

Filtering the Bible and Filtering Spam

Dr. Gene B. Chase, Professor of Mathematics and Computer Science,
Messiah College, Grantham, PA 17027

Abstract: I argue that John Craig (1663?–1731) is the first to do Bayesian statistics. Filtering email spam today using Bayes’s analysis of 1763 is a new application of an old theorem. Craig 67 years before Bayes’s theorem used subjective probabilistic reasoning to argue that Jesus would return in the year 3150, because the Bible would eventually come into disrepute (become spam?) then.

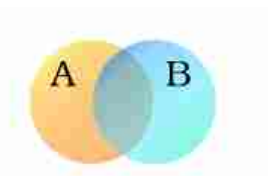
1. Introduction.

The theorem of Thomas Bayes (1702?–1761) for which he is best know was published in his “Essay Towards Solving a Problem in the Doctrine of Chances,” published posthumously in the *Philosophical Transactions of the Royal Society of London* in 1763. Bayes’s theorem is sometimes left out of introductory statistics courses because it calculates probabilities as outputs from estimated probabilities as inputs, and so it seems circular. If we can estimate probabilities, why not just do so and be done with it? However, it is a natural choice for calculating probabilities in artificial intelligence and other decision support applications, where machine learning algorithms can update prior probabilities. In this era of email spam, Bayes’s theorem has become widely known for its use in spam filters. Sixty-seven years prior to Bayes’s theorem, John Craig (1663?–1731)—also spelled “Craige”—in his 1696 work *Mathematical Principles of Christian Theology* (published in 1699) calculated posterior probability based on prior probability in a way that we would today call an application of Bayesian reasoning. Craig used these calculations to argue, under certain assumptions, that the Bible would eventually come into disrepute in the year 3150 AD. As a Christian, Craig argued that Christ would not return until then.

2. Review of Bayes’s Theorem.

Consider two events A and B in a universe or sample space U. Define the conditional probability of A given B, $\Pr(A | B) \equiv \Pr(A \cap B) / \Pr(B)$, assuming of course that $\Pr(B) \neq 0$. Then symmetrically, $\Pr(B | A) = \Pr(B \cap A) / \Pr(A)$, similarly assuming that $\Pr(A) \neq 0$. It’s easy to show that $\Pr(\cdot | B)$, as a function of its first argument, satisfies the axioms for a probability measure. Intuitively, we merely change our mind about the universe, from the set U to the set B.

Bayes’s theorem then follows simply from the fact that $A \cap B = B \cap A$. Changing A and B to the suggestive letters H for hypothesis and E for evidence, we can show



$$\Pr(H|E) = \Pr(E|H) \Pr(H)/\Pr(E). \tag{1}$$

In English, the probability of a hypothesis H given evidence E can be calculated by multiplying the probability of the evidence given the hypothesis by the ratio of the probabilities of the hypothesis and the evidence. Since coming up with a value for $\Pr(E)$, the probability of the evidence, is not easy, we can in fact eliminate it from the equation by using the fact that

$\Pr(U) = 1$, and replacing $\Pr(E)$ with $\Pr(H) \Pr(E|H) + \Pr(H') \Pr(E|H')$ where H' is $U-H$, the complement of H in U . A little rearrangement of the given equations produces the following form of Bayes's theorem:

$$\Pr(H|E) = \frac{\Pr(E|H)\Pr(H)}{\Pr(H)\Pr(E|H) + \Pr(H')\Pr(E|H')} \quad (2)$$

This equation can be interpreted loosely as saying, If I behave rationally (i.e. according to the axioms of probability), then I should change my belief in H based on new evidence E in exactly the way described by (2). Let's apply Bayes's theorem to filtering spam from my email. In the following example, I will use in place of the denominator the equivalent $\Pr(E \cap H) + \Pr(E \cap H')$ in terms of sets. But the important point to notice about (2) is that it does **not** involve set theory. That is to say, (2) is a statement about probabilities and conditional probabilities, however they may be defined or calculated. In particular, the key fact to take away from this section is that **we estimate conditional probabilities based on prior experience**. Bayes's theorem is a statement about subjective probabilities.

3. Filtering Email.

Let E be the event "an email contains the word NIGERIAN." Let H be the hypothesis that an email is spam. Then H' is the event "an email is ham" (i.e. not spam). The value of $\Pr(H|E)$ will tell me whether to file the email in my spam folder automatically, as follows.

Let's assume that 400 out of 3000 spam that I have collected during a training phase contain the word NIGERIAN, but only 5 out of 300 of ham that I have collected do. I as a human make the decision as to which are spam. $\Pr(H \cap E) = 400/3000$ spam. $\Pr(H' \cap E) = 5/300$ ham. So using Bayes's theorem,

$$\Pr(H|E) = 400/3000 / (5/300 + 400/3000) = 8/9 = 0.8889. \quad (3)$$

So the probability is currently a rather high 88.89% that a new email coming my way with NIGERIAN is spam, given the evidence collected so far.

To use Bayes's theorem to make predictions, I must make two decisions, one about training and another about risk. First, I must monitor all the words of all the emails that I get for several weeks and subjectively decide by hand whether the mail containing them is spam or ham as a training step before I turn such an automated system loose. I assumed that, above. Second, I must subjectively decide what my comfort level is for declaring something to be spam. I am a very low risk taker so I would rather see a lot of my spam, rather than miss even one ham by misclassification. I prefer to see new mail unless there is a 99% chance that it is spam. I don't want to miss any ham. I want to move any new email containing the word NIGERIAN that I decide is spam into my spam folder manually, thereby changing E , until $\Pr(H|E)$ exceeds 99%.

Suppose however that I am happy with a threshold of 88%. Then my spam filter has been trained. I trust it to decide automatically whether or not to put new email in the spam folder. Suppose a new email comes along that also has NIGERIAN in it. $\Pr(H|E)$ has changed because the evidence E has changed. $\Pr(H|E)$ is now $401/3001 / (5/300 + 401/3001)$ or 88.91%. Thus the automatic classification of a new email as spam increases slightly the subjective, conditional

probability that further email containing the word NIGERIAN would be considered spam. Had the fraction of ham containing NIGERIAN been 1/300 instead of 5/300, I would be able to be even more certain that the first-mentioned email was spam, since $400/3000 / (1/300 + 400/3000) = 97.6\%$.

Of course in a real situation my computer is monitoring every word of my emails, not just NIGERIAN, and it is monitoring other structural features of my email like whether the email is all pictures, all capital letters, comes from a known spam source, or uses colored text. This computation is more involved, but it is not conceptually more difficult. One simply replaces scalar evidence by a vector of evidence.

The key fact to take away from this example is that **it is not so much the absolute size of the probabilities of spam and ham that is important, but their relative sizes.**

4. Filtering the Bible.

John Craig in 1696 used something that he called “probability” that we would not call probability, regardless of whether we are referring to an axiomatic, a relative frequency, or a subjective definition of probability. In this section, we will use the term “believability” instead, but will keep the variable's name P. We will define it more carefully later. Craig says that the believability, P, that an account of an event will be faithful to historical facts depends on the believability, z, that the account was faithful when first told; on the time, T, since the event; on the distance, D, from the event; and on the number of times, m, that the event has been retold. [Nash 1991, 23] Craig argues that z is much larger for oral retellings than for written accounts, so he actually offers a date of 800 AD for the disappearance of credible oral accounts of the history of Jesus, if that were all we had. Given that we have written accounts, Craig computes the year 3150 AD as the earliest date after which the written Gospels will no longer be believed.

In this section we will look at Craig's model and his calculations. In the next section we will discuss it from a Bayesian perspective. Craig's formula, simplified slightly, is

$$P = c z + f (m-1) + k T^2 + q D^2 \quad (4)$$

where f, k, and q are dimensional constants of proportionality that allow each term to be dimensionless, and c is the number of witnesses. [Nash 1991, 24] f, k, and q are negative (regarded as “suspicions” to use Craig's term), so Craig's believability P decreases with time and distance over the interval that he cares about. The quadratic terms arise because Craig assumes that it is the velocities of the suspicion terms for time and distance that are linear.¹ Craig then integrates the suspicions geometrically to get the decay terms for believability.

Here is a sample of Craig's arithmetic calculations. Craig, calculating in the year 1696, considered the following special case of the event of Jesus' life as told in the Gospels: Let x be an arbitrary base believability for a single faithful oral account, a kind of prior believability.

(i) Assume that written accounts are 10 times as faithful as oral accounts, $z = 10 x$.

(ii) There are four Gospel writers, so $c = 4$. (However, we know that Luke was not an eyewitness, and we know more generally that the Gospel accounts are not independent—facts

¹ In this respect, Craig is being consistent with the second half of his book, for which see Section 10 below.

overlooked by Craig in this analysis.)

(iii) Assume that a written text will last 200 years without being retold. So between 0 AD and 1696 AD there would need to be $m = 1696 / 200 = 34/4$ retellings. (However, we know that Jesus was an adult before He began his ministry, so a 0 AD start might need to be adjusted.)

(iv) Assume that $D = 0$. Craig has in mind D as a physical distance on the earth; we might better regard it as a linguistic distance, representing through how many languages has the written report been translated. For example, $D = 0$ for Greek, $D = 1$ for a translation directly into English; $D = 2$ for a translation generated from the English translation. (Compare [Stigler 1985].)

(v) Assume that the time scale for the constant k is $t = 50$ years. Then $T = 1696 \text{ years} / t = 1696 \text{ years} / 50 \text{ years} = 34$.

(vi) Assume that $f = k = -x/100$. That is to say, assume that the speed with which suspicion grows is only 1/100 of the believability x of a single faithful oral account on which we are basing this analysis.

Craig therefore concludes:

$$P = 4 \cdot 10 x - (34/4 - 1) x / 100 - 34^2 x / 100 = 11346 x / 400 \approx 28 x. \quad (5)$$

Restating Craig's argument in English, we conclude: Under the assumptions (i) through (vi), in order for the believability that the Gospel accounts in 1696 be as faithful to history as they would be if they were contemporary, one would need the same initial believability as one could obtain by hearing 28 oral accounts as a contemporary of Christ's. [Nash 1991, 69] The believability dropped from $40x$ to $28x$ as the memory of the original events faded.

How does Craig arrive at the date of 3150 for the time when the Gospels will no longer be believed? He regards this date as being the date of the return of Christ, based on Luke 18:8, "When the Son of Man comes, will He find faith on the earth?" He assumes that when $P = 0$, the world will be without faith.

Using a general T instead of 1696, and the same assumptions as above, we get

$$P = 40x - (T/(4t) - 1) x/100 - (T^2/t^2) x/100 \quad (6)$$

where $t = 50$ years, as in assumption (v). For what value of T is $P = 0$? We calculate 3156 years, which Craig calculated approximately to be 3150 AD, in error perhaps because of premature rounding of intermediate results.

Shall we agree with Augustus DeMorgan that Craig's work is just "very silly"? Or as Laplace claimed "quite bizarre" [Stigler 1986, 884; 1999, 253], or others have said, "an insane parody of Newton's *Principia*," "perfectly arbitrary," [Stigler 1999, 253], or a "travest[y]." [Hacking 1984, 72] Craig's conclusions seem so sensitively dependent on seemingly arbitrary hypotheses!

There is an irony in such remarks coming from Laplace, since Laplace follows Bayes on the issue of retaining priors (rather than eliminating them by assuming equal likelihood). About that I will say more below. [Jaynes 1974, 138] Stigler underscores this irony with an irony of his own. Stigler applies Craig's reasoning successfully to determining the date of Laplace's birth! [Stigler 1985; 1986; 1999]

Certainly, Craig's conclusion contradicted John Napier's (d. 1617). In Napier's 1593 book, *A Plaine Discovery of the Revelation of St John*, Napier showed from the book of Revelation

that the world would end in 1786 and that the Pope was the Antichrist, which made him “a respected theologian throughout Protestant Europe.” [“John Napier” 2005] Some Protestant contemporaries of Craig, spurred on by Napier's views, expected Jesus to return any day. No math, no controversy!

We so far see one thing in common between our Bayesian spam example and Craig’s example. Neither talk about absolute believabilities; **both quantify the relationship between a prior and a posterior believability.**

5. A Modern Reinterpretation of Craig.

Craig's P is certainly not a probability. But it is a statistic. That is to say, P is a quantity measured from experimental data. What kind of statistic is P in modern terms? In 1985 Stigler argued that it can be understood as a kind of log likelihood ratio. Let me summarize Stigler's argument. [Stigler 1999, 255]. Consider the expression Q,

$$Q = \ln \left(\frac{\Pr(E|H)}{\Pr(E|H')} \right). \tag{7}$$

Apply Bayes's theorem: If $\Pr(H)$ is the a priori probability of H, independent of any testimony E, then the posterior odds in favor of H on the evidence E are just

$$\frac{\Pr(H|E)}{\Pr(H'|E)} = \frac{\Pr(H)}{\Pr(H')} \cdot \frac{\Pr(E|H)}{\Pr(E|H')}. \tag{8}$$

[Stigler] Now take the logarithm of both sides of (8), and substitute Q from equation (7) to get

$$\ln \left(\frac{\Pr(H|E)}{\Pr(H'|E)} \right) = \ln \left(\frac{\Pr(H)}{\Pr(H')} \right) + Q \tag{9}$$

Let Q be approximated by Craig's expression $P = cz + (m-1)f + T^2k + D^2q$, where, to quote Stigler again, “m, T, and D are explanatory variables.” [Stigler 1999, 256] In this discussion, H is held constant. Define the constant α to be $\ln [\Pr(H) / (1-\Pr(H))] + cz - f$. Then we can rewrite equation (9) as

$$\ln \left(\frac{\Pr(H|E)}{\Pr(H'|E)} \right) = \alpha + mf + T^2k + D^2q \tag{10}$$

where the right hand side is Craig's P up to a constant summand, which is not important for Craig, since he is only concerned with differences. Stigler asks us to view α as embodying all prior information. Craig assumed that the number m of retellings is proportional to time T, more precisely, $m = T/4$ (because m is T/200 in years, but we are scaling T so that t=50 years represents 1 unit of time). Therefore, we can simplify (10) further by replacing f with a new constant of proportionality, $s = f / 4 = -x/400$ to give (11):

$$\ln\left(\frac{\Pr(H|E)}{\Pr(H'|E)}\right) = \alpha + Ts + T^2k + D^2q \quad (11)$$

Craig's P can be seen to be a quadratic approximation to “the change in the log odds due to the available testimony,” E over time. [Stigler 1999, 255] Equation (11) then says that Craig's P is not a probability, but is a quadratic approximation to a posterior log odds, parametrized by time and distance. (We are more accustomed to parametrizing statistics by mean and standard deviation, so you may have to think about this some more.)

The fact that Craig started with an arbitrary x and calculated everything in terms of it shows that he was not interested in the probability of something as we use the term “probability” today, but in the change in P, where P is a kind of decaying believability that additional evidence would bring about, or additional time, distance, and recopying would bring about. Craig used an additive model for this change in believability, which gives a problem when P drops below 0, but he was on the right track. In fact, even today there are debates about whether a ratio or a difference is the best statistic to use when calculating with conditional probabilities. [Joyce 2003, Tables 1, 2, 5, and 6] Craig's model is additive in log space, hence multiplicative in physical space, as I shall argue below.

Craig's quadratic approximation is nearly linear. Consider equation (6), in which the denominator of the quadratic term in T is 250,000, small enough to be neglected for many millennia. For simplicity in what follows, we will assume additionally that D = 0, as Craig did in the example that I provided above. If you are interested in an example for D ≠ 0, see Stigler's application of Craig's model to reports of Laplace's birthdate. [Stigler 1985; 1986; 1999] Stigler shows that Craig's model is both useful and robust. Thus in what follows, I will limit my discussion to the following simpler linear model of log posterior odds, Equation (12). The full generality offers no additional insight.

$$\ln\left(\frac{\Pr(H|E)}{\Pr(H'|E)}\right) = \alpha + Ts \quad (12)$$

If $\alpha = 40x$ and $s = -x/400$, we can rewrite (12) as the following exponential function of time, T, where the evidence E is assumed to be parametrized by time T:

$$\Pr(H|E) / \Pr(H'|E) = Ab^{-T/400} \quad (13)$$

for suitably chosen constants A and b depending on x. Thus Craig's result in modern dress says that **the posterior odds of the hypothesis that the Bible is to be believed decays exponentially with passing time, in a way parametrized by the prior odds.** Again we see the subjectivist approach: We use odds in preference to probability, and we describe posterior probability in terms of priors.

6. Using a Bayes factor.

I initially used natural logarithms in the above even though the base of logarithms wouldn't matter until the constant coefficients were specified, because I hoped to do a sensitivity analysis of Craig's results, and so I wanted to use familiar statistical tables. Weisstein says,

[Calculation:] Comparison of $[-2 * \ln(\text{likelihood ratio})]$ to the critical value of the chi-squared distribution with $n-n'$ degrees of freedom then gives the significance of the increase in likelihood.

But it doesn't work. This calculation requires several assumptions to hold (terms to be defined below): (a) the mentioned likelihood ratio must be a maximum likelihood ratio; (b) the likelihood ratio must be of so-called nested hypotheses; (c) we have a large sample of data; and (d) the natural logarithm can be accurately enough approximated by merely two terms of its Taylor series (a fact having nothing to do with Equation (12), appearances to the contrary). [Hogg & Craig 1995, 422] So there is no way that we can use this approximation.

A maximum likelihood ratio assumes that the hypotheses are expressed as parametrized probability density functions, and the maximum referred to is the maximum as the parameters vary. Stigler in his discussion of Craig assumes a maximum likelihood ratio. Yet nothing in any of my above discussion, either of spam or of Craig, or even of Charles S. Peirce's later interpretation of Craig [Nash 1991, 25], requires that the likelihood ratio in Equation (13) need be a maximum likelihood ratio. (It would have been an anachronism for Peirce to have considered maximum likelihood ratios anyway.) Jaynes [1974, 62] says that maximum likelihood estimates are “exactly derivable from Bayes' theorem ... with uniform prior probability.” Furthermore, how different can a likelihood ratio be from the maximal likelihood estimator? Not very different, as Jaynes establishes. Assuming maximum likelihood is therefore not necessary.

Is there a way around the remaining assumptions so that we can arrive at a test of significance? For subjectivist probability theory there is. Harold Jeffries defines a *Bayes factor* K involving two hypotheses H_1 and H_2 and experimental data E to be

$$K = \Pr(E | H_1) / \Pr(E|H_2), \quad (14)$$

[“Bayes Factor”]. Usually in practice the hypotheses are *nested*, which is to say that H_1 specializes H_2 , $H_1 \subset H_2$. They need not be in order to apply reasoning about Bayes factors. Substitute in (14) using $H_1 = H$, and $H_2 = H'$ (which are clearly not nested). The Bayes factor becomes the odds of getting E assuming H . Now restate Bayes's theorem as

$$\Pr(E|H) = \Pr(H|E) \Pr(E)/\Pr(H), \quad (15)$$

by interchanging the roles of H and E in Equation (1) above. Substitute (15) in (14) to get

$$K = \Pr(H|E) / \Pr(H'|E) * \Pr(H')/\Pr(H). \quad (16)$$

Thus we can see that the Bayes factor K is not dependent on our ability to calculate the probability of the evidence E , and that K is the odds ratio for the hypothesis H on the evidence E multiplied by $\Pr(H')/\Pr(H)$, which we will call β .

For the spam example of Equation (3), with H = “a new piece of mail containing the word NIGERIAN is spam,” when the first new email arrives, $\Pr(H|E) = 8/9$ and $\Pr(H'|E) = 1/9$, so $K = 8.00 * \Pr(H')/\Pr(H) = 8.00 \beta$.

In the absence of any prior reasons to believe H, we might model the priors H and H' as equally likely. We are indifferent beforehand about the hypothesis that the next email containing NIGERIAN is spam. Then $\beta \approx 1$, and $K \approx 8.00$, which according to Harold Jeffrey, gives “positive” support for the hypothesis H [“Bayes Factor”]. See Table 1.

| K | Strength of evidence |
|-----------|-----------------------------|
| < 1 | Negative (supports H') |
| 1 to 3 | Barely worth mentioning |
| 3 to 12 | Positive |
| 12 to 150 | Strong |
| > 150 | Very strong |

Table 1 Bayes Factor values for competing hypotheses according to H. Jeffrey

Assign that email to the spam category. Then, under the same hypothesis H, a new Bayes factor K_1 can be calculated, under the new evidence E_1 of yet another spam email containing NIGERIAN.

$$K_1 = \Pr(H|E_1)/\Pr(H'|E_1) \beta = 88.91\%/(1-88.91\%) \beta \approx 8.02 \beta. \tag{17}$$

We see that the strength of the evidence as measured by the Bayes factor has increased slightly from 8.00β to 8.02β .

Our spam example and Craig’s example have this in common: Neither is concerned about quantifying the evidence absolutely; both are concerned about **quantifying the change in the evidence relative to prior knowledge** (x for Craig; β for spam). They also have this in common: **They are about the state of mind of a knowing subject, not about sampling from a randomized collection.**

7. Expressing both spam and Craig's result using the Weber-Fechner law.

In the spam example above, from Equation (16) above, the Bayesian factor K went from 8.00β to 8.02β with the arrival of one new spam. We saw that in the absence of prior knowledge of the probability of the hypothesis H = “the next email will be spam” β can be taken to be 1. In the example of Craig's believability of eyewitness accounts of Jesus, his measure went down from $40x$ to $28x$, where x is an arbitrary base believability. I have used the term “believability” in the above discussion to affirm that it is a kind of measure of subjective belief.

Jaynes suggests that believability should be reported in decibels! [1976, 66] He suggests this for the same reasons that the Weber-Fechner law of psychological response holds experimentally [“Weber-Fechner Law”; Jaynes 1974, 67]. Experimental evidence is available for all but the last of the relationships in Table 2 between the psychological variables on the left and the physical variables on the right. Jaynes claims that believability should take its place in this table analogously. Jaynes defines the “evidence” for an hypothesis H as follows. So as not to confuse it with E, we

| Psychological variable | Physical variable |
|------------------------------------|--------------------------|
| ... is the logarithm of ... | |
| tactile pain | pressure |
| sound pitch | sound frequency |
| sound loudness | sound intensity |
| light brightness | light intensity |
| earthquake feeling (Richter scale) | earthquake energy |
| believability | posterior odds |

Table 2 The Weber-Fechner Law

shall continue to use the word “believability.” The believability of the hypothesis H given the evidence E is defined to be $10 \log_{10}(\text{odds of H given E})$. Jaynes uses base 10 and the multiplier of 10 so that it will be in decibels (ten bels), which is a dimensionless quantity used for sound loudness. The definition of decibels is as follows:

$$\text{dB} = 10 \log_{10} (I/I_0) \tag{18}$$

where I is a physical variable like frequency or intensity, and I_0 is a “threshold” value of I where no effect is felt (so the resulting psychological measure is 0 dB). You can see that the fit of using decibels to measure believability is not perfect, since the odds of H given E is a ratio of conditional probabilities neither of which is a base or threshold quantity. Still, I claim that $10 \log_{10} (K)$ is a more psychologically plausible measure of believability than K. Compare [Jaynes 1974, 67] In the earthquake example of Table 2, an increase of 1 on the Richter scale feels like an additive amount more of earthquake activity, but in fact it is a factor of 10 times as much amplitude. See Table 3 for the thresholds of Table 1 restated in decibels.

| K | $10 \log_{10}(K)$ |
|-----|-------------------|
| 1 | 0.00 dB |
| 3 | 4.77 dB |
| 12 | 10.79 dB |
| 150 | 21.76 dB |

Table 3 How Jeffrey's K thresholds appear in decibels

We apply this to our spam example. Since we are only interested in change of evidence, we compare $10 \log_{10} (8.02\beta)$ with $10 \log_{10} (8.00\beta)$. They differ by only $10 \log_{10} (8.02/8.00) = 0.01 \text{ dB}$. Note that β disappears, and that the difference is very small, as we might have guessed.

We apply this to Craig's example. Following Stigler, Craig's polynomial P is already a logarithm, the log (posterior odds of H given E). Since we can scale the polynomial P however we want by appropriate choices of its dimensional coefficients without loss of generality,

we can write

$$P = 10 \log_{10} (\text{posterior odds of H given E}). \tag{19}$$

In Equation (5), we saw P drop from 40x at $T = 0$ to 28x at $T=1693$. Under this rescaling of P, these numbers won't be 40x and 28x, but they will be 40y and 28y where y incorporates the scaling constant. The relative magnitudes of P remain the same. The believability of the written accounts of Jesus' life on Craig's model has dropped to 28/40 or 70% of its original believability, as shown qualitatively in Equation (6). **Believability is a psychological variable, not a physical variable.**

8. Assessing Craig's contribution to probability theory.

In a Bayesian (subjectivist) approach to Statistics, we take odds as primitive and define probability in terms of odds. This may seem like a small point because given a probability p, we can define $p/(1-p)$ as odds, and conversely, given any odds ratio o, we can define a probability as $p = o/(1+o)$. But to assume that odds and probability are inter-derivable in this way is to beg the question of whether one of them is based on subjective experience and the other is based on some supposed randomized state of the world. [Jaynes 1974] Without some additional reductionist axioms, subjective probability and axiomatic probability do not even share a

common framework of meaning. Craig was called a crackpot. His contemporaries might have said, “You didn't get anything out of your formula that you didn't already put into it beforehand.” We now know better.

Our discussion of spam began with an axiomatic development of Bayes's theorem, and then applied it in a purely frequentist way. Craig's result can be seen as Bayes's theorem applied in a purely subjectivist way. The subjectivist approach to probability begins with Craig in the late 17th C. Ian Hacking would say more strongly that the two strands began to be integrated then, but I can't find an explicit awareness of both strands actually described as complementary until John Venn (1834–1923) or Charles Sanders Peirce (1839–1914). For example, Peirce calls these two strains “materialist” (frequentist) and “conceptualist” (subjectivist). [Hacking 1984, Chapter 2; Nash 1991, 25]

Craig's contribution to this synthesis, according to Nash, is that he translated uncertainties of the universe— such as when Jesus would return—into uncertainties of the mind, which is to say, into a measure of degree of believability in when Jesus would return. [Nash 1991, 30] Craig did not have a fully developed modern notion of subjective probability, but he came close.

Craig's work prefigures that of Bayes and that of a quadratic fit to the log posterior odds for a hypothesis. Craig defined a measure of believability based on odds, a good Bayesian move, and came to a conclusion that—if we were to write it multiplicatively instead of additively—would be a decaying exponential model. Craig's creation of a mathematical formula for decay of probability is a discovery of Malthusian proportions. (All puns intended.)

Craig was regarded as being bizarre for the same reason that utilitarian philosopher David Hartley in the 18th C. might have been thought bizarre for attempting to write a mathematical formula for love between two people: $W = F^2/L$, the love of the world is directly proportional to the square of the fear of God and inversely proportional to love for God. [Kline 1953, 338] Newton encouraged his analytic method to be extended to societal issues. Subsequently hope arose that quantification could give rise to a science of society—Condorcet (1743–1794) being a prime example of a proponent of this program. The resulting “Newtonianism” went far beyond Newton, so much so that Newton cannot properly be considered a Newtonian.

9. Relating Filtering the Bible to Filtering Spam.

I illustrated the classical approach to Bayesian analysis by a spam filtering example. Craig's work can be considered a forerunning of Bayesian analysis, on the following grounds, which summarize the above discussions.

First, Craig was not as concerned to calculate absolute “probabilities” as he was to calculate the relationship between two “probabilities,” about quantifying the change in evidence.

Second, Craig was addressing the state of mind of a knowing subject, not sampling from a randomized collection. This accounts for why Craig thought additively rather than multiplicatively, as the Weber-Fechner law helps us to explain.

Third, in Stigler's words, Craig's work contains “a formula tantamount to a logistic model for posterior odds: that is, Craig's probability should be understood as the logarithm of the ratio of the probability of the historical testimony as received at the present time, given the historical hypothesis in question, to the probability of the same testimony, given the negation of that hypothesis.” [Stigler 1986]

Craig was far from a crackpot despite a the litany of criticism against him. He was far more appreciated in Germany. A friend of Newton, Craig was ranked by *Acta Eruditorum*, the

leading German intellectual journal of Craig's day, as below Leibniz but above Newton in his role in the invention of the Calculus! [Dale 2003, as cited by O'Connor & Robertson 2005] Craig's formula about believability can be transformed into a statement about probability with an exponential decay, a kind of compound interest with a negative rate of compounding. Craig did not see any connection between compound interest or logarithms and his work. But he came close.

Did Craig have a modern notion of logarithm as a function? In 1698 he published a paper integrating the logarithmic curve, so the logarithm as a curve at least was accessible to him. [Craig 1698]

Craig was aware of one of Napier's motives in inventing logarithms: To simplify calculation of compounded interest. Craig was not aware of a general pattern of transform-solve-invert that would relate logarithms and exponentials as functions. Logarithms in fact might very well be a first example of the transform-solve-invert pattern within Mathematics.² Not until the 19th C. did transform-solve-invert examples begin to multiply and generalize. That applies first to applied mathematics (the transforms of Fourier, 1807, or of Poisson-Laplace, 1815, for example) and then to pure Mathematics (conjugacy in groups). Craig did not have in view the mathematical modeling cycle from real world to mathematical model and back to the real world as another example of transform-solve-invert. [Melzak 1976, 372] I am being anachronistic in even describing Craig's work in terms of mathematical modeling.

Since we cannot adduce Bayes's formula from Craig's formula, a careful mathematical analysis of Craig's formula requires a great deal more mathematics than was available to anyone in his time. Little wonder that he was criticized!

10. Craig's approach to Pascal's wager.

The argument for Jesus' return in 3150 takes up the first half of Craig's book. In the second half of his book, Craig takes up a defense of Pascal's famous Wager (found in Pascal's *Pensées*, published in 1670 after Pascal's death). Pascal's Wager says that it is rational to trade an earthly life of happiness and separation from the presence of God for an earthy life of suffering and an eternity of joy in the presence of God. This is a kind of utility theory argument, not at all cast in terms of probability. Craig approaches Pascal's argument geometrically, just as he did the believability of Jesus' return. The argument is similar to that used by Nicole Orême three centuries earlier, according to Fernando Gouvêa's invited talk in these Proceedings:

If an intensity of pleasure is graphed along one axis, and the duration of the intensity is measured along another, then the pleasure is the area under the resulting curve. The value to a human of a pleasure can be represented by its magnitude (area). A positive pleasure of infinite duration will be more than a pleasure of any finite duration. In an honest game, one wagers one's odds. But since the odds are infinite in favor of eternal life, a Christian is the wisest person.

² One might think of an earlier example. Euclid divided a given line segment into n parts by transforming it into n equal pieces of another line segment, then by parallel lines inverting the division so as to divide the original segment. The transforming and the solving collapse in this example. I suspect that non-collapsing example must await Descartes (1596–1650) for the foundation of being explicit about solving hard problems by transforming them into easier problems.

Set aside circularities, like if there were no God, what guarantee is there of an honest game. Set aside competing worlds which this reasoning models, such as the claim of other deities to offer eternal happiness as well.

We can now offer a one-word summary of Craig's book: **hope**. Hope for Christians that Jesus will return as He said (Part 1), and hope that heaven awaits those who believe in Him (Part 2).

References

Anthony, Doctor. "Logarithms: History and Use." 1996.

< <http://mathforum.org/library/drmath/view/52469.html> >. Last accessed May 30, 2005.

Bayes, Thomas. "Essay Towards Solving a Problem in the Doctrine of Chances." *Philosophical Transactions of the Royal Society of London*. 1763.

"Bayes Factor," Wikipedia. < http://en.wikipedia.org/wiki/Bayes_factors >. Last accessed May 27, 2005.

Shows that the maximum likelihood ratio and the Bayes factor can give radically different predictions as to which of two hypotheses to believe.

"Bayesian Statistics." < <http://www.bayesian.org/bayesian/bayes.html> >. Last accessed December 16, 2004.

For photographs of Bayes and his tomb.

Craig, John. "The quadrature of the logarithmic curve." *Philosophical Transactions of the Royal Society* [of London]. 1698.

Dale, Andrew I. "Craig, John." *Dictionary of National Biography*. Oxford, 2004.

Hacking, Ian. *The Emergence of Probability*. Cambridge: Cambridge University Press, 1984.

Hogg, Robert V. and Allen T. Craig. *Introduction to Mathematical Statistics*, 5th edition. Englewood Cliffs, NJ: Prentice-Hall, 1995.

Jaynes, E. T. *Probability Theory With Applications in Science and Engineering*. Washington University. 1974.

Credibility and evidence should be measured in decibels. As cited in

< <http://yudkowsky.net/bayes/bayes.html> >. Most of the book is available at

< <http://bayes.wustl.edu/etj/science.pdf.html> >. Last accessed May 26,, 2005.

"John Napier." < http://en.wikipedia.org/wiki/John_Napier > January 8, 2005. Last accessed January 11, 2005.

Joyce, James. "Bayes's Theorem." *The Stanford Encyclopedia of Philosophy* (Winter 2003 Edition). Edward N. Zalta (ed.). < <http://plato.stanford.edu/entries/bayes-theorem/> >. Last accessed 17 December 2004.

Contains a discussion of whether to use a ratio or a difference of conditional probabilities.

Kline, Morris. *Mathematics in Western Culture*. NY: Oxford U. Press, 1953.

Lauritzen, Steffen. "More on the sequential probability ratio test. BS2 Statistical Inference, Lecture 15, Michaelmas Term 2004." December 1, 2004.
< <http://www.stats.ox.ac.uk/~steffen/bs2siMT04/si15c.pdf> >. Last accessed January 17, 2005.

"Likelihood-ratio test information," Absolute Astronomy < http://www.absoluteastronomy.com/encyclopedia/l/li/likelihood-ratio_test.htm >. Last accessed May 30, 2005.

"Bayesian criticisms of classical likelihood ratio tests focus on two issues:

"1. the supremum function in the calculation of the likelihood ratio, saying that this takes no account of the uncertainty about θ and that using maximum likelihood estimates in this way can promote complicated alternative hypotheses with an excessive number of free parameters;

"2. testing the probability that the sample would produce a result as extreme or more extreme under the null hypothesis, saying that this bases the test on the probability of extreme events that did not happen.

"Instead they put forward methods such as Bayes factors, which explicitly take uncertainty about the parameters into account, and which are based on the evidence which did occur."

Nash, Richard. *John Craige's Mathematical Principles of Christian Theology*. Journal of the History of Philosophy Monograph Series. Carbondale, IL: Southern Illinois University Press. 1991.

English translation with extensive commentary of John Craig's 1699 pamphlet, *Theologiae Christianae Principia Mathematica*.

O'Connor, J. J. and E. F. Robertson. "John Craig." 2005.
< <http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Craig.html> >. Last accessed March 15, 2005.

Stigler, Stephen M. "John Craig and the Probability of History." Chapter 13, *Statistics on the Table; The History of Statistical Concepts and Methods*. Cambridge, MA: Harvard University Press, 1999.

-----. "John Craig and the Probability of History: From the Death of Christ to the Birth of Laplace." *Journal of the American Statistical Association*, 81, 396, Dec. 1986: 879-887.

This article provides some technical statistical details omitted from the popular account in [Stigler 1999].

-----. "John Craig and the Probability of History: From the Death of Christ to the Birth of Laplace." Technical Report 165, Department of Statistics, University of Chicago, 1985.

"Weber-Fechner Law." Wikipedia.

< http://en.wikipedia.org/wiki/Weber-Fechner_Law >. Last accessed, May 27, 2005.

Weisstein, Eric W. "Likelihood Ratio." *MathWorld—A Wolfram Web Resource*.

< <http://mathworld.wolfram.com/LikelihoodRatio.html> >. 1999. Last accessed 20 December 2004.

Relates log likelihood ratio to chi square test.

Melzak, Z. A. *Companion to Concrete Mathematics. Volume 2: Mathematical Ideas, Modeling, and Applications*. Appendix. New York, NY: John Wiley, 1976.