Tina Memo No. 2007-008
Internal[1].

Tutorial: Defining Probability for Science.

Neil A. Thacker.

Last updated
28 / 11 / 2014

This document forms part of the **Statistics and Segmentation** Series (2008-001) available from www.tina-vision.net.

2007-008	Tutorial: Defining Probability for Science.
2001-007	Performance Characterisation in Computer Vision: The Role of Statistics in Testing and Design.
2002-007	The Effects of an Arcsin Square Root Transform on a Binomial Distributed Quantity.
2001-010	The Effects of a Square Root Transform on a Poisson Distributed Quantity.
2004-004	Shannon Entropy, Renyi Entropy, and Information.
2002-002	Validating MRI Field Homogeneity Correction Using Image Information Measures.
2004-001	Empirical Validation of Covariance Estimates for Mutual Information Coregistration.
2004-005	The Equal Variance Domain: Issues Surrounding the Use of Probability Densities in Algorithm Design.
2009-008	Avoiding Zero and Infinity in Sample Based Algorithms.
2001-008	Derivation of the Renormalisation Formula for the Product of Uniform Probability Distributions and Extension to Non-Integer Dimensionality.
2001-005	Model Selection and Convergence of the EM Algorithm.
2003-007	Noise Filtering and Testing for MR Using a Multi-Dimensional Partial Volume Model.
2002-004	A Novel Method for Non-Parametric Image Subtraction: Identification of Enhancing Lesions in Multiple Sclerosis from MR Images.
2001-014	Bayesian and Non-Bayesian Probabilistic Models for Image Analysis.
1997-001	The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data.
1999-001	The Bhattacharyya Measure requires no Bias Correction.
1999-004	B-Fitting: An Estimation Technique With Automatic Parameter Selection.
2005-008	Tutorial: Beyond Likelihood.

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

[1]This document is dedicated to my grandmother, who taught me to play cards and who died on 21/10/2007, aged 95.

Defining Probability for Science.

Preface

For many years I have recommended the short reference on statistics [2] and similar introductory texts to my students and researchers. Barlow's book can be considered a fair reflection of the main stream view of the topic [14]. Yet despite this, I generally gave the caveat that I disagreed with some parts of the book and indeed the conventional view. On a recent reading of the book I concluded that many of my objections were confined to chapter 7, and in particular the discussion regarding definitions of probability. I therefore decided to write this document to provide a record of how I would have presented these sections, for distribution within our group. The document explains a physics based approach motivated by scientific considerations of uniqueness, falsifiability and quantitation. These considerations are intended to eliminate aspects of 'black magic' or arbitrariness, a view which seems to me to be important yet lacking from general texts. It summarises what I regard as the reasons I work as I do when designing and testing algorithms and systems for computer vision and image processing.

Although this document is self contained, the interested reader might wish to look at the original version first, before reading mine. You would then be in a good position to decide if you want to continue to take the conventional view of the topic, or take the rather bold step of being more critical and forming some conclusions.

Introduction

The accepted method for all scientific data analysis is probability and statistics. Although all statistical techniques are related to probability the very definition of this word is not generally comprehended. In my view, the resulting lack of clarity on this issue leads to misunderstanding and inappropriate application of techniques. In subject areas where the analysis of data is the dominant activity, real progress is hampered by a lack of concensus regarding best scientific practice. This situation is not generally improved if we look in standard statistical reference texts for more clarity. The dominant attitude to statistical methods being that we can largely pick various measures out of thin air and worry about how they behave on data afterwards, rather than deriving techniques from principles based upon the characteristics of the data. There is in my view an over-tolerance of contradictory opinions and general failure to resolve fundamental issues.

To do better we must start with a solid understanding of probability. Even here the common perception that there can be multiple (often contradictory) definitions does little to help general understanding. In the sections below I will give a quick summary of the issues which a short investigation is likely to uncover, together with an attempt to explode some of the more distracting arguments.

There are multiple definitions of probability in common use. It has been said that each has its strengths and shortcomings. However, what I aim to show here is that the motivation for some of these seems to be based upon flawed logic. Further, for purposes of quantitative analysis of data, only one approach has any real relevance and would appear to support everything we need to be able to do. This is not just an academic exercise, as if the conclusions of the analysis are accepted it has implications for large numbers of papers published every year in many areas of research.

Mathematical Probability

For a set of mutually exclusive set of events E_i we can define probabilities $P(E_i)$ such that;

- $P(E_i) >= 0$
- $P(E_i \text{ or } E_j) \ = \ P(E_i) \ + \ P(E_j)$
- $\sum_{\forall i} P(E_i) \ = \ 1$

These are Kolmogorov axioms, which can be found in equivalent forms in many standard texts. It has been observed that although you can use these axioms to derive several common results, for example;

$$P(E_i) \ = \ 1 \ - \ P(\tilde{E}_i)$$

they are devoid of any real meaning. The definitions below do not contradict these axioms but aim to provide the concept of probability with meaning.

The Limit of Frequency

Defining probability as a ratio of events is often referred to as the frequentist definition and is the one with which scientists will be most familiar. For an example, if an experiment is performed N times and a certain outcome E_i occurs in M of these cases then as $N \to \infty$ we can say $M/N \to P(E_i)$. The set of N experiments was called the collective by Von Mises in his influential work.

As simple as this definition appears it embodies certain features which have led to confusion and criticism. In the first place the defined probability is not just the product of a single experiment but a joint property of both the derived data and the collective. We can illustrate this by a simple example, inspired by Von Mises original exposition. We can observe the life spans of male members of a population (say German) and from this deduce the probability of an individual dying between their 40th and 41st birthdays. However, if this individual was a taxi driver we could equally have accumulated statistics and compute the probability over a collective of taxi drivers. Clearly this would result in a different probability. The observation that there are multiple ways of defining the probability of the same outcome poses a problem for scientific uses, and can be seen to be one of the motivations behind alternative definitions of probability. However, appropriate use of frequentist probability requires some care. We need to first consider if this is a valid motivation, or just a failure of understanding. I believe that it is the latter, as will now be explained.

A change in the estimates of probabilities between two scenarios, such as the taxi driver/German should come as no surprise to us. What this example illustrates is the difference that information can provide when interpreting data. Ultimately if we knew everything possible about an individual, including having access to medical records, we could get a very good idea of the prognosis for the individual. We need to have a rational way to make use of varying sources of data in order to build predictive systems. It appears simply naive to believe that probabilities should always be the same for the same outcome when we are attempting to make predictions based upon knowledge (see J.G. comments). Thus, although the 'probability' appears to change as a function of the collective, the concept would be of no value to us if it did not. We cannot therefore use this as a criticism of frequentist probability as it appears to be a necessary characteristic. This also contradicts Jaynes' lengthy (and highly influential) exposition on probability which concludes with the criticism "probabilities change when we change our state of knowledge, frequencies do not. ". The frequencies change if the cohort has changed, to say otherwise suggests a pre-conception of what the probability was regarding. A common illustration of this is provided by "Bertrand's paradox", where a poorly defined system leads to multiple estimates of the probability for the length of a randomly chosen chord within a circle. This isn't really a paradox at all, but simply an illustration of how not to use probability. The key here is to realise that the information we use must be consistent with the collective defining our probability, but there is no need (at least at this point) to introduce a fundamental break between frequencies and probabilities.

Scientific Considerations: Repeatability

According to Von Mises, the experiment must be repeatable, under identical conditions, with different possible outcomes in order to observe an empirical probability. In situations where these restrictions are not met, such as the phrase "It will probably rain tomorrow", which embodies an event which can only occur once, Von Mises was highly critical of the use of the word probability, regarding it as unscientific. Though it is right to consider scientific validity, we can make observations here regarding the generality of this conclusion. As with the previous example, I will make an equivalence between probability and prediction in order to demonstrate this.

Imagine that we construct a mathematical model of the weather, there is no reason why this cannot be done according to our best scientific understanding. We could then run a series of predictions under the allowable variations within the uncertainties of our knowledge, and so generate a cohort of "experiments". We can then define the probability of rain $P(rain)$ in a manner fully consistent with Von Mises' requirements which results in the phrase "It will probably rain tomorrow". This process is known in science as running a Monte-Carlo simulation. If we know enough about analytic forms of various perturbations, and can manipulate the resulting mathematical description, we may even be able to predict the results of the Monte-Carlo from analytic expressions. We would call this a statistical theory.

Interestingly, this probability is now equivalent to the process of placing a bet on an outcome, which others have repeatedly associated with the concept of subjective probability (see below). Yet, if we also have a good model then we would expect to find that the probability predictions would prove to be quantitatively valid, ie: it would be wrong in probable rain predictions $1 - P(rain)$ of the time. The idea (which has been called "honesty" [6, 12]) is consistent with a frequentist approach. The only thing we must accept here is that our definition of probability is driven by the uncertainties in our knowledge and embodies our predictive capabilities rather than some more ethereal concept of the chances of seeing precipitation. The perceived problems with regard to the

restrictive requirements of probability therefore become an inability (once again) to conform to a preconception of what we considered our probability to be regarding, rather than a genuine failing of the frequentist definition. This observation is fundamentally important to those who try to build predictive (you might also say intelligent) systems for a living, as we do not need to abandon a frequentist definition of probability when we are designing them.

Mathematical Considerations: Convergence

Mathematicians have been reluctant to adopt the frequentist based definition of probability. They have shown there are sequences of numbers, which otherwise conform to mathematical definitions of random variables, that will not converge to the required probability fractions in the limit of large sequences. However, such theoretical problems are at odds with the practical observation of "Poisson", "Multi-Nominal" and "Binomial" distributions, which can be found to converge exactly as expected in real world problems. You could therefore reasonably make it a condition of the definition of frequentist probability that the associated random number sequence must conform to a process of convergence. The convergence issue then becomes a problem for mathematicians to sort out between themselves. Though this might seem a heavy handed dismissal of an important mathematical issue, it has prior precedence in topics such as the use of Dirac delta and Heaviside step functions in physics. Here physicists used these techniques while mathematicians continued to work out the formal basis.

The easiest way out of this is to say that the very concept of a mathematical sequence is inconsistent with the idea of true randomness, and thereby probability. Genuinely random sampling systems will converge in accordance with the formulae from conventional statistics, which can be derived without the need for an assumption of convergence [11]. Of course, once we say that a random number is not describable by a mathematical sequence we severely restrict the opportunities for the mathematical analysis of probability. However, avoiding this intellectual bear-trap also allows us to go on to relate conventional definitions of probability to limiting cases of the accepted statistical distributions. Class labels will have probabilities defined by the Binomial process and probability densities can be considered as an extension of Poisson sampling. We can then eliminate many of the contradictory methods used for the comparison of probability density distributions found in the area of statistical pattern recognition [19]. Without this step, many of the choices for the selection of appropriate probability similarity functions remain arbitrary. As a consequence there will also be many (potentially infinite) ways of defining a computational task. A scientist might see such arbitrariness as enough reason to dismiss the convergence issue, and effectively this is precisely what anyone using modern statistical theories of physics has already done.

Objective Propensity

The philosopher Karl Popper worried about the convergence of frequentist probability, but also had difficulty with the concept of a collective. In particular he said that science could predict only probabilities and not certainties and this would lead to problems. The often quoted example is quantum mechanics, where you could say that if the computed probability were to change as a function of the collective, chosen by the experimenter, then the particles they describe could have no real properties, behaviour or existence. Such observations led Popper to propose an objective probability, or propensity, which exists in its own right (ie: it is unique) with the only observable effect being to drive the observed frequency limit. Objective probability seems a very reasonable strategy when considering simple unique cases, such as the throwing of a die. In such situations the introduction of a collective in order to describe the system does seem to be excessive. Popper was clearly correct to consider this issue, and a non-unique theory must be considered unscientific, but if his main concern was quantum mechanics, he need not have worried. The possibility of one being able to redefine the collective is made impossible by the use of quantum numbers and quantum states. These provide a complete description of the physical process and leave no room for subjectivity regarding the collective. Indeed a theory which could make multiple predictions would already have been considered unscientific by the physics community and rejected at an early stage, as would any other statistical physics model such as statistical mechanics or quantum electro-dynamics. The requirement that physics theories should make unique predictions has been well understood since Einstein used the principle of equivalence to develop general relativity. This is a very powerful test of any theory, and demands that any theoretical predictions must not change (it must be invariant) following arbitrary re-definitions by the scientist of experimental circumstances (such as metric units, or the bending of space-time due to selection of co-ordinate frame). But then we could not expect Popper to have known this, he was a philosopher and not a physicist.

If my reading of the arguments for objective probability is correct then propensity is nothing more than the frequentist definition of probability, restricted to cases where there can only be one collective. As there appears to be nothing at all wrong with the other cases, and on the contrary we need these other cases if we are to understand how to construct predictive systems, this suggests that propensity has nothing to offer us.

What this episode in confusion does illustrate once again, is that when defining and computing probabilities we must take care to declare everything, and not just brush implicit assumptions under the carpet. Otherwise we might make the mistake of believing two expressions should be equivalent when they are not. If we accept that differences in assumptions lead not just to differences in numerical values, but probabilities of different circumstances, then there is no contradiction in the frequentist approach. As we will see, one way to keep track of these assumptions is via the use of conditional probability.

Conditional Probability

We will now introduce the concept of conditional probability and explain how this can be used to eliminate the ambiguities often arising during the design of algorithms. Also, we cannot discuss the concept of subjective probability without first defining conditional probability.

The conditional probability $P(E_i|X)$ is the probability of getting event E_i given that X is true. If we consider the example above of the probability of mortality for a German male, what we see is that the notation is different if we make the change from a German male X to a taxi driver Y, and in general we have $P(E_i|X) \neq P(E_i|Y)$. This notation prevents us from making the mistake of believing that these two quantities could ever have been the same. The use of conditional probability therefore eliminates a large source of confusion when we are trying to construct probability based systems.

Bayes theorem is generally attributed to Rev. Thomas Bayes (1763), though it was work unpublished in his lifetime and published posthumously by his daughter[2]. It uses the construction:

$$P(a|b)P(b) = P(ab) = P(ba) = P(b|a)P(a)$$

to give

$$P(a|b) = P(b|a)P(a)/P(b)$$

We should appreciate that the derivation of Bayes theorem is entirely consistent with a frequentist definition of probability and we do not need to rush into a subjective interpretation. It is simple enough to construct situations in which Bayes Theorem can be shown to be quantitatively valid, for example the use of Bayes Theorem in pattern classification [8]:

$$P(C_j|data) \;=\; P(data|C_j)P(C_j)/\sum_i P(data|C_i)P(C_i)$$

where each C_j is the potential generator of the observed data. Each term in this expression can be determined from samples of data. As we shall now see however, not all uses of Bayes theorem have the same property.

The Restrictions of Conditional Notation

There is a fundamental assumption pertaining to the characteristics of real systems which is easily overlooked. Once both a and b are known to be true Bayes Theorem makes no distinction regarding the order in which they occurred. In many real world situations events don't happen simultaneously but in a specific time order. If there is a causal relationship between a and b, such that a can only occur if b has already happened, then $P(a|b)$ has an obvious physical interpretation which follows the implied order of conditional notation (the probability that a will occur as a consequence of b) but $P(b|a)$ does not (the probability b was the cause of a). Otherwise identical notation therefore gives two different physical interpretations. Conditional notation describes correlations, not causal (physical) processes. This renders a strict physical interpretation impossible if we do not know the causal interpretation a-priori. Worse, for systems of mixed causality (a causes b *and* b causes a) any resulting expressions are physically meaningless. In fact, and as a consequence, conventional conditional notation can only ever be able to describe non-ordered sets of events [3], and is therefore poorly placed to construct models of causal systems. If we decide to use the order of the terms in expressions to encode causal processes we will observe immediately that in general $P(ab) \neq P(ba)$, this has consequences for our derivation of Bayes theorem. In comparison, the various methods of Markov modelling deal with this issue directly by adopting a state to state transition process such as used in statistical physics, which may explicity allow bi-directional causality with different rates and also specifies the cause. Such a process may be fundamentally important if we are ever to construct intelligent systems [17].

With some work, the above observation can be used to explain Jaynes' conclusion regarding frequentist probability. His arguments attack the inability of standard frequentist models with regard to their inability to embody prior

[2]The statistician Fisher seemed to believe that the work was left unpublished by her father for a reason [7].

[3]A limitation understood by Kolmogorov, but not appreciated by Popper [11], who attempted to define probabilities for sequential statements.

knowledge in the prediction of sequences of numbers. These examples are valid but if you look at his arguments carefully[4] these observations can be interpreted, not as a flaw in the definition of frequentist probability, but as the inappropriate treatment of ordered sequences as non-ordered sets. Mathematicians have introduced a principle which is intended to avoid such misuse of conditional notation which is referred to as the "precluded gambling system". This simply states that correct use of probabilities should prevent the specification of an alternative calcualtion which is capable of beating the computed odds in practical use. Moreover, such problems can be overcome by specifying any conditional probability relating to sequences so that it includes the temporal context along with associated knowledge of the data generation process , as for example in the use of an "embedding" (where a sufficiently long sequence of values from a deterministic sequence is used to predict data from a kinematic model). This results in a more specific cohort which is again consistent with a non-ordered set. What we see is that appropriate use of conditional notation is essential if we are to avoid such problems.

While we are on this point, we need to discuss the case where we have no conditional, ie: the prior probabilities. According to Popper (se P.B. comment 1) the only ways to define a prior independent of a collective would appear to be to either base it on no information at all i.e. use the equally likely definition, in which case expressions derived from Bayes Theorem revert to manipulations of likelihoods, or to base it on all possible information that has, or could ever, be obtained. In this case, the prior can only take the values 0 (the theory is wrong) or 1 (the theory is correct). This conclusion is valid regardless of which definition of probability you choose. Any concept of actually measuring a prior, either subjectively or objectively based on previous data, results in a prior which is itself a conditional on the collective used to define it. We must therefore consider the consequences of generalising this concept, either to include uniformtive (fixed value) scaling priors, or informative (varying over the parameter itself). From a frequentist point of view, an uniformative prior (a number) such as those used in pattern classification, can be used in a way which is consistent with measured data. We will discuss the problems associated with prior distributions below.

Subjective Probability

Subjective probability has been advocated as a way of solving three problems;

- providing a meaning for phrases such as "It will probably rain tomorrow".
- as a method to incorporate prior knowledge into our analysis.
- a way of describing the how humans think.

The fact that we can already resolve the first with frequentist probability has already been discussed above. The second and third motivations will be covered in the discussion.

The subjective interpretation of probability (sometimes also called Bayesian Statistics) makes no claim other than we can no-longer expect our probabilities to be quantitatively related to data. It reduces probability to a non-quantitative ranking process. Note that in order for the idea to have merit, it should not be considered as a simple linear rescaling of frequentist probability. This is a common step found in many frequentist analyses, including Likelihood.

The idea of subjective probability is exemplified by the interpretation of evidence in an experiment, such that new evidence $P(data|theory)$ is used to update our degree of belief $P(theory)$ (making it stronger or weaker), according to [5]:

$$P(theory|data) \;=\; \frac{P(data|theory)}{P(data)} P(theory)$$

The above formula looks at first sight to be analogous to the previous Bayesian formula for pattern classification, which is frequentist. Whaever our definition of probability, the term $P(data)$ must be a constant for a fixed data sample. If we were to attempt to interpret this as a frequentist process we must take $P(theory|data)$ to be the probability of the theory being the generator of the observed data in comparison to the other theories in the set implied by our prior knowledge $P(theory)$. As we know that in reality, there is only one correct theory, we can reasonably say that this process is non-physical. This is an act of imagination similar to a Monte-Carlo simulation of the weather, except as our initial degrees of belief ($P(theory)$) do not correspond to anything we can objectively

[4] Specifically his discussion of the uninformed robot which forms the centrepiece of chapter 9. This is just one aspect of Jaynes' work, and it is not my intention to attempt to write an entire book which analyses the validity or otherwise of every point. Instead I will later prove his central conclusion to be false.

[5] Why $P(data|theory)$ should be defined according to a frequentist process but $P(theory)$ is allowed to become a "degree of belief" has always been beyond my comprehension.

define, they are arbitrary. Adoption of a subjective definition for probability deprives us of the ability to test the validity of any expression empirically, so that we must expect to have to rely entirely on mathematical consistency. This is not as simple as it sounds.

The Logic of Subjectivity

There appear to be several limitations associated with combining subjectivity and conditional notation. Significantly, the notation of conditional probability makes no provision to record the prior probability which was used in the construction of a particular degree of belief. If we need at any point to equate two expressions of the form $P(b|a) = P(b|a)$, or to take differences between differing hypotheses $P(b|a) - P(c|a)$ [22] then any resulting algebraic expression is meaningless unless we can be sure we used equivalent prior assumptions in all previous steps. We have re-introduced the problem of multiple possible values for the resulting probability (which we had only just eliminated by introducing conditional notation). Advocates of the physiological motivation for "degrees of belief" should pay particular attention to the latter limitation, as it implies that there are situations in which we cannot assess the "degrees of belief" for alternative possible actions in order to achieve an outcome.

There is another related issue which deserves mentioning and also has implications for Jaynes' earlier conclusion. The idea of using an equivalent prior is closely related to the notion of consistency in logic. As our ideas of probability must be considered as an extension of logic, any result from the theory of logic must also apply to probability. In the area of predicate logic there is a well known result which states that you can prove any assertion (including contradictory ones) if your data base of facts is not consistent [9] (ie: contradictory). That is, logical consistency of the system is a property both of the mathematical statements and the knowledge database. If we are free to define prior probabilities arbitrarily there will be some *mathematically* consistent statements which are not *logically* consistent. Thus we can not rely solely on mathematical rigour to ensure that expressions we write are meaningful. Under the frequentist definition of probability this situation can be avoided by demanding that our probabilities are descriptive of the real world, which must be consistent [6]. We should never accept that it is safe to break this link between probability and observation. Neither can you expect that the formal notation of mathematics will tell you there is a problem if you do.

The situation gets worse for strong Bayesians. The same reasoning also implies that if we are using degrees of belief we cannot write $P(a|b)P(b) = P(b|a)P(a)$, in order to derive Bayes Theorem. As the priors on either side of this expression can not be guaranteed to be consistent we can not expect the statement to have general validity. Popper had a much shorter argument which is applicable here, it simply states that once you have taken the subjective definition of probability no further algebraic manipulation is possible, as there is no meaningful expression to manipulate [7] (see comment 1 by P.B.). All treatments of Bayes Theorem I have seen sidestep the problem by deriving the result for frequentist systems and switching to a subjective definition after (as in this document). If you want to be a strong Bayesian you must assert Bayes theorem by fiat, thereby destroying any argument of intrinsic validity (see the discussion regarding the inevitable use of subjective priors)[8].

Discussion

We have already seen enough to suspect the validity and general utility of subjective probability. By replacing a well constructed physical (frequentist) model with an intuitive (subjective) process you could argue that we can broaden the scope of our theory and therefore its utility. However, unlike Popper's fears regarding quantum mechanics, this really is an arbitrary process which correctly deserves Von Mises' earlier criticism of being un-scientific. We also arrive at the same conclusion from arguments of logical consistency.

Motivation for the use of subjective probability based upon difficulties with understanding the sentence "It will probably rain tomorrow" and convergence problem seem to have no foundation. As these are due to a poor understanding of what the probability is describing, then in general there is no problem to solve. We must now consider other popular justifications to see if we must modify our opinion.

[6] For example; $P(abc) = P(ab|c)P(c) = P(ac|b)P(b) = P(bc|a)P(a)$, requires a consistency constraint between $P(a)$,$P(b)$ and $P(c)$.

[7] I know, it is difficult to accept that entire communities of mathematicians, engineers and scientists could have overlooked this criticism for so long. Perhaps Popper's statement was just too subtle.

[8] Could this be the reason why Bayes didn't get around to publishing the idea before he died?

It Seems to Work

The subjective definition is often put forward as the only way to make use of prior knowledge. The first objection many will have to any criticism of the use of subjective probability for the combination of evidence is that "It seems to work". This is the same level of justification which was made for "fuzzy logic" in its heyday. Eventually, people began to realise that all successes of this approach could be attributed to its approximation to probability[9].

Be aware, we do not need to believe that it is necessary to adopt a subjective definition of probability in order to combine data. Alternatively, notice that if our prior knowledge was in the form of a previous measurement $P(previous|theory)$ we could have performed an equivalent (but unique) combination process such as:

$$P(result, previous|theory) \;=\; P(result|theory)P(previous|theory)$$

This has analogous structure to Bayes Theorem (you get an updated result by multiplication with a term derived from 'prior' knowledge), but is the conventional basis for the combination of measurements. We don't need much imagination to figure out what will happen if otherwise arbitrary Bayesian priors are selected in order to optimise a quantitative (therefore **frequentist**) performance measure. Bayesian estimation becomes a mis-understood re-invention of Likelihood. Persisting in the Bayesian interpretation may help get the ideas published but just prevents us from developing a valid understanding of the reason for the apparent success. After all, what is wrong with simply saying that our prior knowledge should be interpreted as the Likelihood term from a previous measurement? We can generally perform tasks such as data fusion using several different frequentist approaches, including covariance combination [10], hypothesis combination [4] and probabilistic mapping [18]. Understanding how to interpret this approach in the context of a scientific experiment requires knowledge of confidence intervals.

When considering the problem of defining prior probabilities, attempts to obtain them from sample data looks more like a frequentist approach but cannot be expected to solve this problem. Firstly, we have the difficulty of getting a non-arbitrary data sample. Believing this is even possible often amounts to wishful thinking. Also, for continuous parameters (which is where people generally want to apply this approach), there is a problem which is related to our earlier observation regarding invariance of theories. This is that we can always apply arbitrary non-linear transformations to our parameter definitions (without changing the theory), which will change any sample distribution we measure. Barlow comments that when applying Bayesian methods, the differences you get when changing for example between mass m and m^2 are inevitable, and need to be remembered. In fact a sample distribution is strictly a probability density in the chosen measurement domain. It only becomes a useful probability when we can uniquely define the interval over which to integrate this density[10]. Given an interval the transformation of parameters has no effect on the computed probabilities,

$$P(x_1 < x < x_2) \;=\; P(f(x_1) < f(x) < f_(x_2))$$

as the interval itself transforms to preserve the result. $P(theory|result)$ can never be considered a physical theory unless we insist that $P(theory)$ is constructed using an appropriate interval.

I have yet to meet a strong Bayesian who can explain to me whether the subjective definition of probability does not necessarily imply that densities and probabilities are interchangeable. The conventional approach found in many papers does not contain a distinction. This is perhaps one of the observations which prompts some eminent researchers to say that the use of probability theory inevitably results in some degree of arbitrariness. However, this is simply not true (ask any physicist), it is the unquestioned use of probability densities as probabilities which causes arbitrariness. This is why our own group always uses a notation which expresses probabilities with upper case P and densities in lower case p for derivations.

Inherrent Subjectivity

Many people have made the argument that frequentist approaches to the analysis of data hides subjective decisions which Bayesian approaches make explicit. This argument is exemplified by the comment that Likelihood, for example, is simply Bayes theorem with the arbitrary choice of a uniform prior. The firm belief that this is the best description of Likelihood continues and dominates in many research areas. These arguments are based upon the subjective interpretation of Bayes theorem, which justifies the existence of a subjective distribution over possible theories as an inevitable necessity for the calculation of $P(theory|data)$. This interpretation is preferred by many even though these probabilities are not testable within the framework of the scientific method. However, If Bayes

[9] This is not a random choice of topic for comparison. I do not believe that it would be too inaccurate to describe conventional use of subjective probability as fuzzy logic with conditional notation.

[10] For an explanation of how this problem is avoided in Likelihood construction by maintaining a formal link between probability and probability density see [19].

theorem cannot be derived for subjective probability, then a subjective $P(theory|data)$ is meaningless and there is no grounds for demanding the existence of a meaningful subjective prior.

We must not forget the frequentist interpretation of Bayes theorem, and this can be invoked to account for meaningful (non-subjective) priors (see P.B. comment 1). This is logically consistent with the use of conditional notation, provided that all priors are uninformative. The valid transformation of probabilities also ensures they are uninformative in all descriptions of the system, even in the absence of a defined interval. Such a process is necessary if we are to accommodate quantitative forms of prior knowledge which fall directly from the theory, like the number of parameters and their ranges for example. Under this interpretation, the uninformative prior allows us to compute $P(theory|data)$ from the Likelihood $P(data|theory)$. However, for the reasons given above, one can not and should not argue from this clear frequentist case to the more general validity of arbitrary (informative) priors. Unlike the Bayesian approach, the frequentist case for the use of the Likelihood term is unambiguous, it simply contains the evidence provided by the data towards the theories under consideration. This fits well with the application of the scientific method via the use of confidence intervals. Additional prior knowledge on parameters in a model can also be obtained from previous data in the form of additional likelihood terms. This interpretation is consistent with the arguments regarding misinterpretation above. It has all of the properties that the advocates of subjectivity want so that they can get on with designing algorithms, with none of the penalty regarding scientific interpretation. It does however, require that these terms are derived in ways which are consistent with the use of quantitative probability and this generally requires some degree of training. This counts against widespread understanding of the techniques in comparison to the subjective approach, which can have no associated methodology beyond the mastery of conditional notation. If we use our inability to understand frequentist probability as an excuse to abandon it we are loosing our most powerful experimental tool, the ability to test.

Unfortunately, as I mentioned in the introduction, the true situation regarding subjectivity is not particularly helped by the available texts. In [13] we find the following:

In the end, even a strict frequentist position involves subjective analysis... The **reference class problem** *illustrates the intrusion of subjectivity. Suppose that a frequentist doctor wants to know the chances that a patient has a particular disease. The doctor wants to consider other patients who are similar in important ways - age, symptoms, perhaps sex - and see what proportion of them had the disease. But if the doctor considered everything that is known about the patient - weight to the nearest gram, hair colour, mother's maiden name, etc. - the result would be that there are no other patients who are exactly the same and thus no reference class from which to collect experimental data. This has been a vexing problem in the philosophy of science.*

This argument, and ones like it, are quite common in reference texts. It implies two different definitions for the word subjective, and in any case we cannot justify subjective probability just because we think frequentist probability is flawed. We should identify any meaningful criticisms and show how an alternative approach answers them.

However, we can see from this example that there is no criticism to answer. Each conditional definition can be legitimately treated as its own quantitative assessment of probability. Then, as the passage states, the set of factors we should take into account are those which will be most important (informative). We need only to progress a little further with the analysis to realise that identification of the best predicting system is objective, not subjective [21][11]. The reference class problem is entirely due to our preconception of what probabilities should do for us, not the definition of probability itself. In the case above we are encouraged to assume that the *only* probability possible should automatically solve our diagnosis task optimally. Whereas in fact, as we have seen, different quantities of data will (indeed must) allow us to form more or less informative decisions. As for the curse of the dimensionality in pattern recognition, any solution (here the choice of conditional statements) must factor in the resulting sample size so that, for example, we will never choose the empty reference set.

The Way We Think?

The strongest proponents of the use of degrees of belief often make the claim that this is the way that people think. As people do not perform reasoning tasks well this at least sidesteps the observation that subjective probability is unscientific. If this claim were true then AI researchers would ignore this approach at their peril. However, I believe that the interpretation of brain function in terms of degrees of belief as an explanation for subjectivity might be rather over simplistic. As well as assessing evidence, decision processes must take into account something referred to as "Bayes Risk". This is the amount of importance that we want to attribute to particular categories of outcome such as effort, pain, cost, resource, and time taken. As these outcomes are non-commensurate, any decision which aims to trade off such quantities must be in some respect subjective. Not understanding how a decision is made is different to saying that the decision process involved is not attempting to be quantitative. Apparently completely

[11] For this specific example we can say that the most informative sample will be the one which identifies most clearly the diagnostic choice by minimising the risk associated with a decision.

subjective decisions, e.g.: should I eat an apple or orange?, can be interpreted as attempting to maximise quantities of specific neuro-transmitters. Often, our idea of subjective belief takes on a quantitative (therefore frequentist) consequence which would be precluded from a strict Bayesian interpretation of brain function. For example, someone might have a subjective idea of which horse is likely to win todays race, but the decision to place a bet requires that this is related to quantitative betting odds.

The above arguments are intended to show that any claim for the role or subjectivity cannot be made until we understand the computational processes involved, but this is not the main objection to using a subjective argument to describe brain function. As we have seen above, we can have no reason to believe that the comparison of different degrees of belief as a process of decision making is at all meaningful. We also need to consider how a Bayesian system dealing in degrees of belief could ever get constructed. There are only two possible ways that the parameters in a brain can be established, via long term evolution or through shorter term learning [12].

- Evolution requires that we make decisions which maximise our chances of survival, this task is frequentist by definition. Shouldn't evolution have forced a frequentist solution?
- We believe that learning in intelligent systems proceeds by modification of data stored in memory, subject to observation of data. If the terms in our theory are degrees of belief and cannot be related to samples of data what is left as the theoretical basis for learning?

If human thought processes were entirely based upon subjective probability then the thought process must arrive at decisions which do not maximise the proportion of times we achieve our required outcome, as this is a frequentist definition. However, this does not need to remain a philosophical argument, the claim of subjective decision making processes is testable. Ashby and Perrin performed recognition experiments using line based shapes [1] and were able to show that in this simple case at least the categorisation process was consistent with a frequentist definition. It is also entirely possible to build frequentist models of brain function which are based directly upon the measured behaviour of neurons [15], and these models can be extended to provide learning systems based upon a frequentist interpretation of Bayes Theorem [16], for classification and predictive tasks. For more complicated tasks, it's not enough simply to invoke a definition of subjective probability to cover up problems with non-commensurate objectives and the fact that we are guessing at what would otherwise be frequentist terms. Indeed, this approach is the one adopted by imaging scientists as the only (and pragmatic) way they can see to justify the use of 'Bayesian' methods [3] in practical applications, though they also say that this approach is likely to be unpopular with either the Bayesians or the Frequentists.

The subjective interpretation of brain function may work as an analogy, but simply cannot be considered a scientific theory.

Use of Probability in Science

A basic consideration of the scientific method tells us that probability needs to be quantitative (see P.B. comment 2). A subjective definition of probability should never be allowed to find its way into *any*[13] physical theory. It therefore undoes most, if not all, that frequentist probability did to make Kolmogorov's axioms useful.

For any practical use, the frequentist definition of probability is the one that matters as, in contrast to the subjective approach, it can be applied to circumstances of varying information in a quantitative (and therefore testable) manner. Objective probability appears to be encompassed by this concept. Popper's work [11] was highly influential in its day, and probably remains the most thorough discussion of the topic. However, his analysis concentrated on the logical superiority of falsification over induction as a process and did not appreciate the potential ambiguity of mathematical notation with respect to causality. Since 1950 you will get a better understanding of the use of probability in science from a physics degree. A frequentist interpretation of Bayes Theorem, though possible in some cases (such as mixture modelling based upon Expectation Maximisation [20]), should be quantitatively testable. Researchers need to be vigilant regarding the application of Bayes Theorem to causal systems [4]. Theories must therefore conform to our ideas of physicality in order to be considered valid. The main advantage of this approach is that anyone can do useful research, even those new to algorithmic design, it just requires the effort to test ideas using appropriate Monte-Carlo experiments [5]. Any remaining confusion is eliminated by confirming the predictive capabilites of any theory with data. This has to be the ultimate test of the value of any work and dispels any residual problems we might have with semantics.

[12] Notice I exclude the possibility of arbitrary selection by an external agent such as a researcher.
[13] I take this to include quantitative analysis of data.

Conclusions

I hope this summary of the concept of probability has left the reader with at least some comprehension of the level of confusion which has been allowed to continue with regard to a workable definition. The reader should now be aware that Kolmogorov's axioms alone are of no immediate practical value, and the way that the frequentist definition of probability relates to the objective and subjective definitions. What we have seen is that it is particularly important to be specific regarding the definition of any probability and the assumptions on which it depends. This can be monitored to some extent by the use of conditional probabilities and the associated notation.

In order to understand the problems associated with subjective probability, we do not need to look at large volumes of publications in this area and start looking for mathematical flaws, as our observations regarding logical consistency tell us we would not expect to find any. By adopting a subjective definition of probability, mathematicians have developed a system which cannot be challenged by data or notions of algebraic consistency. Equally, and for exactly the same reasons, it is also of no practical value[14]. Some readers may have already noticed that this is just a new perspective on an idea which dates back to Galileo. We cannot build meaningful theories of the real world without experimental data. If we could we would still be trying to predict planetary orbits using the geometry of perfect solids.

Upon reading this document you might conclude that I am a *frequentist* rather than a *Bayesian*. However, as my demand for a quantitative theory is driven by scientific requirements, rather than the mathematical choice of a specific set of axioms, I prefer to call myself a scientist. This is why I have given this document the title "Defining Probability for Science" [15]. The definition which is appropriate for this task is the frequentist one. The subjective definition fails on not one but three counts; invariance, quantitation and logical consistency. The take home message is that you only have to understand and agree with **one** of these arguments to see that this is a problem. The simplest way to identify these circumstances is to consider the effects of the arbitrary decisions made during the construction of any theory. If simple changes in these decisions (such as co-ordinates) change the final output of the system then the approach has no validity. Any systematic methodology for the use of probability must be free from such issues, wherever it is applied.

Applying probability in a quantitative manner is difficult for most people. The many motivations for a subjective definition (such as the reference class problem) seem to arise when we are unable to see how to get the frequentist definition to work. It is then more convenient to assume that the problem is with the definition of probability rather than our own comprehension. In the research areas of computer vision, image processing and pattern recognition, people throw arbitrary prior probabilities into systems to solve quantitative (frequentist) analysis tasks using a Bayesian (subjective) justification, without appreciating the logical contradiction. In other areas too we see the same process at work, Jeffreys used the argument for *subjective* priors in order to justify the extra terms needed to regain a unique (scientific) interpretation during the *quantitative* analysis of data using Likelihood (note the contradiction) [16] [19]. Others have included prior terms in order to solve the problem of model selection, though the subjective justification is again contradicted by the testing used to quantify generalisation [3]. Others use priors to bias the results of calculations so that they will be closer to those expected (eg: MAP), not realising that in the regions of the model where the process has noticeable effect an honest estimate of the measurement error is so large as to make any fitted value quantitatively useless. We can remove the instabilities of the system in this way, but we do not increase the amount of useful information when it matters (Appendix A). In summary, "priors" are being used to legitimate any additional terms which people want to put into their algorithms to fix the observable consequences of poor methodology.

If we could get people to understand that the concept of "degree of belief" is unscientific the approach above would not be so freely accepted. We might then be able to convince researchers in areas such as computer vision that it is necessary to justify where prior information comes from and to quantitatively confirm the approach using measured distributions. As this raises our level of critical thinking we might expect that in many cases this would identify flaws. I believe this would force us to have to redesign many published systems within a more justifiable (scientific) framework, which could ultimately lead to a theory of human perception. On the other hand, allowing this situation to continue due to an unwillingness to arrive at conclusions might have the research community going around in circles for another 50 years.

Here is a bullet point summary of some of the main issues to help provide focus;

[14] You have probably already spotted the obvious counter argument to this, academic publications can be counted as practical value under the RAE. Indeed such publications will contain material of academic interest, but this isn't what I mean by practical.

[15] If there is a forced choice to be made here I would suggest that it is between *scientist* and *mathematician*, as I would hope only the latter might consider subjective probability to have merit.

[16] Bayes Theorem is not too specific regarding where the prior terms should come from. The one thing you can say is that the $P(theory)$ in $P(theory|data) \propto P(data|theory)P(theory)$ should not be a function of the data.

- Frequentist probability can achieve far more than those who support subjective probability would want to admit. It has solutions for predictive statements and the use of prior knowledge.
- Making explicit all of the assumptions used in the calculation of probability, via the use of conditional notation is needed in order to avoid semantic problems.
- Bayes Theorem is derived for frequentist probability, it can and should be used in the same way, with quantitative testing of resulting theories.
- Bayes Theorem cannot be derived for subjective probability, and therefore can not be used to invoke the inevitable presence of arbitrary priors in the analysis of data.
- Subjective probability is unscientific, as it precludes objective testing it cannot be considered as a valid part of any theory.
- The use of arbitrary priors over continuous parameters in the construction of algorithms has no validity and any quantitative examples of utility must be viewed as a corruption of Likelihood.
- Conditional notation is insufficient to prevent the problems introduced into mathematical expressions by the use of subjective probability.
- *Mathematical consistency* is not a sufficient condition to ensure that subjective probability is *logically consistent.*
- Subjective definitions of probability are neither needed nor useful in explaining the subjective nature of brain function.

If Bayes Theorem with subjective priors is even to be considered valid, it must be derivable from the axioms and the subjective definition of probability, and thus without reference to frequency. Cursory consideration shows this to be impossible. In addition, the limitations of conditional notation associated with causal data are impossible to identify during general use of the theorem once the link to samples of data has been broken. You can not confirm the theoretical validity of an approach by observing that "It seems to work." when there are already theoretical arguments which undermine it.

It is also worth making some general comments regarding the attitude to these sort of issues found in different disciplines;

- Scientists largely ignore the mathematical problem of convergence, because not to do so would halt modern science in its tracks.
- Mathematicians largely ignore the scientific requirements of probability, as they find the possibility of greater freedom interesting.
- Engineers often switch between definitions of probability without knowing it, and are largely unaware that subjective probability can only ever be employed to describe (rather than scientifically develop) the consequently arbitrary choices and structures of resulting algorithms.

Under the scientific method you cannot *logically* prove a theory using data but only refute it, and a broken theory needs no testing. One might expect to apply the same logic to the problem of convergence for the frequentist case, but to me this isn't a broken theory so much as a semantic muddle.

Although you can never logically prove a theory, you can *statistically* corroborate a theory, and quantitative frequentist theories have been shown to match real world data in physics experiments to high degrees of precision, this is a level of testing which far outstrips "It seems to work". The key question which remains is; are there any practical circumstances in which we can apply a subjective definition of probability and a frequentist definition would not be more appropriate and useful?

Finally, I have not fully discussed here the problems of defining probabilities over continuous variables, rather than discrete states. This issue is covered at length in other documents on these web pages in the context of frequentist interpretations of Likelihood, Entropy and Bayes Theorem [19].

A Personal Note to the Reader

I would like to say that there is some freedom for advocates of Bayesian statistics to yet show the ideas have merit, but in all honesty I can't. The practical testing of algorithms which many offer up as evidence for the value of subjective probability is a quantitative and therefore frequentist task. This amounts to changing the definition of probability in mid-stream. Using a strict Bayesian interpretation to include subjective prior knowledge is immediately at odds with the process of quantitative estimation. The continued enthusiasm for subjective techniques is not unsurprising when we consider the fact that many people are still quite satisfied with Kolmogorov's axioms as a definition of probability. In my opinion, the large number of publications in the literature which apply these techniques should not be seen as contradiction of this conclusion. Instead they can be regarded as the consequence of accepting an inability to test without question, as it generates what can be recognised as an academic's 'gold mine', or a licence to print.

You should therefore be warned that a frequentist approach not only suggests how analysis of data should be done, but also restricts the methods which would otherwise appear to be available. If your main criteria at this stage in your career is publishing a lot of papers, taking my view of the subject is likely to slow you down. If you have understood this document, you need to decide; do you want only to publish as much as possible, or to restrict yourself to what you can be sure are scientific problems? I arrived at my own conclusions on this long ago, and believe that scientists won't even accept that there is a choice. Engineers might take the view that they are prepared to adopt a method simply because "It seems to work.". They should then be aware that anything they publish could ultimately be weakened when (as for fuzzy logic) they cannot honestly claim a valid theoretical argument for why they should have taken the approach. This gets us to the last respectable alternative, show how the reasoning in this document is flawed so that you can ignore its conclusions. In this case please drop me an email explaining where I went wrong.

Appendix A

The use of priors in estimation tasks is now widespread across many disciplines. The standard description of Bayesian estimation is the construction of[17]

$$P(\theta|x) \;=\; \frac{P(x|\theta)P(\theta)}{\int P(x|\theta)P(\theta)d\theta}$$

The best value of θ is that which maximises this expression, given the Likelihood distribution $P(x|\theta)$ for measurement x and prior distribution $P(\theta)$.

Bayesian techniques are popular and have many enthusiatic proponents. For example, in the preface of [3] the claim is made that by inlcuding a prior in the estimation of an extraction fraction (to asseses heart efficiency) a noisy measurement can be improved. When challenged on the theoretical basis (and yes there are challenges), most practitioners will resort to "it seems to work" as a justification. The key word here is "seems", and validity of the statement depends on how deeply we are prepared to test this claim. Anyone considering using "priors" to improve their analysis methods should take a some time to understand the consequences. It will be shown below that with a little work it is possible to justify the opposite statement "it doesn't seem to work". With some systems we can considerably simplify the process of incorporating Bayes priors into analysis in order to do this. For Gaussian distributions, the resulting distribution is also a Gaussian and its peak is given by a weighted average of the data and prior mean.

Imagine we have a weighing machine with (Gaussian) accuracy σ, and we use it to measure the weights w_i (our x above) of a group of people with a biological Gaussian distribution (our $P(\theta)$ above) of mean w_0 and standard deviation also σ. Without using Bayes priors to adjust our estimates, the expected distribution of measurements has an S.D. of $\sqrt{\sigma^2 + \sigma^2} = \sqrt{2}\sigma$.

If we apply the Bayes prior then each measurement is effectively combined with another 'prior' measurement of equal accuracy. The combination process will yield the average $w_i' = (w_i + w_0)/2$ for each measurement. Each Bayes estimate will therefore have an apparent reproducibility of $\sigma/2$ while each measurement will also be biased by a factor $(w_i - w_0)/2$ so than the apparent biological variation (for exact measurements) will be $\sigma/2$. The combined apparent distribution across the entire sample with therefore now be $\sqrt{\sigma^2/4 + \sigma^2/4} = \sigma/\sqrt{2}$.

If we are looking at the overall distribution of all sampled data it looks as if we have halved the variance of our system ($\sqrt{2}\sigma \rightarrow \sigma/\sqrt{2}$). We may be very happy with the result, because our measurements are much better

[17] If this were summarising my own work I would normaly try to explain which if these terms are probabilites and which densities, but I am just summarising here what others do.

behaved [18], This is about as far as most people go with their assessment, "it seems to work". But paradoxically, we have also achieved a distribution which, including noise, is narrower than the known biological variation (σ)! This should worry us, we appear to be removing the very thing we seek to measure.

Have we usefully increased the accuracy of our measurement of w_i by replacing it with $(w_i + w_0)/2$? Imagine that we wish to make a measurement for someone who is 3 S.D. heavier than the average, and wish to say something about the probability that the subject is heavier than the average. (This is analogous to detecting pathology in tissue labelling tasks). If we were to take multiple measurements, on average the subject will have a measured weight of $(w_i + w_0)/2 = w_0 + 3\sigma/2$, this is a change of $3\sigma/2$ from the mean with an accuracy of $\sigma/2$ i.e. a 3.D. effect. **This is precicely the same statistical significance we had before using the prior.**

This example is not a specific choice of numbers which provides this behaviour, you can try it for yourself with different starting distributions and values and will see that this is always the case. It is also true for any comparisons between measurements which use the same prior. If we take account of all the changes made to the measurement distribution in our assessment (as we would need to in a scientific summary), any improvement in stability is directly offset by the reduction in measured signal variation. The sensitivity to detect change is not improved by the use of priors. Logically we could argue that it would be a very strange result if it did, as any analysis could be then improved by combining with an arbitrary prior. The statistical equivalent of a free lunch.

The use of a prior has mapped our initial variable onto an equivalent one with reduced variation. On average a measurement of $w_0 + 3\sigma$ will be observed to have a biased value of $w_0 + 3\sigma/2$. This bias may be difficult to appreciate. On the face of it, given an appropriate prior mean, an average of all measurements will have no bias. However, repeated measurements have a systematic bias towards this mean.

The presence of bias will cause problems. Firstly, it will hinder any attempt to combine information from multiple sources. Secondly, it prevents an observation being interpreted as an absolute value. In addition, if the measurement accuracy of the incoming data varies, while the prior distribution is fixed, then the amount of bias will also vary. For engineering applications this may not worry us, and many will point out that we may now have a solution to an ill-posed inverse problem which could not be solved before. But this property has detrimental consequences in quantitative and scientific applications, where **getting a unique estimate is less important than honestly summarising the information content of data**.

You may feel that the weighing machine example is too simple to form a general conlusion, but you only need one example of something not working to disprove the theory (Popper). Strong Bayesians may argue that we have no right to perform quantitative tests, as that requires the frequenstist axiom (i.e. assumed probabilities match real world samples). However, if you are thinking of using this method yourself, you can always construct your own problem and work it through with some numbers.

In conclusion; the standard claim that the noise distribution on measurements is reduced by use of priors, which gives rise to opinions that "it seems to work", does not take account of the accompanying bias (or systematic error). Use of priors in this way converts some of the statistical measurement error (which is directly observable and relatively easy to deal with) into a systematic error (which is not observable and difficult to deal with). In any quantitative application (such as science or medicine), which requires use of; hypothesis tests, parameter covariances, or simply an absolute measurement, "it doesn't seem to work". We are not claiming here that all uses of Bayes' theorem are flawed, only that it is the responsibility of the researcher to see to it that it does something meaningful.

Acknowledgements

Thanks go as always to Paul Bromiley, who just happens to have recently decided to devote himself to reading the original books of Karl Popper, and to Jamie Gilmour and Bill Crum who sent me comments on early drafts of this document. Apologies go to Prof. Roland Wilson, who became an unwilling test subject for some of the arguments presented here, over a curry during BMVC07 at the University of Warwick.

References

[1] Ashby. F. G. and Perrin. N. A., Towards a Unified Theory of Similarity and Recognition. Psychological Review , 95, 1 ,124-150, 1988.

[2] R.J. Barlow. Statistics: A Guide to the use of Statistical Methods in the Physical Sciences. John Wiley and Sons, U.K., 1989.

[18] I use this example because I was once told by a peer at a conference that, as "it seems to work", a scale manufacturer could improve his sales figures by using this method, and therefore he (and anyone else) would need to be stupid not to do so.

[3] H.H.Barrett and K.LMyers, Foundations of Image Science, John Wiley and Sons, Editor. B.E.A Saleh, (see preface xxvii) 2004.

[4] P.A. Bromiley, N.A. Thacker, M.L.J. Scott, M. Pokrić,A.J. Lacey, and T.F. Cootes, "Bayesian and Non-Bayesian Probabilistic Models for Medical Image Analysis", *Image and Vision Computing*, 21/10 pp. 851-864, 2003.

[5] P. Courtney and N.A. Thacker, Performance Characterisation in Computer Vision: The Role of Statistics in Testing and Design, Imaging and Vision Systems: Theory, Assessment and Applications, Jacques Blanc-Talon and Dan Popescu (Eds.), NOVA Science Books, 2001.

[6] A.P. Dawid, Probability Forecasting. Encyclopedia of Statistical Science 7, pp 210-218. Wiley, 1986.

[7] R.A. Fisher, Theory of statistical estimation. *Proc. Cambridge Philosophical Society*, 2:700–725, 1925.

[8] K.Fukenaga, Introduction to Statistical Pattern Recognition, 2ed. Academic Press, San Diego, 1990.

[9] M. Ginsberg, Essentials of Artificial Intelligence, Morgan Kaufman Pub., U.S.A., 1993.

[10] R.M. Haralick, Performance Characterization in Computer Vision, CVGIP-IE 60, pp.245-249, 1994.

[11] K. Popper, The Logic of Scientific Discovery, English Trans, Routledge Classics, 1959.

[12] I. Poole, Optimal Probabilistic Relaxation Labelling. Proc. BMVC 1990, BMVA, 1990.

[13] S. Russell and P. Norvig, Artificial Intelligence, A Modern Approach (secong edition), Prentice Hall, 2003.

[14] A. Stuart, K.Ord and S.Arnold,. Kendall's Advanced Theories of Statistics. Volume, 2A, Classical Inference and the Linear Model, Oxford University Press, 1999.

[15] N.A.Thacker and J.E.W.Mayhew, 'Designing a Network for Context Sensitive Pattern Classification.' Neural Networks 3,3, 291-299, 1990.

[16] N.A.Thacker, I.A.Abraham and P.Courtney, 'Supervised Learning Extensions to the CLAM Network.' Neural Networks Journal, 10, 2, pp.315-326, 1997.

[17] N. A. Thacker, A.J.Lacey and P.Courtney, What is Intelligence?: Generalized Serial Problem Solving. Tina memo, 2001-006, www.tina-vision.net, 2001.

[18] N.A.Thacker, A.J.Lacey, P.Courtney and G. S. Rees, An Empirical Design Methodology for the Construction of Machine Vision Systems. Tina memo, 2002-005, www.tina-vision.net, 2002.

[19] N.A.Thacker, P.Bromiley, The Equal Variance Domain: Issues Surrounding the use of Probability Densities for Algorithm Construction. Tina memo, 2004-005, 2004.

[20] N.A.Thacker., Parameter Estimation for EM Mixture Modelling and its Relationship to Likelihood and EML. Tina Memo 2004-006.

[21] N.A.Thacker., Avoiding Zero and Infinity in Sample Based Algorithms, Tina Memo 2009-008.

[22] E.A. Vokurka., A. Herwadkar, N.A. Thacker, R.T. Ramsden and A. Jackson, Using Bayesian Tissue Classification to Improve the Accuracy of Vestibular Schwannoma Volume and Growth Measurement. AJNR, 23, 459-467, 2002.

Comments from Jamie Gilmour

Is there really an issue about German taxi drivers or rain tomorrow? To me, it becomes clearer if one reframes the question. E.g. What is the probability that it rained in Knoxville yesterday.

Clearly, "the" answer is 1 or 0.

Of my knowledge, I am uncertain unless I look it up. I think it is meaningful for me to say that the probability is about 80%. Of course, what I mean is that yesterday was a summer day in Knoxville and in my experience it rains at least once in about 8 out of 10 summer days. This can be made more scientific, of course, but the principle is the same.

Statements about the probability of one-off events are answered in the form "given what I know about the circumstances, I'd expect conditions like these to result in events of the specified sort about x % of the time"

Did a specified German born in 1960 die between the ages of 40 and 41? I don't know, but given knowledge of the German population I'd assign the idea low probability.

Given that he's also a taxi driver, I'd revise my estimate.

Given that he's buried in Cologne, I'd revise it upward somewhat.

And so on.

The notion that probability is a property only of the event and not of the event plus what we know about such circumstances seems nonsensical to me.

I agree, I think the section on frequentist probability now reflects these views. (Neil)

Comments from Paul Bromiley

Point 1:

Bayes theorem can be derived from frequentist arguments, and this may appear to provide it with some theoretical legitimacy. The numerator takes the form

$$p(data|theory)p(theory)$$

However, taking the next step to introduce subjective probability introduces two errors. First, the above expression results from a perfectly valid manipulation of the notation as long as we adhere strictly to that notation. In particular, the prior term in the above expression is the probability of the theory, conditional on nothing i.e. independent of any collective. The only ways to define a prior independent of a collective would appear to be to either base it on no information at all (as stated by Popper) i.e. use the equally likely definition, in which case expressions derived from Bayes Theorem revert to manipulations of likelihoods, or to base it on all possible information that has, or could ever, be obtained. In this case, the prior can only take the values 0 (the theory is wrong) or 1 (the theory is correct). This conclusion is valid regardless of which definition of probability you choose. Any concept of actually measuring a prior, either subjectively or objectively based on previous data, results in a prior which is itself a conditional on the collective used to define it. Therefore, we obtain the expression:

$$p(data|theory)p(theory|previousdata) \;=\; p(data|theory)p(previousdata|theory) \times 1$$

which is, of course, simply the use of likelihood to combine several sets of experimental data.

I have already mentioned in the text that the use of subjective priors cannot be encompassed by conditional notation. Your argument goes much further, it shows that strict adherence to the notation of conditional probability results in the frequentist interpretation of Bayes Theorem, and Likelihood as the appropriate way to compute $P(data|theory)$. I have included my own version of this argument in the document.(Neil)

Second, for a theory of probability to be considered valid in a mathematical sense, it must be derivable from Von Mises' axioms and the definition of probability we wish to adopt. Many of the familiar proofs (Poisson, Binomial and Gaussian distributions, Central Limit Theorem etc.) incorporate the use of the number of events. This is valid if we define probability in the frequentist sense i.e. as the limiting frequency in a number of events. However, we cannot switch definitions half-way through the derivation of our theory of probability i.e. we cannot start from the frequentist definition, obtain Bayes Theorem, and then switch to a subjective probability definition for the prior term. If Bayes Theorem with subjective priors is to be considered valid, it must be derivable from the axioms and the subjective definition of probability, and thus without reference to frequency i.e. all steps must be N-free (make

no reference to the total number of events). Cursory consideration shows this to be impossible. In fact, as stated by Popper, once the subjective definition has been adopted we can make no progress beyond the original axioms, as there is by definition no objective definition that we can manipulate.

Thanks for that Paul; This accords with my observation that there is no subjective derivation of Bayes Theorem, and that subjective probabilities cannot be related. I have put part of this comment into the conclusions. (Neil)

Point 2:

The conventional, although somewhat idealistic, view of progress in science is that it takes the form of an iteration of theory and experimental test. Failures in current theories are exposed through inabilities to explain experimental data, leading to the derivation of more advanced theories. The process must be based on logic and guided by general principles that have survived extensive experimental testing, such as conservation of energy, in order to avoid a random walk around theory space. The new generation of theory is then tested in turn and either corroborated or refuted. Theories must therefore be testable i.e. take the form of hypotheses rather than assertions. Any departure from this model will, in general, reduce the rate of progress or even reverse it.

The use of subjective probability fails to fit within the above model, on two counts. Most importantly, the use of arbitrary priors over parameters of theories, which are then optimised to fit the model to the data, reduces our ability to identify failures in the fit of experimental data to models, and thus interrupts the iterative process. The failures in the theory are accommodated by the priors, but we have no objective method with which to analyse these priors as they are, by definition, considered to be subjective. Furthermore, the ability to corroborate or refute a theory relies on our ability to calculate the probability that the data agrees with the theory, *and then compare it to another probability*, common choices being 95%, 99% or five-nines in more rigorous experimental disciplines. The definitions of these limits are frequentist, since they refer to the probability that the match between data and model could have resulted from random noise. We therefore have no right to compare subjective probabilities to these limits or any others, since that would be to change the definition of probability within an expression, and we therefore cannot use subjective probability to corroborate or refute a theory.

A clear and simple argument which goes to the core of the issue. Those seeking to use something other than a frequentist definition of probability must also develop a new approach to science if they wish to use it. It implies, for example, that any theory of brain function based upon subjective probability cannot be a scientific theory. (Neil)

Point 3:

The scenario in which prior probabilities are genuinely arbitrary, for example equally likely, could be called "strong" subjectivism. However, an alternative definition in which priors are estimated using frequentist methods, albeit from an arbitrary collective, could also be suggested. I will refer to this as "weak" subjectivism. Such an approach might appear to be more satisfactory as it accords more closely with the frequentist derivation of Bayes Theorem. However, weak subjectivity also has significant failings.

As you have said, we can not take a quantitative (frequentist) interpretation when we are considering distributions over continuous variables. Such methods are popular in medical data analysis and they can appear to improve the performance of an algorithm. However, this is done at the expense of generalisation capability. Any estimate of the parameters will not be consistent with the underlying generator of the data, it will be biased. Performance evaluation applied to algorithms will also tend to underestimate errors, thereby giving a misleading indication of apparent performance. As we have taken some random error and turned it into a systematic error. This is deeply unsatisfactory from the point of view of scientific logic, as performance evaluation should always be aimed at proving that an algorithm does not work, instead of proving that it does, in order to avoid accepting invalid algorithms as valid.

Taking our priors from sample distributions over discrete events is also problematic. Consider a simple example, weighting a Bayesian categorisation according to the demographic of a population. If we can only observe the population and have no independent reason to believe that its statistical characteristics should remain fixed, then we can not know that any probabilities we compute are correct. This is the problem of non-stationarity. In addition, if we attempt to solve this problems by tracking the time variation of the priors, then as you have said, the results are not directly comparable, as they are in effect using different definitions of probability.

However contrary this might appear to common logic, I think it is particularly important that we appreciate how the use of sample data to define prior distributions is not automatically consistent with the quantitative use of probability. (Neil)

Comments from Bill Crum

1. I'm not sure I agree with the argument about P(theory) being non-physical as for real experiments there is only one correct theory. For instance if I have a set of n brains and laboriously measure grey and white matter tissue distributions through some manual means then I can fit a Gaussian model to GM (say) for each case which in generally will have parameters θ_i for the the i'th brain. These parameters will form a distribution across the set of n brains. Then if want to use this information as a prior for the segmentation of GM in the (n+1)th brain, don't I then have a (frequentist) prior which gives me P(theory). Maybe this is the point - that I have to define my prior in such a frequentist way for it to be valid?

For any one of the brains you are analysing there is only one correct value for θ_i. You can choose to look at a distribution across different brains if you wish but why do you believe that this distribution tells you anything about a specific case? As Paul says in his comments, using a sample to define a prior produces not a pure prior but $P(theory|cohort)$. This is generally going to be arbitrary (unscientific) unless you stick with Poppers restrictions on the use of conditional notation.(Neil)

2. I don't understand the point about "applying arbitrary non-linear transformations to our parameter definitions without changing the theory which will change any sample distribution we measure". Can you elaborate or point me to something which explains this statement a bit more?

The choice of how to define our measurements of the world, or the parameters we put in a theory (the domains) are never unique. Suppose for example we choose to model the size of a circle as a radius (r), another scientist might say that he thinks you should use area ($a = \pi r^2$). For the specific case of prior distributions over parameters (which is the section you refer to), if we sample our distributions and compute the prior value directly from the estimated probability density (rather than integrating over an interval), this value will change if we apply a non-linear transformation, like $r \rightarrow a$. Just because these choices are not immediately obvious in specific applications (such as brain modelling and your θ_i), doesn't mean they are not there. Such a prior changes at the whim of the researcher and therefore this gives us a second reason why the methodology can not be said to have any intrinsic validity. A scientifically valid methodology would give the same result regardless of these choices. (Neil)

3. The issue about maximising probability densities as opposed to integrating over intervals is interesting. I wondered if people "get away" with it in practice because if you represent your data as a histogram then a single bin of width 1 is the smallest interval you can define. Then if you integrate the discrete density over a single bin you get back numerically the probability density. Do I need another coffee or is this an example of where discrete versus continuous representations can trip you up?

Yes, this is a situation where the discrete case would not be ambiguous. Why should an interval of 1 be considered appropriate, and how would we know even if it was? People aren't really getting away with it if that means they are unaware of the issue and only use an appropriate domain by accident. (Neil)

4. Actually this report has made me want to go back and read some of your other memos for clarification. Is there a more concrete standard example from the "contradictory methods used for the comparison of probability density distribution found in the area of statistical pattern recognition" which could be invoked in several places in the text to show where "standard" approaches get it wrong.

I am reluctant to make this document any longer than it already is by including examples. However, memo 2004-005 already describes alternative probability similarity measures (Kullback Liebler Divergence (KL), Matusita Measure... etc.). In the conventional presentation none of these have associated domains of applicability, ie: you are free to choose whichever you like as the basis for an algorithm. This situation is tolerated due to the vagueness surrounding probability. (Neil)