33

Figure 33 emphasizes the relationship between the Poisson and Exponential. With a Poisson λ = 5, I used Excel to generate the probabilities of counts of events. The left-hand side of the figure shows both the PMF and the CMF (this is a discrete distribution so this is a PMF). For that same λ, 5, we use the formula shown in the right-hand side to calculate the PDF and CDF for an Exponential distribution.

Figure 34, seen below, emphasizes the fact that in the Exponential formula shown in the title of that chart, we use the time gap between events as X and the lambda that is calculated from the Poisson.

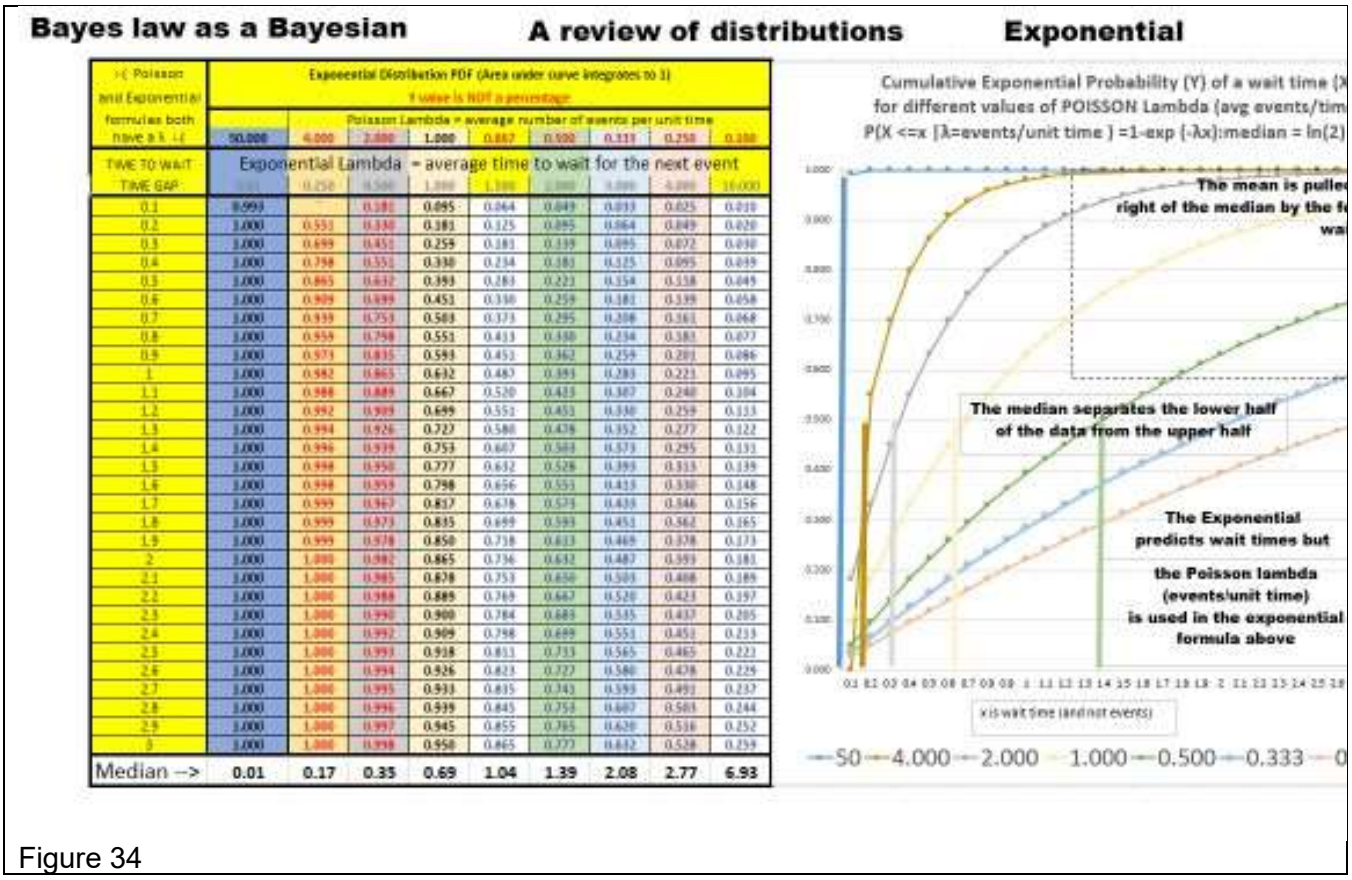


Figure 34

A CONCEPTUAL REVIEW OF THE GAMMA DISTRIBUTION

The Gamma distribution is also important in Bayesian analysis. If the Exponential distribution describes the waiting time to the 1st event the Gamma distribution can be used to describe the waiting time to the nth event.

Using the most common Bayesian formulation, the insight into the Gamma is that the α parameter is the number of events (though alpha can be fractional) we are waiting for.

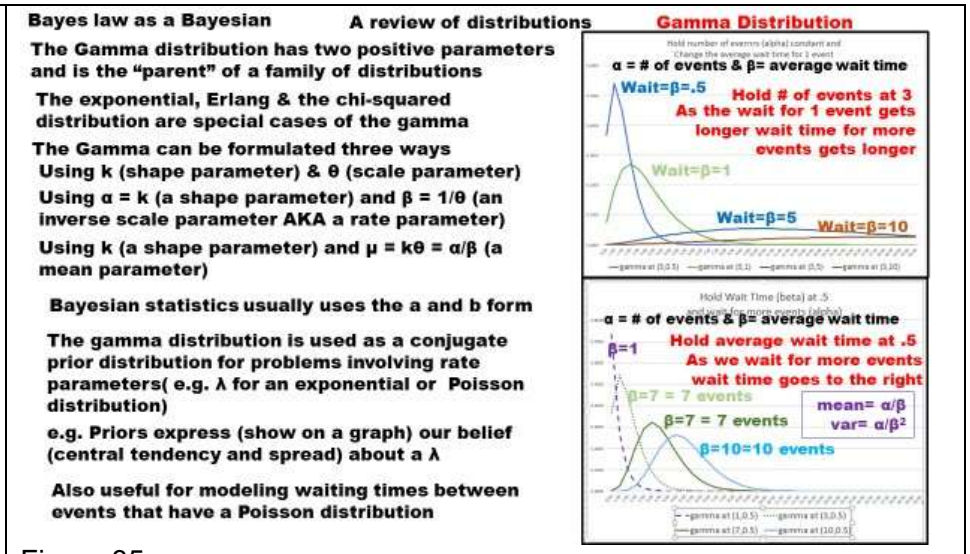


Figure 35

The β parameter is the average wait time for one event.

With that as an understanding (and formulation), it's easy to see that if either α or β gets larger the distribution will shift to the right. If either α (# of events to wait for) or β (average wait time for 1 event) gets larger, an observer must wait longer.

Unfortunately, the Gamma is a bit confusing because it can be parameterized, or formulated, in at least three different ways:

using K as a shape parameter and Θ as a scale parameter
using alpha = K as a shape parameter and β equals 1/Θ as an inverse scale/rate parameter.
using K as a shape parameter and μ=kΘ = α/β as a mean parameter.

…and other ways as well. Thankfully, most discussions in Bayesian statistics use the α and β parameterization.. This is what Excel uses

Some examples of using the Gamma shift from one parameterization to another in the same problem.

Even one Gamma formula is a bit confusing because the word Gamma is reused. The formula for the Gamma distribution has the Gamma function as a component.

$$\text{Gamma} = P(X=k) = \frac{\lambda^k * e^{-\lambda k} * x^{k-1}}{\Gamma}$$

The mean of the Gamma distribution is α/β and the variance is α/β^2 and these can be used to help create a graphic illustrating the distribution that describes a client's belief in some parameter being modeled using the Gamma. The mode is $(\alpha-1)/\beta$ for $\alpha \geq 1$. As β increases the PDF become steeper, since the variance has β^2.in the denominator.

There is quite a tangled relationship between the Gamma and several other distributions The Exponential, the Erlang and the Chi-squared distribution are special cases of the Gamma.

Additionally the Gamma function is a generalization of the factorial process. The factorial process only works for integers but the Gamma function is considered to be a parallel for positive real numbers. If you use integers to create a series of points in X, Y, for the formula y=(x-1)!, you can fit a Gamma function smoothly through all those points. Since the Gamma is valued for non-integer values of X, people consider the Gamma to be a generalization of the factorial.

EXCEL EXAMPLE: A CLOSED FORM SOLUTION FOR UPDATING A PRIOR PERCENTAGE
Figure 36 shows an image of an Excel spreadsheet that is given to seminar attendees. Is interactive and it is intended to mimic the discussion one might have with a client.

It is assumed that the client is a person with a prior knowledge of the percentage and this Excel spreadsheet is intended to help the client express that prior belief in some percentage and then to examine the issue from a few different viewpoints. It is hoped that the interactive nature of this tab on the workbook will allow the reader to become familiar with different parts of this process.

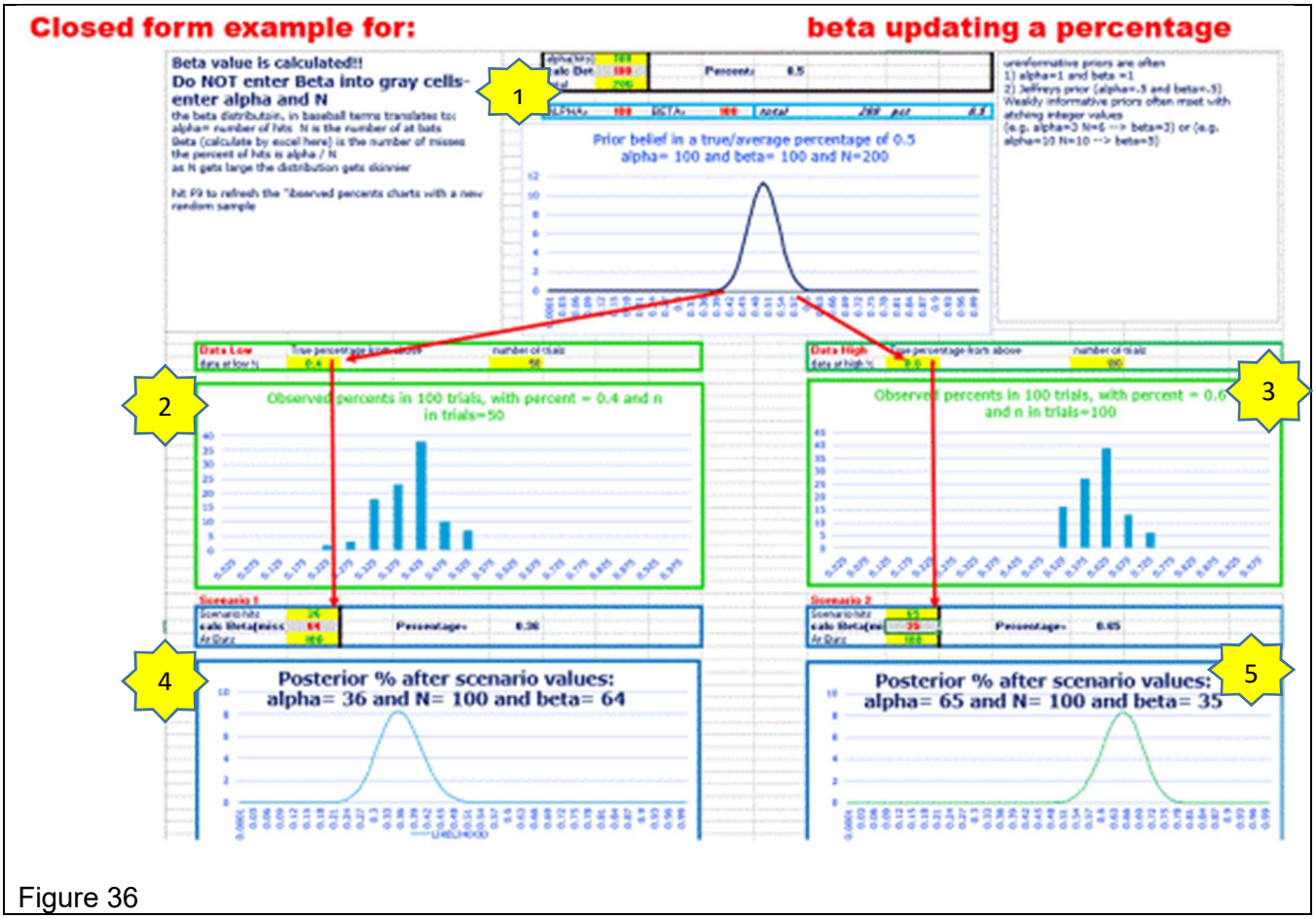


Figure 36

In step 1 we create a picture of a prior distribution of the percentage. We know that the client has some percentage in mind and the easiest way to center this percentage distribution on that number is to enter the percentage number (as integer) as alpha and enter 100 as N. The spreadsheet will automatically calculate beta as N-α.

Remember that, if were thinking in terms of baseball, α is the number of hits. N is the number of at-bats. Therefore; β will be the number of misses and is easily calculated as N-α. That centers the distribution at the proper point.

To make the distribution wider or narrower, mimicking the client's degree of belief in that percentage, one just has to increase/decrease N while keeping the ratio of α/N a constant. Increasing N will make the distribution narrower and decreasing N will make the distribution wider.

At this point we have modeled the clients' belief about a percentage. It is thought that this process would be new to many clients who might be more accustomed to seeing the data rather than some abstracted "distribution about the true percentage that generates the data".

Steps 2 and 3 are very much linked to the prior distribution and are intended to show the client the implications of his idea of the true percentage. It is thought that a client will likely have little experience with mentally imagining data from a distribution of possible percentages and can benefit from adding some concreteness to the problem. It is thought that a manager must often plots his percentage of "conversions per day" as a KPI and might recognize the distributions in step 2 or step 3

In step 2 we take the lowest percentage that the client thinks is possible, the place where the prior distribution touches 0 on the left-hand side, and use that to show the client what the data might look like at that end of his belief system. Step 2 requires that you input the percentage, where the curve in step one touches 0 on the left, into a cell in the Excel SS and Excel will generate random data for that percentage. The thought is that the analyst can show this to the client and say something like "so on your worst month of sales you expect the daily sales percentages to look like this".

In step 3 we take the highest percentage that the client thinks possible and reproduce the steps in step 2. The thought is that the analyst can show this to the client and say something like "so on your best month of sales you expect the daily sales percentages to look like this".

In steps 4 and 5 we allow the client to "see a bit into the future". You can input the clients expected percentage from step one and then say if our "to be collected" data percentage looks like this (might be 20% might be 50%) this is how it will update your belief about the true percentage (of customers who convert on your website). There are two charts here so that the client has the ability to play a little bit of "what if" to see how sensitive the update is to things that he could imagine might happen.

A CONCEPTUAL REVIEW OF THE MCMC ALGORITHM

As has been said before, the calculations required to do a Bayesian analysis are so complicated that they held back progress in the field until computers became powerful enough, and algorithms became good enough, so that the problems could be solved and produce practical results.

The only closed form solutions for Bayesian problems come from the Exponential family and require that someone derive a conjugate prior for a certain distribution. This is extremely difficult to do and has not been done many times.

The solution for this difficulty is to stimulate the Bayesian process and this can be done with many different software packages and many different algorithms. The MCMC algorithm is a classic and worth studying to get the concepts behind using computers and simulation to solve a Bayesian problem.

Figure 37 Is a link between what we've done and what we will do.

The denominator in the examples we did with discrete tables had numbers on the margins of the table. The marginal numbers represented our prior belief in something without having looked at any additional data.

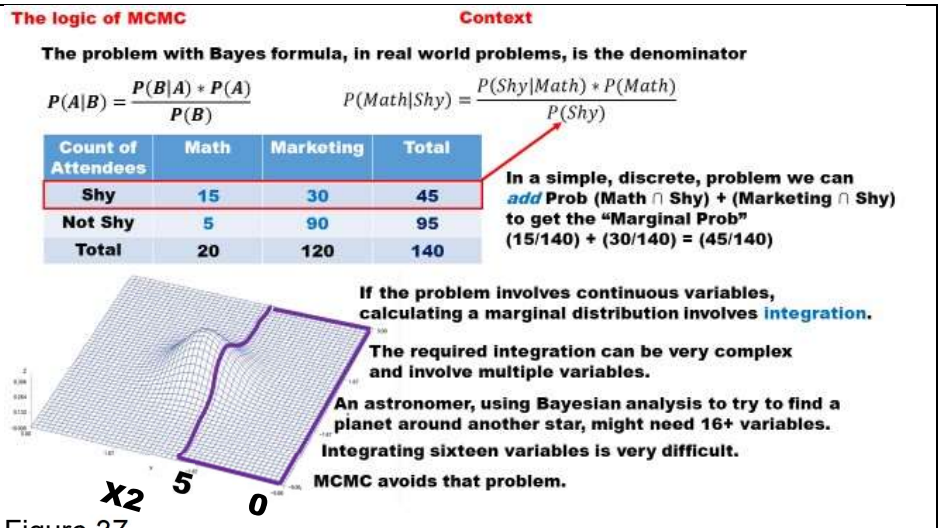


Figure 37

When working with a two dimensional table calculating the marginals is easy – even when they're not on the margin as they were with the smoking example. When one moves to a continuous distribution the summing up of cells has to be replaced by an integration process as is shown in the figure above. In the 2x2 table examples we summed the cells to get a denominator (the total number of people who were shy.) and threw away the rest of the table. Figure 37 shows an analogous process for a continuous variable. While I made up the need to be interested in between 0 and 5, the picture shows us throwing away "information and concentrating on only part of the data". We see a picture of a continuous problem where we might be saying "given X2 is <= 5".

Astronomers do use Bayesian analysis to try and find planets rotating around distant stars in the denominator in that calculation can have 16 to 20 different variables. That integration is difficult if not impossible. Fortunately there is a solution.

Figure 38 links us into something we've seen before. We've all use the normal distribution in one of our stat classes the formula is easy and is something divided by a constant.

That constant is called a normalizing constant and makes the area under the curve integrate to one and use area under the curve's as probabilities of an event happening because probabilities must sum to one.

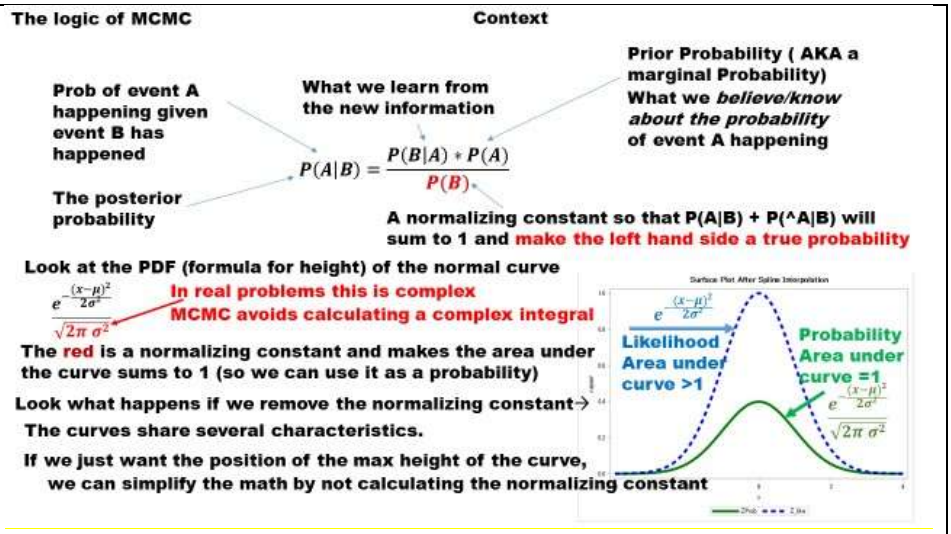


Figure 38

However; if we are just interested in relative probabilities, we can often get answers from a curve whose area does not integrate to one. If were only interested in relative probabilities, and not absolute probabilities, we can dispense with the calculation of the normalizing constant and this can make the calculations (and our life) a bit easier. Many real-world Bayesian calculations dispense with the calculation of the normalizing constant.

Figure 39 lists the steps in the MCMC algorithm for the Metropolis Hastings algorithm.

The goal is to find the highest point in some "accuracy/likelihood" space. The space is defined by the parameters that describe your distribution.
If one were trying to do a Bayesian simple regression the space defined by the two parameters for that regression are βo and β1 as is seen to the right.

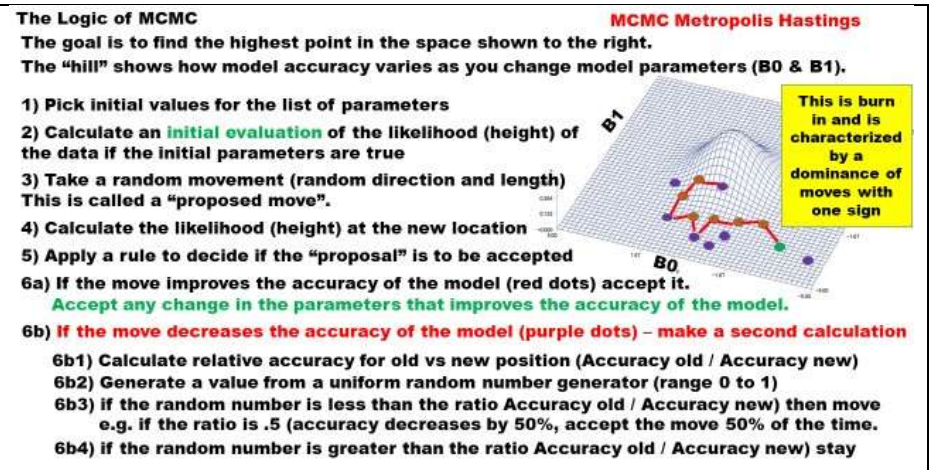


Figure 39

If we can create a distribution that is "highest" at the point where βo and β1 have maximum accuracy we can climb that mountain to find a solution to our problem using the algorithm below.

The Metropolis Hastings algorithm has the following steps:

1) Pick random initial values for your list of parameters (obviously, a good starting point helps a lot).
2) Calculate some initial evaluation of the likelihood of the data if these parameters are true.
3) Propose taking a random jump in βo and β1 (use a bivariate normal to generate length and direction)
4) Calculate the likelihood (likelihood is something like accuracy) of the data at this new point
5) If the model is better at the new point, then make that move.
6) If the model is not better at the new point perform additional calculations

6a) Calculate the ratio of $likelihood_{new}$/ by $likelihood_{old}$.
6b) We are willing to go in the "wrong direction" so that we can explore the space.
6c) Exploring the space keeps us from getting trapped in local minimum.
6d) Generate a random number.
6e) If the random number is smaller than the ratio, accept the jump and move; else stay where you are.

If the jump would put us in a position that's a little bit worse than where we are we should be willing to make that jump with great frequency – just so that we can explore the parameter space. If the move would put us in a position that's much worse than where we are we should be reluctant to make that jump but should occasionally make those jumps to explore the parameter space.

Rejecting jumps keeps us in one place and keeps us from exploring the space quickly. Accepting jumps can move us far in the correct, or in the incorrect, direction.

The picture in the figure above shows a pattern of jumps through parameter space.
It should be noted that, at the beginning, the algorithm will largely be climbing up the hill and the direction will largely be upward.

SAS automatically does some "tuning" of the process of accepting/rejecting "a jump that puts us in a less advantageous place" and does this tuning as a way to reduce run time. Simulations have shown that the algorithm is relatively insensitive to the percentage of jumps that are accepted and rejected but SAs still adjusts. SAS is not just doing the MCMC- Metropolis Hastings algorithm for us, it is removing some of the onerous and time-consuming sub-tasks that used to be performed by hand.

As the algorithm is climbing up the hill the preponderance of the jumps will be in one direction - upwards.

If we plotted the size of the jumps as the algorithm is searching for the hill we would see it moving in one direction – or tending to move in one direction.

The normal curve will tend to make the jumps small. This means that many jumps will be accepted but the hill climb will be slow.

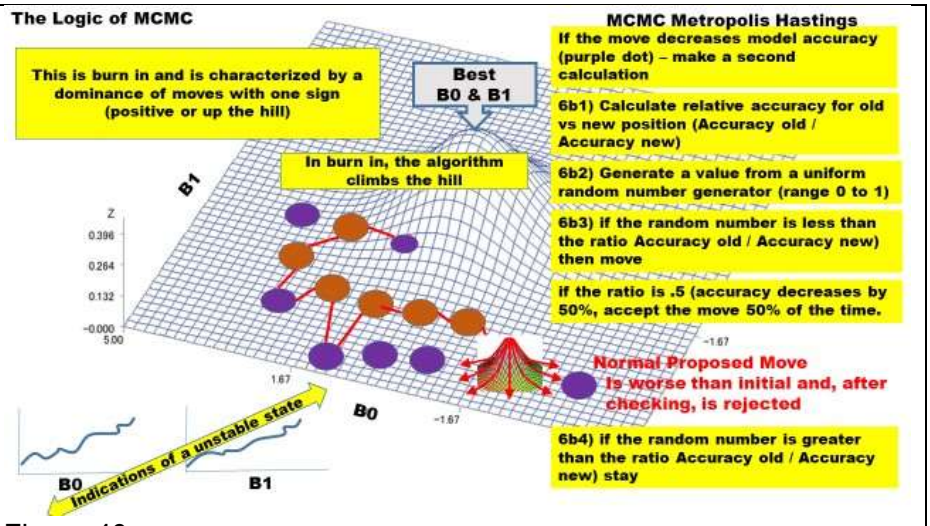


Figure 40

Large jumps explore the space quickly but can actually jump over a high point.

Figure 41 Shows what happens when the algorithm reaches the high point. SAS will tend to jump over and around the maximum point.

Metropolis Hastings, unlike some of the gradient descent algorithms, does not converge to an ideal point.

It converges to an area in parameter space.

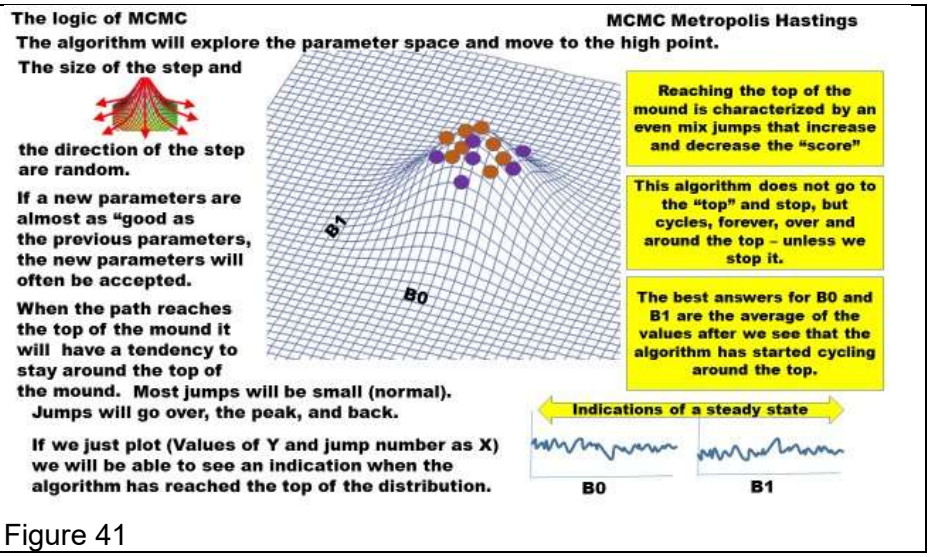


Figure 41

When it has reached a high point a plot of the jump direction and size will look like what can be seen in the bottom right-hand corner of this figure. Plots of the jump direction and size are the best indication of convergence.

Both β0 and β1 must converge, and both must show trace plots similar to what you can see in the picture above. If the parameter space is more than two dimensions, all trace plots must indicate convergence.

The Metropolis Hastings is only one of many algorithms that can be used in Bayesian Analysis. To take advantage of the increased computing power of multi-core chips, another algorithm is the Metropolis Coupled Markov Chain Monte Carlo or MCMCMC. It achieves the goal of exploring parameter space by having multiple "searcher chains" searching in parallel. At the end of every jump, an evaluation is made as to which searcher is in the "best position" and the searcher chain with the best score is allowed to write to the output file.

Selecting the search algorithm is quite a complicated subject and beyond the scope of this seminar. SAS, because it is trying to remove onerous tasks from a user, has very good defaults and will shift from one algorithm/set of settings to another without any input from the user.

Effective Sample Size

Effective sample size is often discussed as a way of measuring how well the Markov chain is "mixing". $ESS=n1 + 2\sum(n-1)k = 1\rho k(\theta)$ where N is the total sample size (number of "jumps") after burning is finished and $\rho k(\theta)$ is the autocorrelation of leg K for θ.

The closer ESS is to N the better the mixing, though people think that in ESS of around 1,000 is sufficient for estimating a posterior density. If you're looking to estimate the tails of the posterior density plots with some precision, you might want to have more than that 1000. ESS is not a significance test but it's more of a criteria that you can use to figure out whether it's reasonable to use the estimate on your printout. .

A CONCEPTUAL REVIEW OF GIBBS SAMPLING

While it's never too wrong to just take the defaults that SAS has set, it is worthwhile learning about another algorithm – because SAS uses it and because many people discuss it. That algorithm is Gibbs sampling and it is implemented in SAS.

Figure 42 starts an explanation of Gibbs sampling.

MCMC is very efficient if the distribution you're selecting from is the posterior distribution – but it's hard to tell what that posterior distribution is until the algorithm converges. Gibbs, like the Metropolis Hastings is trying to jump through parameter space to find a high point of a distribution.

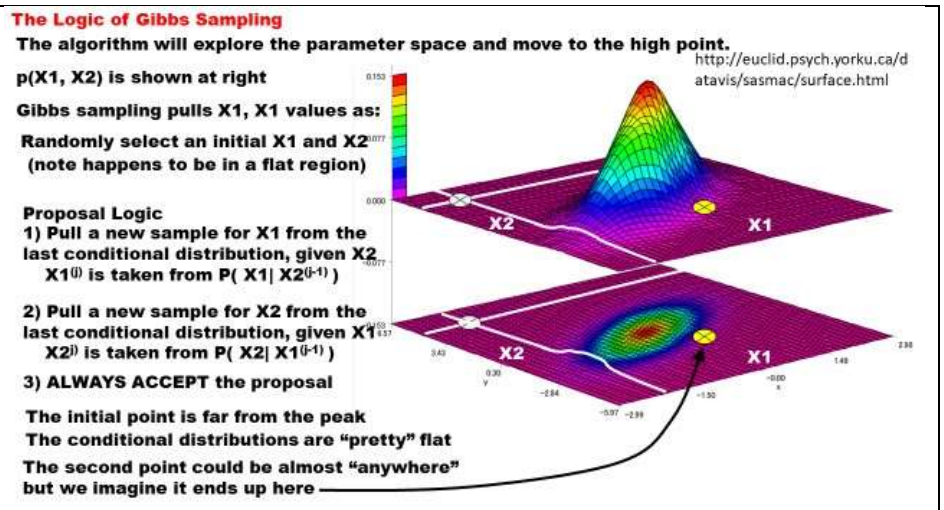


Figure 42

In this example parameters are called X1 and X2. As a 1st step, Gibbs will randomly select values of X1 and X2 (see white circle) and calculate the likelihood. Gibbs will now want to jump to a better position. Metropolis Hastings generated proposals using a normal distribution. Gibbs wants to sample from the posterior distribution but does not know what the posterior is.

As an approximation, Gibbs pulls a new sample for X1 from the last distribution, given that X2 is the value that it had in the last distribution (see white lines). It then repeats the process for X2. Gibbs does not know the posterior distribution but samples from "what it already knows". The idea is that whatever it already knows about the shape of the posterior is likely better than a normal distribution. These distribution are pretty flat and the first jump could be to anywhere. The first jump is to the yellow circle.

Figure 43 shows the next step for the Gibbs sampler. Gibbs pulls a new sample for X1 from the last conditional distribution holding X2 constant (see gold lines). It then repeats the process for X2.

The proposed conditional distributions are shown by the gold lines.

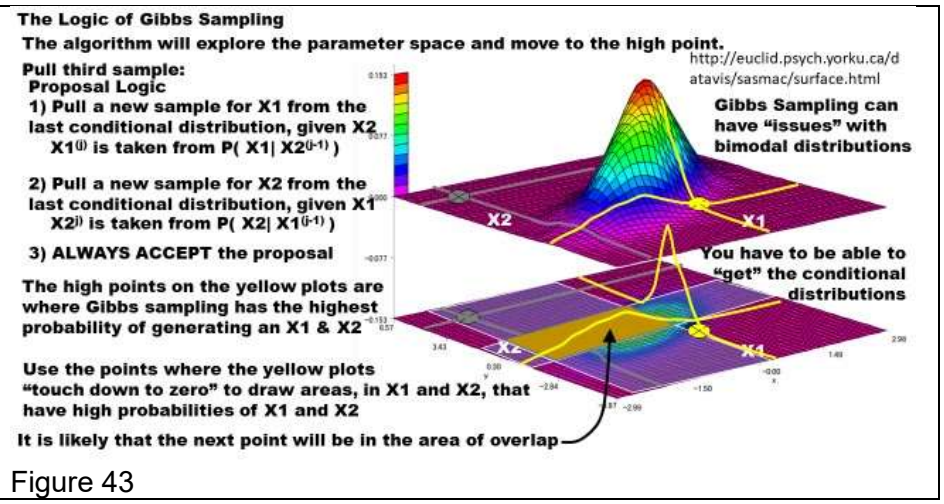


Figure 43

One of the limitations, in using the Gibbs, is that the algorithm must be able to calculate the conditional distributions, and this is not always possible.

Please look at Figure 43. If the next sample is going to be pulled from the place where the distribution for X1 and X2 are high it will likely be in the gold area and that is a significant improvement. Gibbs always accepts its proposed jumps. The process repeats

SAS EXAMPLE 1: MULTIPLE REGRESSION: GLM GENMOD AND MCMC

With a sufficient sample size, when you update the priors with “the data”, the data can overwhelm the influence of your prior belief, and Bayesian and frequentist analysis get to the same answer. In this section were going to do a compare and contrast between frequentist and Bayesian analysis in SAS and we expect similar answers.

PROC MCMC is the most powerful, customizable and complicated Bayesian procedure in SAS but there are other procedures that also give good Bayesian answers and are easier to use. PROC GENMOD is much easier to use than PROC MCMC. Let’s compare outputs from GLM, GENMOD and MCMC.

Figure 44 shows code and output for PROC GLM and PROC GENMOD. The actual code is below the figure.

All that is needed to get Bayesian out of PROC GENMOD.is a bayes statement. .

The answers produced by the two procedures are very similar.

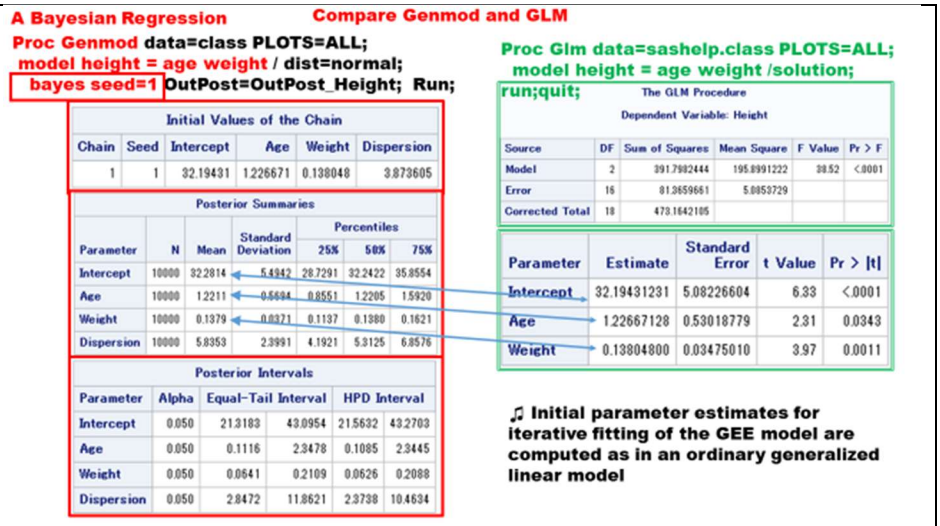


Figure 44

In place of confidence intervals, the Bayesian analysis will produce posterior credible intervals, that are kind of like confidence intervals. The code for the figure above is immediately below.

Proc Glm data=sashelp.class PLOTS=ALL; model height = age weight /solution; Figure 45	**Proc Genmod data=class PLOTS=ALL;** **model height = age weight / dist=normal;** **bayes seed=1 OutPost=OutPost_Height;** **Run;**

PROC GENMOD is smart enough to know when a conjugate exists for a particular model and will, because SAS wants you to get quick run times, use the conjugate formula rather than a simulation if it is possible.

The GENMOD Bayesian analysis also produces some diagnostics that are shown at right.

Most importantly are the plots on the right-hand side. They're describing the changes in the betas, caused by jumps, after "burn in" is finished and the jumps have reached the "top of the hill".

.

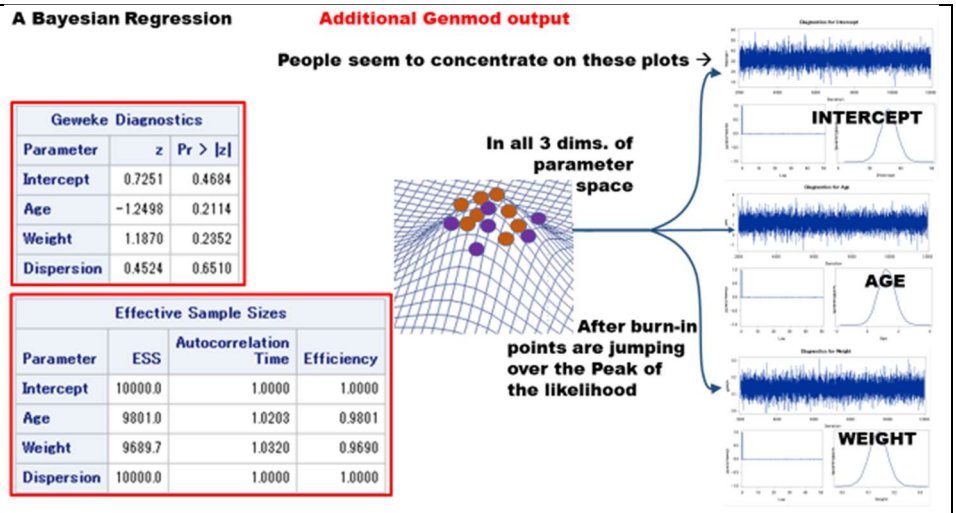


Figure 46

The plots you see at above show a mixture of small, and large, positive, and negative, jumps and are characteristic of the algorithm jumping, over and around, the top of the hill

The PROC MCMC code, and output, are shown below.
Data specifies the input data set.
Outpost specifies where the output should go.
NBI tells the number of jumps to be used as burn in. Burn-in jumps are not shown in the trace plots. We hope to reach the top of the hill before this number of jumps.
NMC is the number of jumps after burn in. The show up in the trace plots that can be seen in the figure above.
NThreads will allow you to take advantage of a multiple CPU computer.
Thin n only shows one n^{th} of the observations on the trace plot. This reduces correlation but I'm not sure if it helps much.
Seed starts a random number generator so that results can be reproduced.

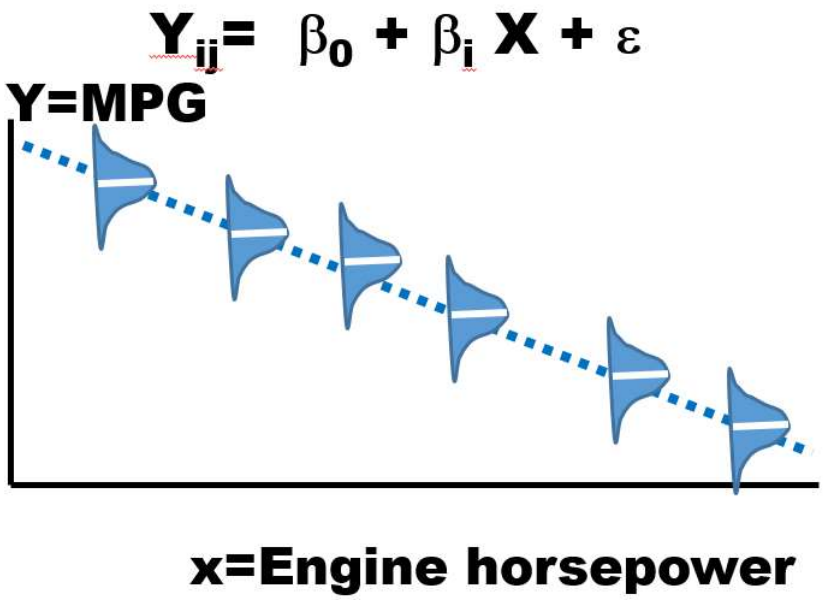
```
Proc MCMC data=SASHelp.class outpost=A03_MCMC_MltReg
  NBI=5000  /*# of burn-in jumps*/
  NMC=10000 /*# of MCMC jumps, excluding burn-in*/
  NThreads =3
  thin=5
  seed=246810;
  parms beta_Int 0 beta_Age 0 beta_Weight 0;
  parms sigma2 1;
  prior
     beta_Int beta_Age beta_Weight ~ normal(mean=0, var = 1e6);
  prior sigma2 ~ igamma(shape = 3/10, scale = 10/3);

  mu = beta_Int + beta_Age*Age + Beta_Weight * weight;

  model Height ~ n(mu, var = sigma2);     run;   ods graphics off;
```

Figure 47

Parm specifies the parameters and sets their initial values. You can have multiple parm statements.

The 1st parm statement specifies initial values for the X variables in our problem. We don't have much of an idea of what the beta values should end up being so were going to start them off at 0. If we have enough observations in our data set, the initial values will have very little impact on the final result.

The 2nd parm statement says we think the value of the standard deviation should start out at 1 (see Figure 47).

The prior statements provide more information about the prior distributions of the parameters. The betas for the X variables are expected to have a mean of 0 and the variance that's very large, making this a pretty uninformative prior. The beta for the error around the fitted regression line will be modeled using an inverse gamma with a shape and scale that makes this an uninformative prior. The mu statement and the model statement go together.	 Figure 48

Remember that regression predicts the conditional mean Y value for a set of Xs. A regression line predicts the conditional means of Y, given that the independent variables have some value. Regression also assumes that the errors around those means are normal.

The mu statement specifies the model relationship. The model statement says that height is a function of the relationship specified in the mu statement. It says that, if the hypothesized relationship, defined by the mu formula, is true, then the data should be normally distributed for a particular value of Y. Please go back a look at the dice example for a parallel.

The MCMC output, seen to right, matches the output from the previous two procedures.

The three PROCs gave the same answer.

Note: This model might benefit from a longer burn-in. The ESS numbers could be improved.

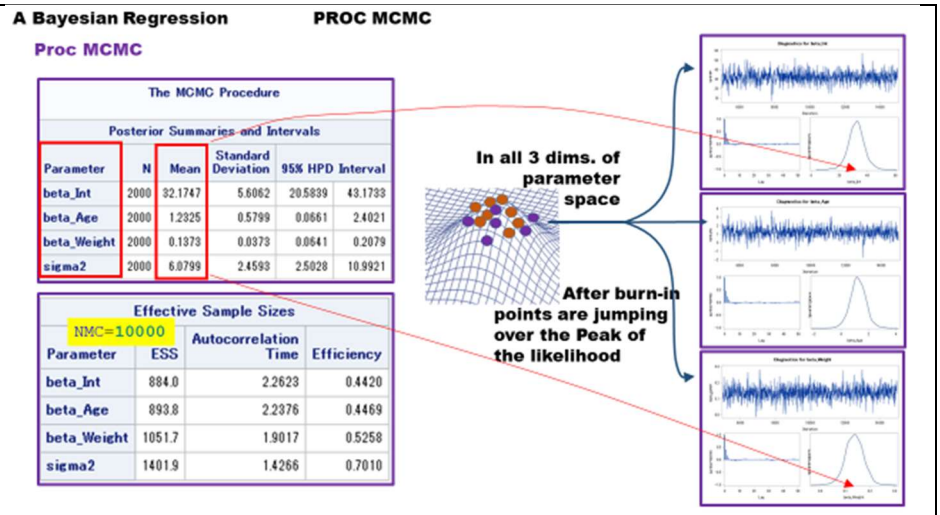


Figure 49

SAS EXAMPLE 2: LOGISTIC REGRESSION: GLM GENMOD AND MCMC

Figure 49, to right, shows the code and output from a logistic regression using PROC Logistic.

The code is simple and easy to understand.

PROC Logistic *data=vaso descending plots=ALL;*
model resp(Event="1") = lvol lrate / link=logit;

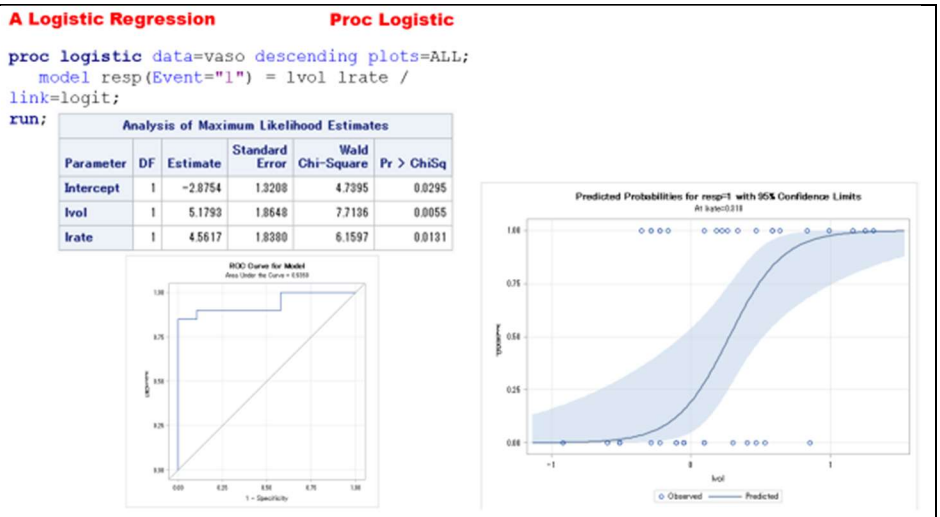


Figure 50

I think SAS ODS output is not only helpful but very professional and ready for inclusion in any sort of report.

The PROC GENMOD code is also easy to understand and options are similar to those in PROC MCMC.

PROC GENMOD
data=vaso
descending;
ods select
PostSummaries
PostIntervals;
model resp = lvol lrate
/ d=bin link=logit;
bayes seed=**17**
coeffprior=jeffreys
nmc=**20000** thin=**2**;
run;

A Bayesian Logistic **Proc Genmod**

```
proc genmod data=vaso descending;
   ods select PostSummaries PostIntervals;
   model resp = lvol lrate / d=bin link=logit;
   bayes seed=17 coeffprior=jeffreys
        nmc=20000 thin=2;
run;
```

```
NOTE: The PLOTS= option is ignored for a Bayesian analysis.
NOTE: The default sampling algorithm is the Gamerman algorithm,
which is different from the default in SAS/STAT 9.3 and
    earlier releases. To revert to the previous behavior, specify the
SAMPLING=ARMS option in the BAYES statement.
NOTE: PROC GENMOD is modeling the probability that resp='1'.
NOTE: Algorithm converged.
NOTE: The scale parameter was held fixed.
```

The SAS System

The GENMOD Procedure

Bayesian Analysis

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
Intercept	10000	-2.8773	1.3213	-3.8821	-2.7326	-1.9097
lvol	10000	5.2059	1.8707	3.8535	4.9574	6.3337
lrate	10000	4.5525	1.8140	3.2281	4.3722	5.6643

Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
Intercept	0.050	-5.7447	-0.6877	-5.4593	-0.5488
lvol	0.050	2.2066	9.4415	2.0729	9.2043
lrate	0.050	1.5906	8.5272	1.3351	8.1152

Figure 51

The seminar ends with the next, and last, example for PROC MCMC. **PROC MCMC is its own programming language, like the data step**. A detailed explanation of this language is beyond the scope of this seminar. The manual for PROC MCMC is a bit over 300 pages long and makes assumptions that the reader has a pretty good understanding of Bayesian statistics and of SAS programming.

With that caveat, it would probably be helpful to take a high-level walk-through of the code below.

```
%let n = 39;
PROC MCMC data=vaso nmc=10000 outpost=MCMCLogistic seed=17;
  array beta[3] beta0 beta1 beta2;    array m[&n, &n];          array x[1] / nosymbols;
  array xt[3, &n];                    array xtm[3, &n];         array xmx[3, 3];
  array p[&n];

  parms beta0 1 beta1 1 beta2 1;

  begincnst;
    if (ind eq 1) then do;
      rc = read_array("vaso", x, "cnst", "lvol", "lrate");
      call transpose(x, xt);
      call zeromatrix(m);
    end;
  endcnst;

  beginnodata;
  call mult(x, beta, p);            /* p = x * beta */
  do i = 1 to &n;
    p[i] = 1 / (1 + exp(-p[i]));    /* p[i] = 1/(1+exp(-x*beta)) */
    m[i,i] = p[i] * (1-p[i]);
  end;
  call mult (xt, m, xtm);           /* xtm = xt * m       */
  call mult (xtm, x, xmx);          /* xmx = xtm * x      */
  call det (xmx, lp);               /* lp = det(xmx)      */
  lp = 0.5 * log(lp);               /* lp = -0.5 * log(lp) */
  prior beta: ~ general(lp);
  endnodata;

  model resp ~ bern(p[ind]);  run;
```

Figure 52

In the first section of code we see array statements. Array statements will associate a name to a list of variables or constants. Note that, in MCMC, the array statement is similar to, but not identical with, the array statement in a typical SAS data step. The array statement in MCMC is the same as the array statements in NLIN, NLP, NLMIXED and model procedures. Note that the array statement in PROC MCMC is more limited than the array statement in the data step. The statement SCSUG(5) creates variables with the names SCSUG1 to SCSUG5.

The parm statement lists the names of the parameters to be used in the model and can also specify initial values for these parameters. You can specify multiple parm statements and this is useful when you want to apply different options to different parameters. Each parm statement defines a “block of parameters” and the Metropolis Hastings algorithm will attempt to jump all the parameters in a block simultaneously.

MCMC is faster if we put correlated parameters in the same “parameter block”. However; if many parameters are updated at one time, it is likely that one of the parameters in the block will end up being a large jump in one (and bad) direction – thereby lowering the ability of the “group of the parameters” to predict and causing PROC MCMC to reject all of the moves for parameters in that block. There’s a trade-off involved in blocking parameters.

Putting five parameters in five different blocks increases the run time by a factor of five, though it does increase the chance of jumps being accepted.

An important issue, in using PROC MCMC, is having a good mixture of the sequence chains. The jump points should quickly explore the parameter space and the blocking of parameters can be very important in this. The common practice is to block, on the same parm statement, small groups of correlated parameters.

The BEGINCNST-ENDCNST block causes PROC MCMCM to skip unnecessary evaluation in order to reduce the run time. Statements inside this block execute only during the set up stage of the simulation and are very useful to define constants or to load variables into arrays. If you have program statements, remember PROC MCMC is itself a programming language, statements that do not need to be evaluated for each loop through the simulation should be put in the BEGINCNST-ENDCNST. PROC MCMC will evaluate the programming statements inside this block once for each observation in the data set and ignore the statements during the rest of the run.

PROC MCMC can use multiple data sets as input and sometimes users need to store variables in arrays and use those arrays to specify the model. The read_array function, seen in this block of code is an easy way to read a data set into an array. PROC MCMC provide users with over 10 call routines that allow the user to perform matrix operations on arrays inside PROC MCMC.

PROC MCMC does not have a class statement and so design matrices must be created using PROC TransReg and imported into PROC MCMC.

The BEGINNODATA-ENDNODATA block defines statements that get executed without stepping through the entire data set. These statements are executed only two times: at the first and at the last observations in the entire data set. Any calculations that would be identical for every observation in the data set should be enclosed in this block. This block is run for the 1st observation to ensure that values have been calculated correctly. It’s run for the last observation because PROC MCMC executes the statements and then adds results to the output dataset.

Each parameter must have a prior. Inside the BEGINNODATA-ENDNODATA we see that this program uses the colon operator to initialize all of the beta values. This syntax is: PRIOR; one or more parameters; a tilde (~) and then the distribution to be applied with all of the distributions parameters. The name of the distribution to be applied is “general”.

The model statement specifies the conditional distribution of the data given the parameters. This is also the likelihood function given the hypothesized model. The model statement assumes observations are independent of each other and conditional on model parameters.

If you do not have data that’s independent of each other, SAS provides other procedures.

The model statement must come after any SAS programming statements that define, or modify, arguments used in the construction of the log likelihood in the model. This model statement uses bern, for Bernoulli indicating a binary distribution with having P as the probability of success. Ind evaluates to the probability of success.

The output from the PROC MCMC is below.

The PROC MCMC output starts with the information in Figure 53.

All of the observation that were read word used and it tells that the three parameters in the model were in one block and had an initial value of one.

The code that requests this can be seen in Figure 52.

Logistic Regression Model with Jeffreys Prior

The MCMC Procedure

Number of Observations Read	39
Number of Observations Used	39

Parameters

Block	Parameter	Sampling Method	Initial Value	Prior Distribution
1	beta0	N-Metropolis	1.0000	general(lp)
	beta1		1.0000	general(lp)
	beta2		1.0000	general(lp)

Figure 53

Figure 54 shows the estimated coefficients. The column titled mean, to the right, are the beta values. The estimated value for the intercept is -2.85.

The 95% HPD interval is a Bayesian version of a confidence interval. HPD stands for highest posterior density. If you think of a normal curve, there are many different cutpoints one could select such that 95% of the area under the normal curve is between the two cutpoints.

That same phenomenon happens in the Bayesian world. The 95% HPD interval is the narrowest interval that can be created that contains 95% of the observations.
Many statisticians suggest using the 95% HPD interval.

The ESS values are greater than 1000 and that gives us some comfort.

Logistic Regression Model with Jeffreys Prior

The MCMC Procedure

Posterior Summaries and Intervals

Parameter	N	Mean	Standard Deviation	95% HPD Interval	
beta0	100000	-2.8513	1.3031	-5.4622	-0.5151
beta1	100000	5.1554	1.8517	1.7829	8.8176
beta2	100000	4.5236	1.8129	1.1834	8.0820

Logistic Regression Model with Jeffreys Prior

The MCMC Procedure

Effective Sample Sizes

Parameter	ESS	Autocorrelation Time	Efficiency
beta0	3314.5	30.1708	0.0331
beta1	3409.0	29.3341	0.0341
beta2	3254.1	30.7308	0.0325

Figure 54

Here is the report for beta 0.

The curve in the lower right hand corner shows the belief in the value for beta 0. The high point of the curve is not above the -2.85 that we saw in Figure 54. You will notice that this distribution has a long tail to the left and that will pull the average to the left of the mode.

The mixing shown at the top indicates that we have converged on the top of our distribution.

It would be hoped that the correlations were less severe but they do not impact the accuracy of the estimate and can be tolerated.

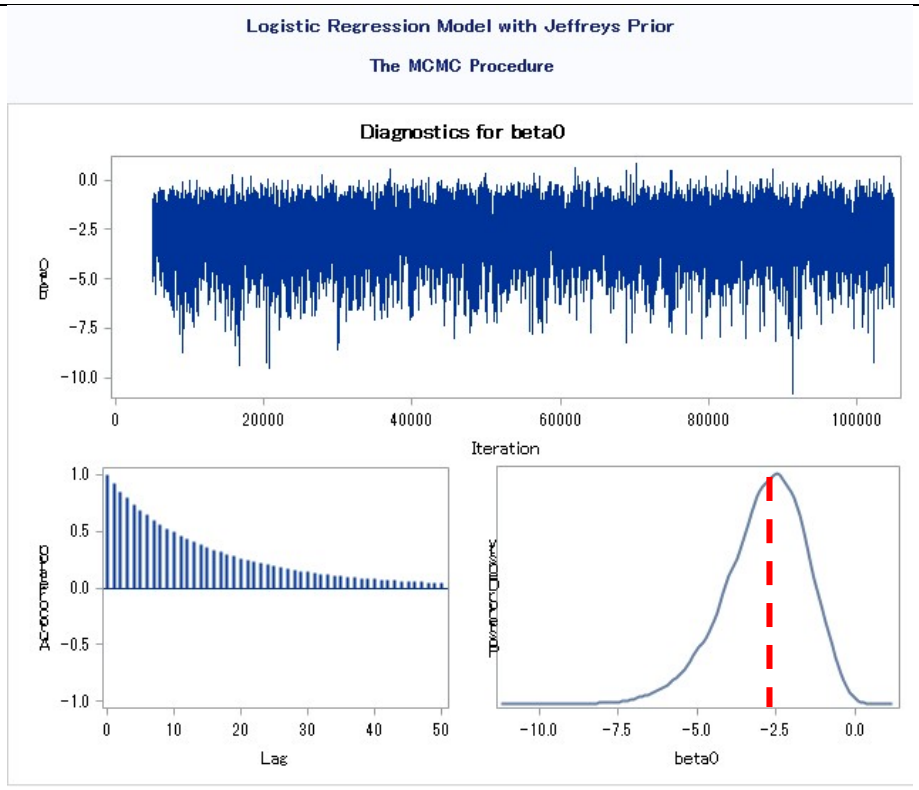


Figure 55

To the right is the report for beta 1.

The interpretation of this is very similar to the interpretation in Figure 55.

The numbers that we can see in Figures 55, 56, and 57 are very close to the numbers that were reported using other techniques on this data.

Those numbers are reported in figures above.

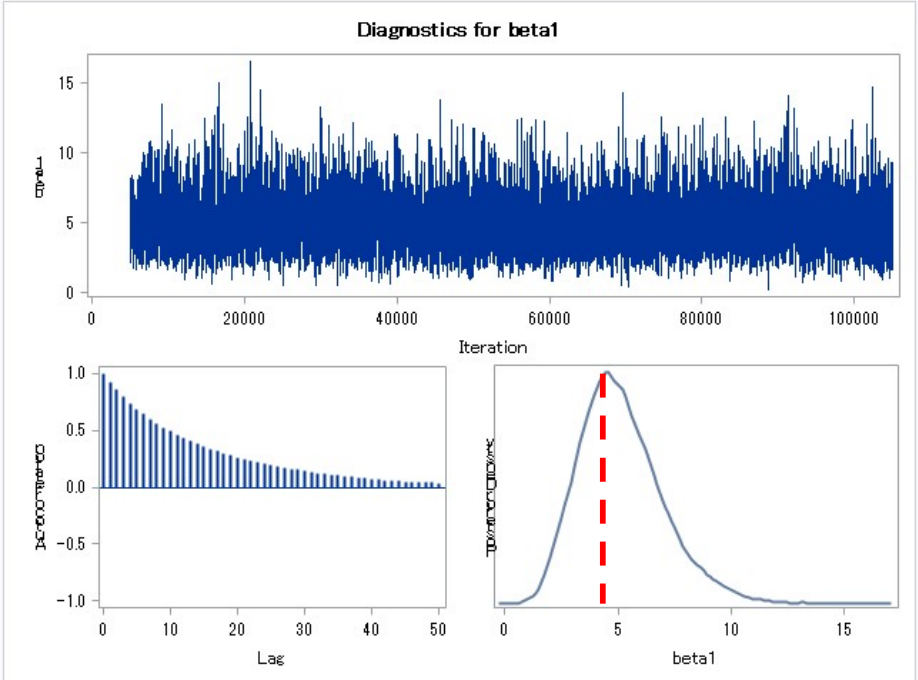


Figure 56

Figure 57 shows the output from PROC MCMC for beta 2.

The interpretation of this figure would be the same as for the previous two figures.

In summary, the three procedures that we used on the same data set all gave very similar answers.

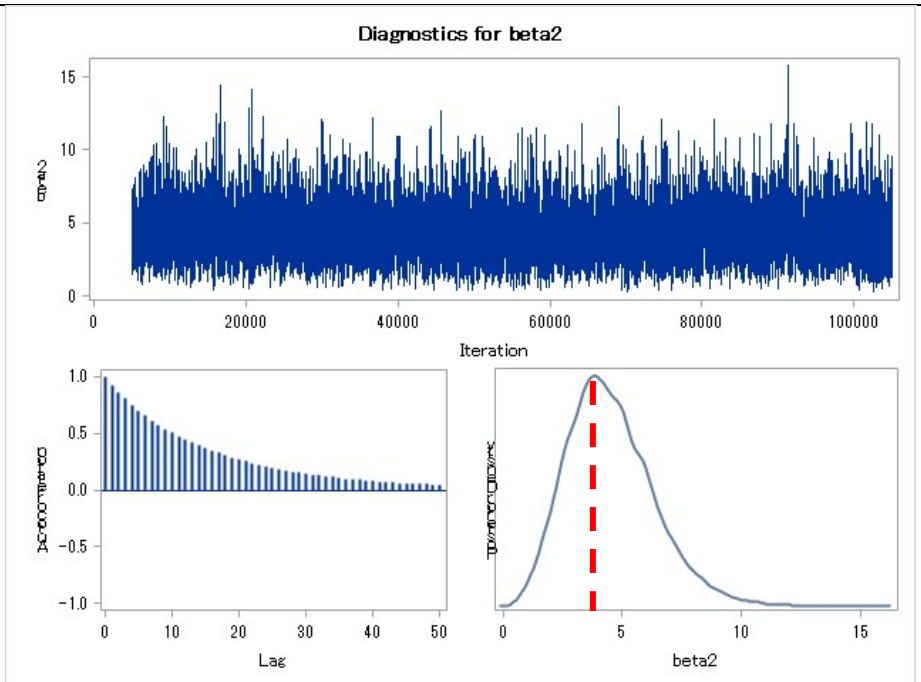


Figure 57

While PROC MCMC is very versatile and very powerful, Bayesian analysis, in a great many situations, can be done with procedures that have a syntax more familiar to people with experience with PROC GLM and would require less study.

CONCLUSION

Thank you for attending this seminar and I hope this has been helpful to you.

Russ Lavery