

Expanded Notes on Probability

Excerpts from Chapter 3 - Boccio - Quantum Mechanics

1. Probability Concepts

Quantum mechanics will necessarily involve probability in order for us to make the connection with experimental measurements.

We will be interested in understanding the quantity

$$P(A|B) = \text{probability of event A given that event B is true}$$

In essence, event B sets up the conditions or an environment and then we ask about the (conditional) probability of event A given that those conditions exist. The $|$ symbol means "given" so that items to the right of this "conditioning" symbol are taken as being true.

In other words, we set up an experimental apparatus, which is expressed by properties B and do a measurement with that apparatus, which is expressed by properties A. We generate numbers (measurements) which we use to give a value to the quantity $P(A|B)$.

(a) Standard Thinking

We start with the **standard** mathematical formalism based on axioms. We define these events

$A = \text{occurrence of } A$

(denotes that proposition A is true)

$\sim A = \text{NOT } A = \text{nonoccurrence of } A$

(denotes that proposition A is false)

$A \& B = A \text{ AND } B = \text{occurrence of both } A \text{ and } B$

(denotes proposition A and B is true)

$A \vee B = A \text{ OR } B = \text{occurrence of at least one of the events } A \text{ and } B$

(denotes proposition A or B is true)

and standard Boolean logic as shown below:

Boolean logic uses the basic statements AND, OR, and NOT. Using these and a series of Boolean expressions, the final output would be one TRUE or FALSE statement.

This is illustrated below:

If A is true AND B is true, then $(A \text{ AND } B)$ is true

If A is true AND B is false, then $(A \text{ AND } B)$ is false

If A is true OR B is false, then $(A \text{ OR } B)$ is true

If A is false OR B is false, then $(A \text{ OR } B)$ is false

or written as a "truth" table:

A	B	$(A \wedge B)$	$(A \vee B)$
1	1	1	1
1	0	0	1
0	1	0	1
0	0	0	0

where 1 = *TRUE*, 0 = *FALSE*.

We then set up a theory of probability with these axioms:

$$(1) P(A|A) = 1$$

This is the probability of the occurrence A given the occurrence of A. This represents a **certainty** and, thus, the probability must = 1. This is clearly an obvious assumption that we must make if our probability ideas are to make any sense at all.

In other words, if I set the experimental apparatus such that the meter reads A, then it reads A with probability = 1.

$$(2) 0 \leq P(A|B) \leq P(B|B) = 1$$

This just expresses the sensible idea that no probability is greater than the probability of a certainty and it make no sense to have the probability be less than 0.

$$(3) P(A|B) + P(\sim A|B) = 1 \text{ or } P(\sim A|B) = 1 - P(A|B)$$

This just expresses the fact that the probability of something (anything) happening (A or $\sim A$) given B is a certainty (=1), that is, since the set A or $\sim A$ includes everything that can happen, the total probability that one or the other occurs must be the probability of a certainty and be equal to one.

$$(4) P(A \& B|C) = P(A|C)P(B|A \& C)$$

This says that the probability that 2 events A, B both occur given that C occurs equals the probability of A given C multiplied by the probability of B given (A&C), which makes sense if you think of them happening **in sequence**.

All other probability relationships can be derived from these axioms.

The nonoccurrence of A given that A occurs must have probability = 0. This is expressed by

$$P(\sim A|A) = 0$$

This result clearly follows from the axioms since

$$P(A|B) + P(\sim A|B) = 1$$

$$P(A|A) + P(\sim A|A) = 1$$

$$P(\sim A|A) = 1 - P(A|A) = 1 - 1 = 0$$

Example: Let us evaluate $P(X \& Y | C) + P(X \& \sim Y | C)$.

We use axiom (4) in the 1st term with $A = X, B = Y$ and $C = C$ and in the 2nd term with $A = X, B = \sim Y$ and $C = C$ to get

$$\begin{aligned} P(X \& Y | C) + P(X \& \sim Y | C) &= P(X | C)P(Y | X \& C) + P(X | C)P(\sim Y | X \& C) \\ &= P(X | C)[P(Y | X \& C) + P(\sim Y | X \& C)] = P(X | C)[1] \quad \text{using axiom (3)} \end{aligned}$$

and finally

$$P(X \& Y | C) + P(X \& \sim Y | C) = P(X | C)$$

which says probability that X is true regardless of whether Y is true, is the sum of the probabilities of X and Y for all possibilities associated with Y (Y and $\sim Y$ in this case).

Now let us use this result with $X = \sim A, Y = \sim B$. This gives

$$P(\sim A \& \sim B | C) = P(\sim A | C) - P(\sim A \& B | C) = 1 - P(A | C) - P(\sim A \& B | C)$$

Expanding the last term using $X = B, Y = A$ we then have

$$P(B \& \sim A | C) + P(B \& A | C) = P(B | C)$$

or

$$P(\sim A \& B | C) = P(B | C) - P(B \& A | C)$$

which gives

$$P(\sim A \& \sim B | C) = 1 - P(A | C) - P(B | C) + P(A \& B | C)$$

Now

$$P(A \vee B) = 1 - P(\sim(A \vee B) | C) = 1 - P(\sim A \& \sim B | C)$$

and since

$$(\sim(A \vee B)) = (\sim A \& \sim B)$$

i.e., we can construct a 'truth table' as shown below, which illustrates the equality directly

A	B	$(\sim(A \vee B))$	$(\sim A \& \sim B)$	
1	1	0	0	
1	0	0	0	(this is the "truth table")
0	1	0	0	
0	0	1	1	

we finally get

$$P(A \vee B) = P(A | C) + P(B | C) - P(A \& B | C)$$

This is a very important and useful result.

If we have $P(A \& B | C) = 0$, then events A and B are said to be **mutually exclusive** given that C is true and the relation then reduces to

$$P(A \vee B) = P(A | C) + P(B | C) \tag{3.1}$$

This is the rule of **addition of probabilities for exclusive events**.

Some other important results are:

$$\text{If } A \& B = B \& A \text{ , then } P(A|C)P(B|A \& C) = P(B|C)P(A|B \& C) \quad (3.2)$$

$$\text{If } P(A|C) \neq 0 \text{ , then } P(B|A \& C) = P(A|B \& C) \frac{P(B|C)}{P(A|C)} \quad (3.3)$$

which is **Baye's theorem**. It relates the probability of B given A to the probability of A given B.

When we say that B is **independent** of A, we will mean

$$P(B|A \& C) = P(B|C) \quad (3.4)$$

or the occurrence of A has **NO influence** on the probability of B given C. Using axiom (4) we then have the result:

if A and B are independent given C,
then $P(A \& B|C) = P(A|C)P(B|C)$

This is called **statistical** or **stochastic** independence. The result generalizes to a set of events $\{A_i \text{ , } i=1,2,\dots,n\}$. All these events are independent if and only if

$$P(A_1 \& A_2 \& \dots \& A_m | C) = P(A_1|C)P(A_2|C)\dots\dots\dots P(A_m|C)$$

for all $m \leq n$.

Now let us think about these ideas in another way that has fundamental importance in modern approaches to quantum theory. The fundamental result in this view will turn out to be the Bayes formula and its relationship to measurements.

(b) Bayesian Thinking

Two Different Axioms

- (1) If we specify how much we believe something is true, then we must have implicitly specified how much we believe it is false.
- (2) If we first specify how much we believe that (proposition) Y is true, and then state how much we believe X is true given that Y is true, then we must implicitly have specified how much we believe that both X and Y are true

We assign **real** numbers to each proposition in a manner so that the larger the numerical value associated with a proposition, the more we believe it.

Only using the rules of **Boolean logic**, ordinary algebra, and the constraint that if there are several different ways of using the same information, then we should always arrive at the same conclusions independent of the particular analysis-path chosen, it is then found that this consistency could only be guaranteed if the real numbers we had attached to our beliefs in the various propositions could be **mapped** (or transformed) to another set of **real** positive numbers which obeyed the usual rules of probability theory:

$$\text{prob}(X|I) + \text{prob}(\sim X|I) = 1 \quad (\text{same as axiom(3)}) \quad (3.5)$$

$$\text{prob}(X \& Y|I) = \text{prob}(X|Y \& I) \times \text{prob}(Y|I) \quad (\text{same as axiom (4)}) \quad (3.6)$$

Equation (3.5) is called the **sum rule** and states (as earlier) that the probability that X is true plus the probability that X is false is equal to one.

Equation (3.6) is called the **product rule**. It states (as earlier) that the probability that both X and Y are true is equal to the probability that X is true given that Y is true times the probability that Y is true (independent of X).

Note that all the probabilities are conditional on proposition(s) or conditioning(s) I , which denotes the relevant background information on hand. It is important to understand that there is no such thing as an absolute probability (without prior information).

Bayes' Theorem and Marginalization

As before, we can use the sum and product rules to derive other results.

First, starting with the product rule we have

$$\text{prob}(X \& Y|I) = \text{prob}(X|Y \& I) \times \text{prob}(Y|I)$$

We can rewrite this equation with X and Y interchanged

$$\text{prob}(Y \& X|I) = \text{prob}(Y|X \& I) \times \text{prob}(X|I)$$

Since the probability that both X and Y are true must be logically the same as the probability that both Y and X are true we must also have

$$\text{prob}(Y \& X|I) = \text{prob}(X \& Y|I)$$

or

$$\text{prob}(X|Y \& I) \times \text{prob}(Y|I) = \text{prob}(Y|X \& I) \times \text{prob}(X|I)$$

or

$$\text{prob}(X|Y \& I) = \frac{\text{prob}(Y|X \& I) \times \text{prob}(X|I)}{\text{prob}(Y|I)} \quad (3.7)$$

which is **Bayes theorem** (as earlier).

Most standard treatments of probability do not attach much importance to Bayes' rule.

This rule, however, which relates $\text{prob}(X|Y \& I)$ to $\text{prob}(Y|X \& I)$, allows us to turn things around with respect to the conditioning symbol, which leads to a reorientation of our thinking about probability.

The fundamental importance of this property to data analysis becomes apparent if we replace X and Y by **hypothesis** and **data**:

$$\text{prob}(X|Y \& I) \propto \text{prob}(Y|X \& I) \times \text{prob}(X|I)$$

$$\text{prob}(\text{hypothesis} | \text{data} \& I) \propto \text{prob}(\text{data} | \text{hypothesis} \& I) \times \text{prob}(\text{hypothesis} | I)$$

Note that the equality in equation (3.7) has been replaced with a proportionality because the term $prob(data|I) = \mathbf{evidence}$ has been omitted. The proportionality constant can be found from the normalization requirement that the sum of the probabilities for something happening must equal 1.

The power of Bayes' theorem lies in the fact that it relates the quantity of interest, the probability that the hypothesis is true given the data, to the term that we have a better chance of being able to assign, the probability that we would have obtained the measured data if the hypothesis was true.

The various terms in Bayes' theorem have formal names. The term $prob(hypothesis|I) = \mathbf{prior}$ probability represents our state of knowledge (or ignorance) about the truth of the hypothesis before we have analyzed the current data. This is modified by the experimental measurements through the term $prob(data|hypothesis \& I) = \mathbf{likelihood}$ function. This product gives $prob(hypothesis|data \& I) = \mathbf{posterior}$ probability representing our state of knowledge about the truth of the hypothesis in the light of the data (after measurements).

In some sense, Bayes' theorem encapsulates **the process of learning**, as we shall see later.

Second, consider the following results from equation (3.6)

$$\begin{aligned} prob(X \& Y | I) &= prob(Y \& X | I) = prob(Y | X \& I) \times prob(X | I) \\ prob(X \& \sim Y | I) &= prob(\sim Y \& X | I) = prob(\sim Y | X \& I) \times prob(X | I) \end{aligned}$$

Adding these equations we get

$$prob(X \& Y | I) + prob(X \& \sim Y | I) = (prob(Y | X \& I) + prob(\sim Y | X \& I)) prob(X | I)$$

Since $prob(Y | X \& I) + prob(\sim Y | X \& I) = 1$ we have

$$prob(X \& Y | I) + prob(X \& \sim Y | I) = prob(X | I) \tag{3.8}$$

which, again, is the same result as earlier.

If, on the other hand, $Y \rightarrow \{Y_k; k=1,2,\dots,M\}$ representing a set of M alternative possibilities, then we generalize the two-state result above as

$$\sum_{k=1}^M prob(X \& Y_k | I) = prob(X | I) \tag{3.9}$$

We can derive this result exactly as we did in equation (3.8).

$$\begin{aligned} prob(X \& Y_1 | I) &= prob(Y_1 \& X | I) = prob(Y_1 | X \& I) \times prob(X | I) \\ prob(X \& Y_2 | I) &= prob(Y_2 \& X | I) = prob(Y_2 | X \& I) \times prob(X | I) \\ &\dots\dots\dots \\ prob(X \& Y_M | I) &= prob(Y_M \& X | I) = prob(Y_M | X \& I) \times prob(X | I) \end{aligned}$$

Adding these equations we get

$$\sum_{k=1}^M \text{prob}(X \& Y_k | I) = \text{prob}(X | I) \left(\sum_{k=1}^M \text{prob}(Y_k | X \& I) \right)$$

If we assume that the $\{Y_k\}$ form a mutually exclusive and exhaustive set of possibilities, that is, if one of the Y_k 's is true, then all the others must be false, we then get

$$\sum_{k=1}^M \text{prob}(Y_k | X \& I) = 1 \tag{3.10}$$

which is a normalization condition. Thus, we get equation (3.9).

If we go to the **continuum limit** where we consider an arbitrarily large number of propositions about some result (the range in which a given result might lie), then as long as we choose the intervals in a **contiguous** fashion, and cover a big enough range of values, we will have a mutually exclusive and exhaustive set of possibilities. In the limit of $M \rightarrow \infty$, we obtain

$$\text{prob}(X | I) = \int_{-\infty}^{\infty} \text{prob}(X \& Y | I) dY \tag{3.11}$$

which is the **marginalization** equation. The integrand here is technically a **probability density function** rather than a probability. It is defined by

$$\text{pdf}(X \& Y = y | I) = \lim_{\delta y \rightarrow 0} \frac{[\text{prob}(X \& y \leq Y \leq y + \delta y | I)]}{\delta y} \tag{3.12}$$

and the probability that the value of Y lies in a finite range between y_1 and y_2 (and X is also true) is given by

$$\text{prob}(X \& y_1 \leq Y \leq y_2 | I) = \int_{y_1}^{y_2} \text{pdf}(X \& Y | I) dY \tag{3.13}$$

and equation (3.11) then follows directly. In this continuum limit the normalization condition takes the form

$$1 = \int_{-\infty}^{\infty} \text{prob}(Y | X \& I) dY \tag{3.14}$$

Marginalization is a very powerful device in data analysis because it enables us to deal with **nuisance parameters**, that is, quantities which necessarily enter the analysis but are of no intrinsic interest. The unwanted background signal present in many experimental measurements, and instrumental parameters which are difficult to calibrate, are examples of nuisance parameters.

2. Probability Interpretations

(a) Standard thinking

In the standard way of thinking about probability in relation to experiments, measured results are related to probabilities using the concept of a **limit frequency**. The limit frequency is linked to

probability by this definition:

If C can lead to either A or $\sim A$, and if in n repetitions, A occurs m times, then

$$P(A|C) = \lim_{n \rightarrow \infty} \left(\frac{m}{n} \right) \quad (3.15)$$

We must now connect the mathematical formalism with this limit frequency concept so that we can use the formalism to make predictions for experiments in real physical systems.

This approach depends on whether we can prove that the limit makes sense for real physical systems. Let us see how we can understand the real meaning of the above interpretation of probability and thus learn how to use it in quantum mechanics, where probability will be the dominant property.

Suppose that we have an experimental measurement, M , that can yield either A or $\sim A$ as results, with a probability for result A given by

$$P(A|M) = p$$

In general, we let any sequence of n independent measurements be labelled as event M^n and we define n_A as the number of times A occurs, where $0 \leq n_A \leq n$.

Now imagine we carry out a sequence of n independent measurements and we find that A occurs r times. The probability for a sequence of results that includes result A r times and $\sim A$ $(n-r)$ times (independent of their order in the sequence) is given by

$$p^r q^{n-r}$$

where

$$q = P(\sim A|M) = 1 - P(A|M) = 1 - p$$

The different sequence orderings are mutually exclusive events and thus we have

$$P(n_A = r | M^n) = \sum_{\substack{\text{all possible} \\ \text{orderings}}} p^r q^{n-r} \quad (3.16)$$

The sum $\sum_{\substack{\text{all possible} \\ \text{orderings}}}$ just counts the number of ways to distribute r A 's and

$(n-r)$ $\sim A$'s, where all the terms contain the common factor $p^r q^{n-r}$. This result is given by the Binomial probability distribution as

$$\frac{n!}{r!(n-r)!}$$

so that

$$P(n_A = r | M^n) = \frac{n!}{r!(n-r)!} p^r q^{n-r} \quad (3.17)$$

Now to get to the heart of the problem. The frequency of A in M^n is given by

$$f_n = \frac{n_A}{n}$$

This is **not necessarily** $= p$ in any set of measurements.

What is the relationship between them? Consider the following:

$$\begin{aligned} \langle n_A \rangle &= \text{average or expectation value} \\ &= \text{sum over [possible values times probability of that value]} \\ &= \sum_{r=0}^n r P(n_A = r | M^n) = \sum_{r=0}^n r \frac{n!}{r!(n-r)!} p^r q^{n-r} \end{aligned} \quad (3.18)$$

We now use a clever mathematical trick to evaluate this sum. For the moment consider p and q to be two arbitrary independent variables. At the end of the calculation we will let $q=1-p$ as is appropriate for a real physical system.

From the Binomial expansion formula, we have, in general,

$$\sum_{r=0}^n \frac{n!}{r!(n-r)!} p^r q^{n-r} = (p+q)^n$$

We then have

$$\begin{aligned} p \frac{\partial}{\partial p} \sum_{r=0}^n \frac{n!}{r!(n-r)!} p^r q^{n-r} &= p \frac{\partial}{\partial p} (p+q)^n \\ \text{or } \sum_{r=0}^n r \frac{n!}{r!(n-r)!} p^r q^{n-r} &= np(p+q)^{n-1} \end{aligned}$$

or

$$\begin{aligned} \sum_{r=0}^n r P(n_A = r | M^n) &= np(p+q)^{n-1} \\ \text{or } \langle n_A \rangle &= np(p+q)^{n-1} \end{aligned}$$

In a real physical system, we must have $p+q=1$, so that we end up with the result

$$\langle n_A \rangle = np \quad (3.19)$$

and

$$\langle f_n \rangle = \frac{\langle n_A \rangle}{n} = p \quad (3.20)$$

This says that p = the average frequency.

This **does not** say, however, that f_n is actually close to p .

Now consider a more general experiment where the outcome of a measurement is the value of some continuous variable Q , with probability density (for its continuous spectrum) given by

$$P(q < Q < q + dq | M) = h(q) dq$$

If we let $h(q)$ contain delta-functions, then this derivation is also valid for the discrete part of the spectrum.

We can now derive the following useful result. If Q is a nonnegative variable, which means that $h(q)=0$ for $q<0$, then for any $\varepsilon>0$

$$\langle Q \rangle = \int_0^{\infty} h(q)q dq \geq \int_{\varepsilon}^{\infty} h(q)q dq \geq \varepsilon \int_{\varepsilon}^{\infty} h(q) dq = \varepsilon P(Q \geq \varepsilon | M)$$

This implies that

$$P(Q \geq \varepsilon | M) \leq \frac{\langle Q \rangle}{\varepsilon} \quad (3.21)$$

Now we apply this result to the nonnegative variable $|Q-c|^\alpha$, where $\alpha>0$ and $c = \text{number}$, to obtain

$$P(|Q-c| \geq \varepsilon | M) = P(|Q-c|^\alpha \geq \varepsilon^\alpha | M) \leq \frac{\langle |Q-c|^\alpha \rangle}{\varepsilon^\alpha} \quad (3.22)$$

which is called **Chebyshev's inequality**.

In the special case where

$$\alpha = 2 \quad , \quad c = \langle Q \rangle = \text{mean of distribution}$$

$$\langle |Q-c|^2 \rangle = \langle |Q - \langle Q \rangle|^2 \rangle = \langle Q^2 \rangle - \langle Q \rangle^2 = \sigma^2 = \text{variance} \quad , \quad \varepsilon = k\sigma$$

we have

$$P(|Q - \langle Q \rangle| \geq k\sigma | M) \leq \frac{1}{k^2} \quad (3.23)$$

or, the probability of Q being k or more standard deviations from the mean is no greater than $\frac{1}{k^2}$ (**independent** of the form of the probability distribution).

Now we return to the n repetition experiment and choose

$$\alpha = 2 \quad , \quad Q = n_A = \sum_{i=1}^n \ell_i$$

$$\text{where } \ell_i = \begin{cases} 1 & \text{if outcome of } i^{\text{th}} \text{ repetition of } M \text{ is } A \\ 0 & \text{otherwise} \end{cases}$$

$$c = \langle Q \rangle = np$$

We then have

$$P(|n_A - np| \geq \varepsilon | M) \leq \frac{\langle (n_A - np)^2 \rangle}{\varepsilon^2}$$

Now

$$\begin{aligned} \langle (n_A - np)^2 \rangle &= \left\langle \left\{ \sum_{i=1}^n (\ell_i - p) \right\}^2 \right\rangle \\ &= \sum_i \sum_j \langle (\ell_i - p)(\ell_j - p) \rangle \end{aligned}$$

Each repetition is independent, which implies that

$$\langle (\ell_i - p)(\ell_j - p) \rangle = \langle \ell_i - p \rangle \langle \ell_j - p \rangle = 0 \text{ for } i \neq j$$

so that

$$\langle (n_A - np)^2 \rangle = \left\langle \left\{ \sum_{i=1}^n (\ell_i - p) \right\}^2 \right\rangle \leq n$$

Thus, we have

$$P(|n_A - np| \geq \varepsilon | M) \leq \frac{n}{\varepsilon^2} \quad (3.24)$$

For $f_n = \text{relative frequency of } A = \frac{n_A}{n}$ we then have

$$P(|f_n - p| \geq \frac{\varepsilon}{n} | M) \leq \frac{n}{\varepsilon^2}$$

If we let $\delta = \frac{\varepsilon}{n}$, then we have

$$P(|f_n - p| \geq \delta | M) \leq \frac{1}{n\delta^2} \quad (3.25)$$

This implies that the probability of f_n (the relative frequency of A in n independent repetitions of M) being more than ε away from p approaches 0 as $n \rightarrow \infty$.

This is an example of the **law of large numbers** in action.

This **DOES NOT** say $f_n = p$ at any time or that f_n remains close to p as $n \rightarrow \infty$.

It **DOES** say that the deviation of f_n from p becomes more and more improbable or that the probability of any deviation approaches 0 as $n \rightarrow \infty$.

It is in this sense that one uses the limit frequency from experiment to compare with theoretical probability predictions in quantum mechanics.

From probability theory one derives only statements of probability, not of necessity.

(b) Bayesian thinking

How do we reason in situations where it is not possible to argue with certainty? In other words, is there a way to use techniques of deductive logic to study the inference problem arising when using inductive logic? No matter what scientists say, this is what they are actually doing most of the time.

The answer to this last question resides in the Bayes' rule.

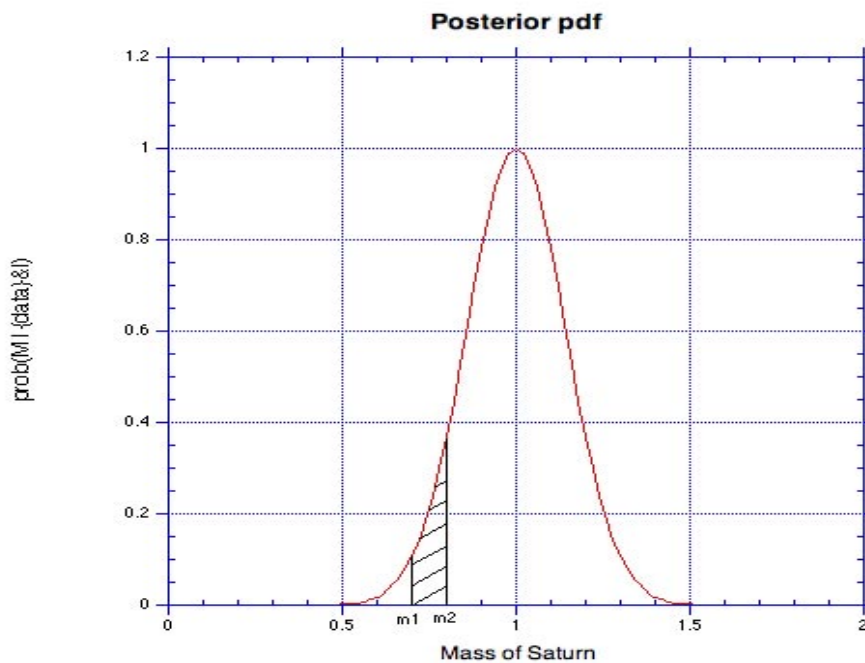
To Bayes (along with Bernoulli and Laplace), a probability represented a "degree-of-belief" or "plausibility", that is, how much one thinks that something is true, **based** on the **evidence on hand**.

The developers of probability (Fisher, Neyman and Pearson) thought

this seemed too vague and subjective a set of ideas to be the basis of a "rigorous" mathematical theory. Therefore, they **defined** probability as the long-run relative frequency with which an event occurred, given **infinitely many repeated** experimental trials. Since such probabilities can be measured, probability was then thought to be an objective tool for dealing with random phenomena.

This frequency definition certainly **seems** to be more objective, but it turns out that its **range of validity** is far more limited.

Let us imagine an experiment to measure the mass M of Saturn. Assume that we are given orbital data (astronomical **measurements**). We then compute the posterior **pdf** for the mass M given the data and all the relevant background information I (laws of classical mechanics, etc) $prob(M|\{data\} \& I)$. Suppose that we find the result shown below:



The shaded area under the posterior pdf curve between m_1 and m_2 is a measure of how much we believe that the mass of Saturn is in the range $m_1 \leq M \leq m_2$.

Clearly, the position of the maximum of the posterior pdf represents a best estimate of the mass; its width, or spread, about this optimal value gives an indication of the uncertainty in that estimate.

How would this data (as in the figure) be interpreted in terms of the frequency definition?

In order to think of the mass of Saturn as a **random variable**, we would have to imagine a large ensemble of universes in which everything remains constant apart from the mass of Saturn, which is a very strange procedure to say the least.

Alternatively, we could think of the data in the figure in terms of the distribution of measurements of the mass in many repetitions of the experiment.

Clearly, having to think of a frequency interpretation for every data analysis problem is rather perverse.

What do we mean by the "measurement of the mass" when the data consist of orbital periods?

Why should we have to think about many repetitions of an experiment that never happened?

What we need to do is make the best **inference** of the mass given the (few) data that we actually have, which is precisely the Bayes' view of probability.

In this view, probability represents a **state of knowledge**. The conditional probabilities represent **logical** connections rather than **causal** ones.

Example:

Consider an urn that contains 5 red balls and 7 green balls.

If a ball is selected at "random", then we would all agree that the probability of picking a red ball would be $5/12$ and of picking a green ball would be $7/12$.

If the ball is not returned to the urn, then it seems reasonable that the probability of picking a red or green ball must depend on the outcome of the first pick (because there will be one less red or green ball in the urn).

Now suppose that we are not told the outcome of the first pick, but are given the result of the second pick.

Does the probability of the first pick being red or green change with the knowledge of the second pick?

Initially, many observers would probably say "no", that is, at the time of the first draw, there were still 5 red balls and 7 green balls in the urn, so the probabilities for picking red and green should still be $5/12$ and $7/12$ independent of the outcome of the second pick.

The error in this argument becomes clear if we consider the extreme example of an urn containing only 1 red and 1 green ball.

Although, the second pick cannot affect the first pick in a physical sense, a knowledge of the second result does influence what we can infer about the outcome of the first pick, that is, if the second ball was green, then the first ball must have been red, and vice versa.

We can calculate the result as shown below:

$Y = \text{pick is GREEN}(2\text{nd pick})$

$X = \text{pick is RED}(1\text{st pick})$

$I = \text{initial number of RED/GREEN balls} = \{n, m\}$

A Bayesian would say :

$$\text{prob}(X | Y \& I) = \frac{\text{prob}(Y | X \& I) \times \text{prob}(X | I)}{\text{prob}(Y | I)}$$

$$\text{prob}(X | Y \& \{n, m\}) = \frac{\text{prob}(Y | X \& \{n, m\}) \times \frac{n}{n+m}}{\frac{n}{n+m} \frac{m}{n+m-1} + \frac{m}{n+m} \frac{m-1}{n+m-1}} = \frac{\frac{m}{(n+m-1)} \times n}{\frac{nm}{n+m-1} + \frac{m(m-1)}{n+m-1}} = \frac{n}{n+m-1}$$

$$n = m = 1 \Rightarrow \text{prob}(X | Y \& \{1, 1\}) = \frac{n}{(n+m-1)} = 1$$

$$n = 5, m = 7 \Rightarrow \text{prob}(X | Y \& \{5, 7\}) = \frac{5}{11} = 0.456$$

Non - Bayesian says :

$$\text{prob}(X | \{5, 7\}) = \frac{5}{12} = 0.417$$

Clearly, the Bayesian and Non-Bayesian disagree.

However, the Non-Bayesian is just **assuming** that the calculated result 0.417 is correct, whereas, the Bayesian is using the rules of probability (Bayes' Rule) to infer the result 0.456 **correctly**.

The concerns about the subjectivity of the Bayesian view of probability are understandable. I think that the presumed shortcomings of the Bayesian approach merely reflect a confusion between subjectivity and the difficult technical question of how probabilities (especially prior probabilities) should be assigned.

The popular argument is that if a probability represents a degree-of-belief, then it must be subjective, because my belief could be different from yours. The Bayesian view is that a probability does indeed represent how much we believe that something is true, but that this belief should be based on all the relevant information available (all prior probabilities).

While this makes the assignment of probabilities an open-ended question, because the information available to me may not be the same as that available to you, it is not the same as subjectivity. It simply means that probabilities are **always conditional**, and this conditioning must be stated **explicitly**.

Objectivity demands only that two people having the same information should assign the same probability.

Cox looked at the question of plausible reasoning from the perspective of logical consistency. He found that the only rules that worked were those of probability theory! Although the sum and product rules of probability are straightforward to prove for frequencies (using Venn diagrams), Cox showed that their range of validity goes much further. Rather than being restricted to frequencies, he showed

that probability theory constitutes the basic calculus for logical and consistent plausible reasoning, which means scientific inference!

Another Example - Is this a fair coin?

We consider a simple coin-tossing experiment.

Suppose that I had found this coin and we observed 4 heads in 11 flips.

If by the word "**fair**" we mean that we would be prepared to make a 50:50 bet on the outcome of a flip being a head or a tail, then do you think that it is a fair coin?

If we ascribe fairness to the coin, then we naturally ask how sure are we that this was so or if it was not fair, how unfair do we think it was?

A way of formulating this problem is to consider a large number of **contiguous** hypotheses about the range in which the **bias-weighting** of the coin might lie. If we denote bias-weighting by H , then $H=0$ and $H=1$ can represent a coin which produces a tail(not a head!) or a head on every flip, respectively. There is a continuum of possibilities for the value of H between these limits, with $H=1/2$ indicating a fair coin. The hypotheses might then be, for example

- (a) $0.00 \leq H \leq 0.01$
 - (b) $0.01 \leq H \leq 0.02$
 - (c) $0.02 \leq H \leq 0.03$
- and so on

Our state of knowledge about the fairness, or the degree of unfairness, of the coin is then completely summarized by specifying how much we believe these various hypotheses to be true. If we assign a high probability to one (or a closely grouped few) of these hypotheses, compared to others, then this indicates that we are confident in our estimate of the bias-weighting. If there was no such distinction, then it would reflect a high level of ignorance about the nature of the coin.

In this case, our inference about the fairness of the data is summarized by the conditional pdf $prob(H|\{data\} \& I)$. This is just a representation of the limiting case of a continuum of hypotheses for the value of H , that is, the probability that H lies in an infinitesimally narrow range between h and $h+\delta h$ is given by $prob(H=h|\{data\} \& I)dH$. To estimate this posterior pdf, we need to use Baye's theorem(eq 3.7), which relates the pdf of interest to two others that are easier to assign:

$$prob(H|\{data\} \& I) \propto prob(\{data\}|H \& I) \times prob(H|I) \quad (3.25)$$

We have omitted the denominator $prob(\{data\}|I)$ since it does not involve bias-weighting explicitly and replaced the equality by a proportionality. The omitted constant can be determined by normalization

$$\int_0^1 \text{prob}(H|\{data\} \& I) dH = 1 \quad (3.26)$$

The prior pdf, $\text{prob}(H|I)$, on the right side represents what we know about the coin given only that I found the coin. This means that we should keep an open mind about the nature of the coin. A simple probability assignment which reflects this is a uniform pdf

$$\text{prob}(H|I) = \begin{cases} 1 & 0 \leq H \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (3.27)$$

This prior state of knowledge (or ignorance) is modified by the data through the likelihood function, $\text{prob}(\{data\}|H \& I)$, which is a measure of the chance that we would have obtained the data we actually observed if the value of the bias-weighting H was given (as known). If, in the conditioning information I , we assume that the flips of the coin were independent events, so that the outcome of one did not influence that of another, then the probability of obtaining the data R heads in N tosses is given by the binomial distribution

$$\text{prob}(\{data\}|H \& I) \propto H^R (1-H)^{N-R} \quad (3.28)$$

According to eq 3.25, the product of eqs 3.27 and 3.28 gives the posterior pdf that we require. It represents our state of knowledge about the nature of the coin in light of the data.

It is instructive to see how this pdf evolves as we obtain more and more data pertaining to the coin. A computer simulation given by the following IDL code

```

;function probhi,h
;z=1.0D0
;return,z
;end

;function probhi,h
;z=exp((-h-0.5D0)^2)/(0.01))
;return,z
;end

function probhi,h,choice
if choice eq 1 then z=1.0D0
if choice eq 2 then z=exp((-h-0.5D0)^2)/(0.01))
if choice eq 3 then z=exp((-h^2)/(0.005))+exp((-h-1.0D0)^2)/(0.005))
return,z
end

function probdhi,h,n,r
z=exp(r*log(h)+(n-r)*log(1.0D0-h))
return,z
end

function coin,h0
h=randomu(seed)
z = h lt h0
return, z

```

```

end

pro bayes1,choice
h0=0.25D0
nn=1000
r=0
n=0
hplot=0.001D0+findgen(999)*0.001D0
red=[0,1,1,0,0]
green=[0,1,0,1,0]
blue=[0,1,0,0,1]
window,0,xsize=800,ysize=400,xpos=50,ypos=50
tvltct,255*red,255*green,255*blue
val=probhi(hplot,choice)*probdhi(hplot,n,r)
plot,hplot,val/max(val),color=3,xrange=[0.0,1.0],yrange=[-0.2,1.2], $
  xtitle='Bias-weighting for head h', ytitle='prob(h|{data},I)', $
  title='Bayes Simulation - Initial Probability Prob(h|I)', $
  xstyle=1,ystyle=1
window,1,xsize=800,ysize=400,xpos=50,ypos=550
plot,hplot,val/max(val),color=0,xrange=[0.0,1.0],yrange=[-0.2,1.2], $
  xtitle='Bias-weighting for head h', ytitle='prob(h|{data},I)', $
  title='Bayes Simulation - Running Probability Prob({data}|h,I)*prob(h|I)', $
  /nodata, background=1,xstyle=1,ystyle=1
for i=1,nn do begin
  r=r+coin(h0)
  n=i
  val=probhi(hplot,choice)*probdhi(hplot,n,r)
  oplot,hplot,val/max(val),color=2
  oplot,[h0,h0],[-0.2,1.2],color=4
  wait,0.01
  oplot,hplot,val/max(val),color=1
  oplot,[h0,h0],[-0.2,1.2],color=1
  ;wait,0.1
endfor
oplot,hplot,val/max(val),color=2
oplot,[h0,h0],[-0.2,1.2],color=4
print,n,r
end

```

allows us to demonstrate what happens in some typical cases.

The code allows for three distinct and very different prior probabilities:

- (1) Uniform distribution
- (2) Gaussian distribution centered around 0.5 with some spread
- (3) Sum of two Gaussians with different centers

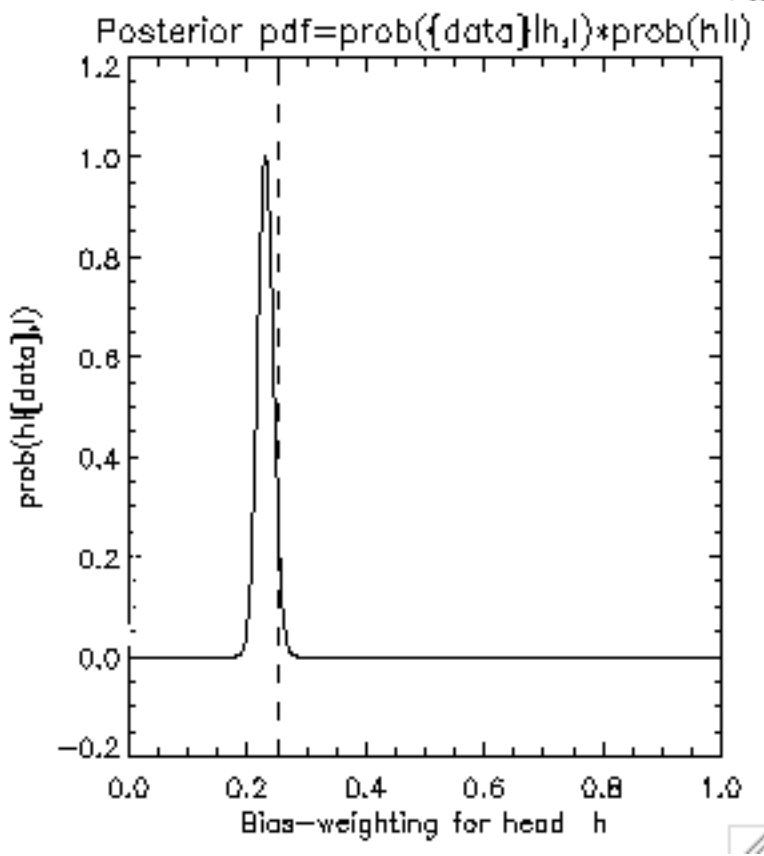
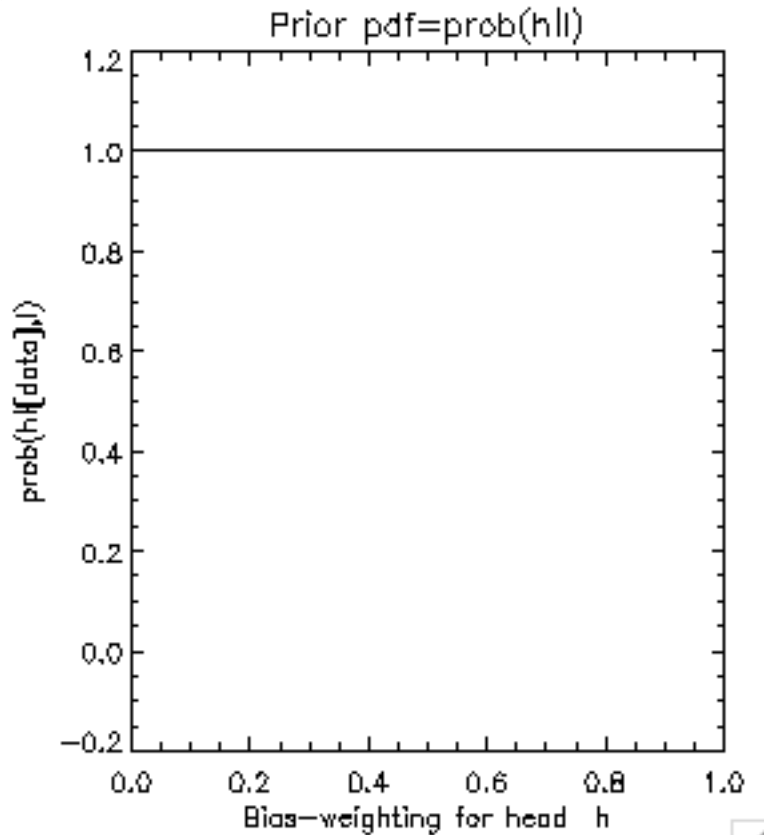
These prior probabilities represent very different initial knowledge

- (1) total ignorance—we have no idea if it is fair
- (2) knowledge that mean is 0.5[with spread]—we think it is fair
- (3) knowledge that it is unfair (either all tails or all heads)
[with spreads]

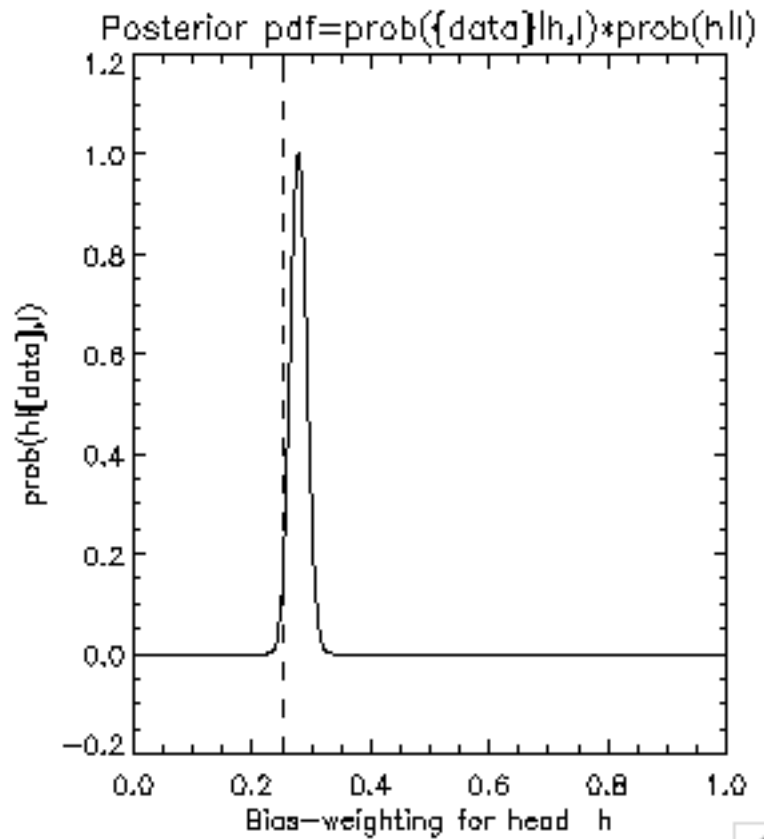
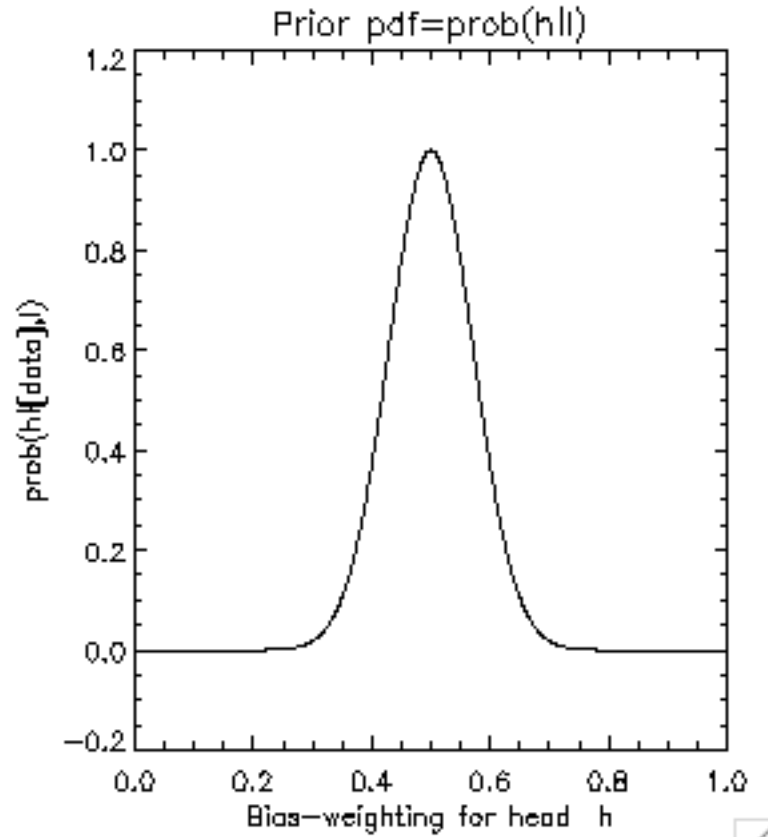
The code also allows us to choose the true mean value (h_0), which is then reflected in the simulated coin tosses (the data).

As can be seen from the images below, the only effect that different prior probabilities have is to change the period of time evolution to the final posterior pdf (which is the same eventually in all cases)!

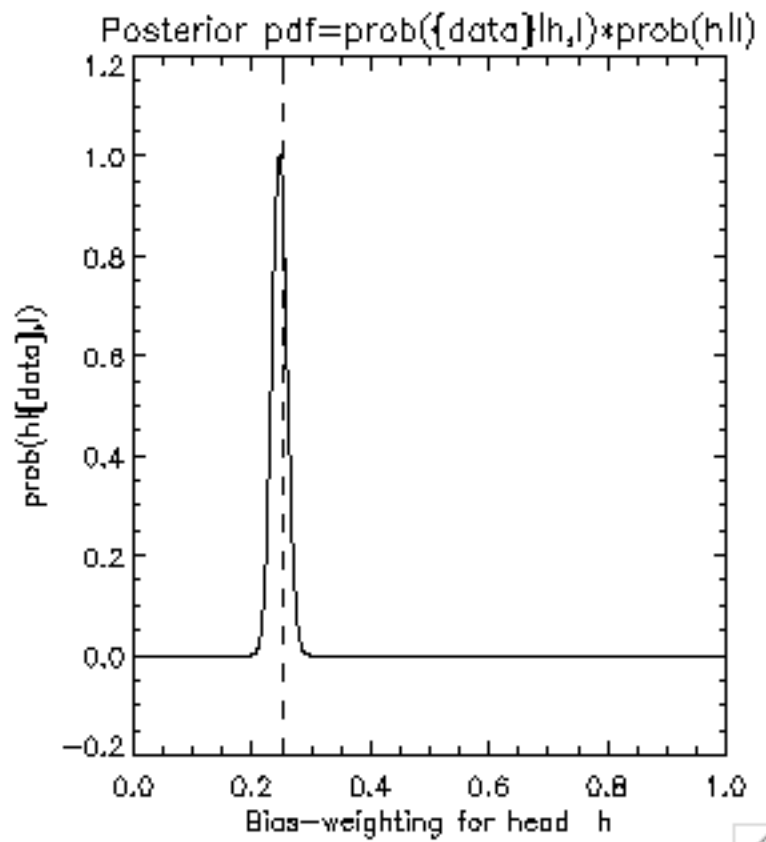
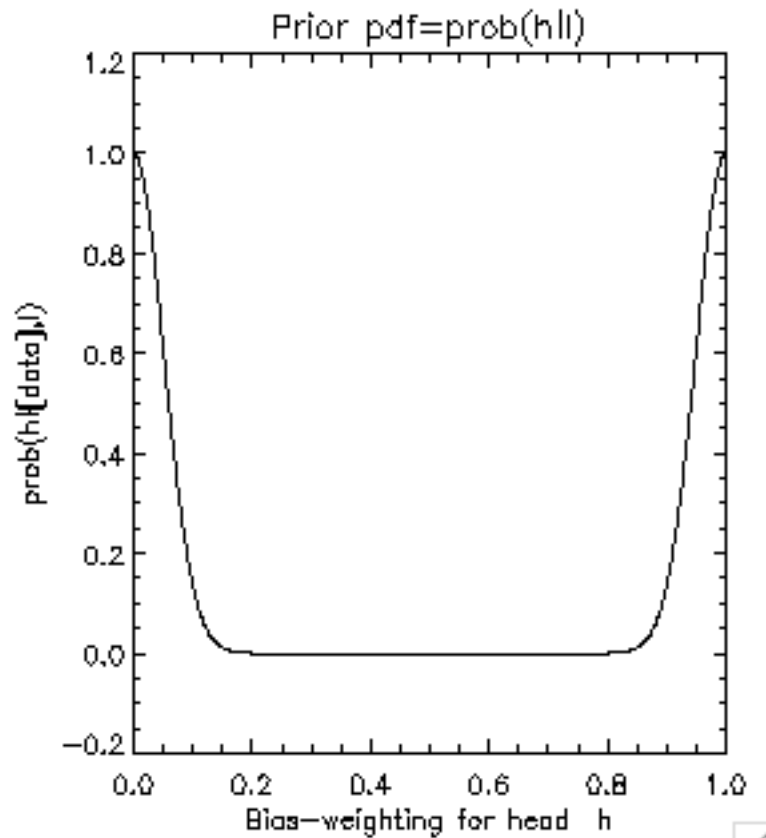
(1) total ignorance - we have no idea if it is fair



(2) knowledge that mean is 0.5 [with spread] - we think it is fair



(3) knowledge that it is unfair (either all tails or all heads) [with spreads]



In each case, the first figure shows the posterior pdf for H given no data (it is the same as the prior pdf) and the second figure shows the posterior pdf after 1000 tosses and they clearly indicate that no

matter what our initial knowledge, the final posterior pdf will be the same, that is, the posterior pdf is dominated by the **likelihood** function(the actual data) and is **independent** of the prior pdf.

The Problem of Prior Probabilities

We are now faced with the most difficult question. How do we assign probabilities based on prior information?

The oldest idea was devised by Bernoulli - the "principle of insufficient reason" or the "principle of indifference". It states that if we determine a set of basic, mutually exclusive, possibilities, and we have no reason to believe that any one of them is more likely to be true than another, then we must assign the same **probability** to each of them. Clearly, this makes sense. Think of flipping a coin with two possibilities, heads and tails. If it is a legitimate coin, then we have no reason to favor heads over tails and we must assign equal probability to each possibility, that is,

$$prob(heads | I) = prob(tails | I) = \frac{1}{2}$$

Let us elaborate on the idea of "not having any reason to believe....". Suppose we had ten possibilities labelled by $X_i, i=1,2,\dots,10$ and we had no reason to think any was more likely than any other. We would then have

$$prob(X_1 | I) = prob(X_2 | I) = \dots = prob(X_{10} | I) = \frac{1}{10}$$

Suppose that we relabel or reorder the possibilities. If the conditioning on I truly represents gross ignorance about any details of the situation, then such a reordering should not make any difference in the probability assignments. Any other statement has to mean that we have other important information besides the simple ordering of the possibilities. For example, imagine that you called a certain side of the coin head and therefore the other side tails. Nothing changes if your friend switches the meaning of heads and tails. This justification of the Bernoulli principle led Jaynes to suggest that we think of it as a consequence of the "requirement of consistency".

This principle of insufficient reason can only be applied to a limited set of problems involving games of chance. It leads, however, to some very familiar and very important results if combined with the product and sum rules of probability theory.

Example 1:

Assume W white balls and R red balls in an urn. We now pick the balls out of the urn randomly. The principle of indifference says that we should assign a uniform prior probability (actually a pdf)

$$prob(j | I) = \frac{1}{R + W}, j = 1, 2, 3, \dots, R + W \quad (3.29)$$

for the proposition that any particular ball, denoted by index j , will be picked. Using the marginalization idea from eq(3.11)

$$\text{prob}(X | I) = \int_{-\infty}^{\infty} \text{prob}(X \& Y | I) dY$$

we have

$$\text{prob}(\text{red} | I) = \sum_{j=1}^{R+W} \text{prob}(\text{red} \& j | I) = \sum_{j=1}^{R+W} \text{prob}(j | I) \text{prob}(\text{red} | j \& I) = \frac{1}{R+W} \sum_{j=1}^{R+W} \text{prob}(\text{red} | j \& I)$$

where we have used the product rule. The term $\text{prob}(\text{red} | j \& I)$ is one if the j^{th} ball is red and zero if it is white. Therefore the summation equals the number of red balls R and we get

$$\text{prob}(\text{red} | I) = \frac{1}{R+W} \sum_{j=1}^{R+W} \text{prob}(\text{red} | j \& I) = \frac{R}{R+W} \quad (3.30)$$

as expected. We have derived this result from the principle of indifference and the product rule. It also follows from the basic notion of probability, that is,

$$\text{prob}(\text{red} | I) = \frac{\text{number of cases favorable to red}}{\text{total number of equally possible cases}} = \frac{R}{R+W}$$

We now assume that after each pick the ball is returned to the urn and we ask the question: what is the probability that N such picks (trials) will result in r red balls?

Using marginalization and the product rule we can write

$$\text{prob}(r | N \& I) = \sum_k \text{prob}(r \& S_k | N \& I) = \sum_k \text{prob}(r | S_k \& N \& I) \text{prob}(S_k | N \& I)$$

where the summation is over the 2^N possible sequences of red-white outcomes $\{S_k\}$ of N picks. The term $\text{prob}(r | S_k \& N \& I)$ equals one if S_k contains exactly r red balls and is zero otherwise so that we need only consider those sequences which have exactly r red outcomes for $\text{prob}(S_k | N \& I)$.

Now we have

$$\text{prob}(S_k | N \& I) = [\text{prob}(\text{red} | I)]^r [\text{prob}(\text{white} | I)]^{N-r} = \frac{R^r W^{N-r}}{(R+W)^N} \quad (3.31)$$

for those S_k that matter and

$$\text{prob}(r | S_k \& N \& I) = \text{prob}(r | N \& I)$$

since we are only considering those S_k which contain exactly r red balls. Finally,

$$\text{prob}(r | N \& I) = \frac{N!}{r!(N-r)!} \quad (3.32)$$

which is the number of sequences (permutations) containing r red balls. Thus,

$$\text{prob}(r | N \& I) = \frac{N!}{r!(N-r)!} \frac{R^r W^{N-r}}{(R+W)^N} = \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r} \quad (3.33)$$

where $p = \frac{R}{R+W}$ = probability of picking a red ball and $q = 1 - p = \frac{W}{R+W}$ = probability of picking a white ball; note that $p + q = 1$ as it should since red and white balls are the only possibilities.

Equation(3.30) allows us to compute the frequency r/N with which we expect to observe red balls. We have

$$\begin{aligned} \left\langle \frac{r}{N} \right\rangle &= \sum_{r=0}^N \frac{r}{N} \text{prob}(r | N \& I) = \sum_{r=0}^N \frac{r}{N} \frac{N!}{r!(N-r)!} p^r q^{N-r} \\ &= \sum_{r=1}^N \frac{(N-1)!}{(r-1)!(N-r)!} p^r q^{N-r} = p \sum_{j=0}^{N-1} \frac{(N-1)!}{j!(N-1-j)!} p^j q^{N-1-j} \quad (3.34) \\ &= p(p+q)^{N-1} = p = \frac{R}{R+W} \end{aligned}$$

as the "expected" or "anticipated" result. Thus, the expected frequency of red balls, in repetitions of the urn" experiment", is equal to the probability of picking one red ball in a single trial.

A similar calculation for the mean-square deviation gives the result

$$\left\langle \left(\frac{r}{N} - \left\langle \frac{r}{N} \right\rangle \right)^2 \right\rangle = \left\langle \left(\frac{r}{N} - p \right)^2 \right\rangle = \frac{pq}{N} \quad (3.35)$$

Since this becomes zero in the limit of large N , it agrees with the result we derived earlier (eq (3.25)). It also verifies that Bernoulli's famous theorem or law of large numbers is valid:

$$\lim_{N \rightarrow \infty} \left(\frac{r}{N} \right) = \text{prob}(\text{red} | I) \quad (3.36)$$

This relationship, which allows prediction of the long-run frequency of occurrence from the probability assignment, goes in a direction opposite to the one we want, that is, we would like to be able to determine the probability of obtaining a red ball, in a single pick, given a finite number of observed outcomes. This is, in fact, exactly what Bayes theorem allows us to do!

How do we generalize Bernoulli's principle of insufficient reason to the case of continuous parameters, that is, when the quantity of interest is not restricted to certain discrete values (heads/tails)?

Suppose we have a variable X which represents the position of some object. We then define a probability as follows. Given the information I , the probability that X lies in the infinitesimal range between x and $x + \delta x$ is

$$\text{prob}(X = x | I) = \lim_{\delta x \rightarrow 0} \text{prob}(x \leq X < x + \delta x | I) \quad (3.37)$$

so that we are treating continuous pdfs as the limiting case of discrete ones. Although it is still awkward to enumerate the possibilities in this case, we can still make use of the principle of consistency which underlies the principle of indifference.

Examples:

A Location Parameter

Suppose that we are unsure about the actual location of the origin. Should this make any difference to the pdf assigned for X ? Since I represents gross ignorance about any details of the situation other than the knowledge that X pertains to a location, the answer must be no; otherwise we must already have information regarding the position of the object. Consistency then demands that the pdf for X should not change with the location of the origin or any offset in the position values. Mathematically, we say

$$\text{prob}(X | I)dX = \text{prob}(X + x_0 | I)d(X + x_0)$$

Since x_0 is a constant, $d(X + x_0) = dX$ so that we have

$$\text{prob}(X | I) = \text{prob}(X + x_0 | I) = \text{constant}$$

so that the complete ignorance about a location parameter is represented by the assignment of a **uniform** pdf.

A Scale Parameter

Suppose that we have another parameter that tells us about size or magnitude, a so-called **scale** parameter. If we are interested in the size L of some object and we have no idea about the length scale involved, then the pdf should be invariant with respect to shrinking or stretching the length scale. Mathematically, the requirement of consistency can be written

$$\text{prob}(L | I)dL = \text{prob}(\beta L | I)d(\beta L)$$

where β is a positive constant. Then since $d(\beta L) = \beta dL$ we must have

$$\text{prob}(L | I) = \beta \text{prob}(\beta L | I)$$

which can only be satisfied if

$$\text{prob}(L | I) \propto \frac{1}{L}$$

which is called **Jeffrey's prior**. It represents complete ignorance about the value of a scale parameter.

Now we must have

$$\text{prob}(L | I)dL = \text{prob}(f(L) | I)df(L)$$

since we are looking at the same domain of values in each case. We then have

$$prob(\log(L) | I) d(\log(L)) = prob(L | I) dL$$

$$prob(\log(L) | I) \frac{dL}{L} = prob(L | I) dL$$

$$prob(\log(L) | I) = L prob(L | I) = \text{constant}$$

So that assignment of a **uniform** pdf for $\log(L)$ is the way to represent complete ignorance about a scale parameter.

3. Testable Information: The Principle of Maximum Entropy

Clearly, some pdfs can be assigned given only the nature of the quantities involved (as we saw above). The methods employed hinge on the use of consistency arguments along with "transformation groups", which characterize the ignorance for a given situation.

For a set of discrete probabilities (finite) the associated pdf must be invariant with respect to any permutation of the propositions (permutation group). In the continuous parameter case, the associated transformations are translation (origin shift) and dilation (shrink/stretch), which are also group transformations.

Now let us move on to a situation where we do not have total ignorance.

Suppose that a die, with the usual six faces, was rolled a very large number of times and we are only told that the average result was 4.5. What probability should we assign for the various outcomes $\{X_i\}$ that the face on top had i dots?

The information I provided by the experiment is written as a simple constraint equation

$$\sum_{i=1}^6 i prob(X_i | I) = 4.5 \quad (3.38)$$

If we had assumed a uniform pdf, then we would have predicted a different average

$$\sum_{i=1}^6 i prob(X_i | I) = \frac{1}{6} \sum_{i=1}^6 i = \frac{1}{6} (21) = 3.5$$

which means the uniform pdf is not a valid assignment.

There are many pdfs that are consistent with eq(3.38). Which one is best?

The constraint equation (3.38) is called **testable information**. With such a condition, we can either accept or reject any proposed pdf. Jaynes proposed that, in this situation, we should make the assignment by using the principle of maximum entropy (MaxEnt), that is, we should choose that pdf which has the most entropy S while satisfying the available constraints.

Explicitly, for case if the die experiment above, we need to maximize

$$S = - \sum_{i=1}^6 p_i \log_e(p_i) \quad (3.39)$$

where $p_i = \text{prob}(X_i | I)$ subject to the conditions

$$\sum_{i=1}^6 p_i = 1 \quad \text{normalization condition} \quad (3.40)$$

and

$$\sum_{i=1}^6 i p_i = 4.5 \quad \text{constraint equation (testable information)} \quad (3.41)$$

Such a constrained optimization is done using the method of **Lagrange multipliers** as shown below:

Define the functions

$$\begin{aligned} f(p_i) &= \sum_{i=1}^6 p_i - 1 = 0 \Rightarrow \frac{\partial f}{\partial p_j} = 0 \\ g(p_i) &= \sum_{i=1}^6 i p_i - 4.5 = 0 \Rightarrow \frac{\partial g}{\partial p_j} = 0 \end{aligned} \quad (3.42)$$

The maximization problem can then be written

$$\frac{\partial S}{\partial p_j} + \lambda_f \frac{\partial f}{\partial p_j} + \lambda_g \frac{\partial g}{\partial p_j} = 0 \quad j = 1, 2, 3, 4, 5, 6 \quad (3.43)$$

where the constants λ_f, λ_g are called undetermined multipliers.

We get the equations

$$-\log_e(p_j) - 1 + \lambda_f + j\lambda_g = 0 \quad j = 1, 2, 3, 4, 5, 6$$

We then obtain

$$\begin{aligned} -\log_e(p_{j+1}) - 1 + \lambda_f + (j+1)\lambda_g &= -\log_e(p_j) - 1 + \lambda_f + j\lambda_g \\ \Rightarrow \log_e \frac{p_{j+1}}{p_j} = \lambda_g &\Rightarrow \frac{p_{j+1}}{p_j} = \beta = \text{constant} \end{aligned}$$

and

$$\begin{aligned} -\log_e(p_1) - 1 + \lambda_f + \log_e(\beta) &= 0 \\ \lambda_f &= 1 + \log_e \frac{p_1}{\beta} \end{aligned}$$

Therefore,

$$\begin{aligned} \sum_{i=1}^6 p_i = 1 &= p_1(1 + \beta + \beta^2 + \beta^3 + \beta^4 + \beta^5) \\ \sum_{i=1}^6 i p_i = 4.5 &= p_1(1 + 2\beta + 3\beta^2 + 4\beta^3 + 5\beta^4 + 6\beta^5) \end{aligned}$$

or

$$\frac{1+2\beta+3\beta^2+4\beta^3+5\beta^4+6\beta^5}{1+\beta+\beta^2+\beta^3+\beta^4+\beta^5} = 4.5$$

$$1.5\beta^5+0.5\beta^4-0.5\beta^3-1.5\beta^2-2.5\beta-3.5=0$$

Solving for β we get 1.449255 so that

$$p_1 = \frac{1}{1+\beta+\beta^2+\beta^3+\beta^4+\beta^5} = 0.05435$$

$$p_2 = \beta p_1 = 0.07877$$

$$p_3 = \beta p_2 = 0.11416$$

$$p_4 = \beta p_3 = 0.16545$$

$$p_5 = \beta p_4 = 0.23977$$

$$p_6 = \beta p_5 = 0.34749$$

is the MaxEnt assignment for the pdf for the outcomes of the die roll, given only that it has the usual six faces and yields an average result of 4.5.

Why should the function specified in eq(3.39) be the choice for a selection criterion?

Let us look at two examples that suggest this criterion is highly desirable and probably correct.

Kangeroo Problem(Gull and Skilling)

The kangeroo problem is as follows:

Information: 1/3 of all kangeroos have blue eyes and 1/3 of all kangeroos are left-handed

Question: On the basis of this information alone, what proportion of kangeroos are both blue-eyed and left-handed?

For any particular kangeroo, there are four distinct possibilities, namely, that it is

- (1) blue-eyed and left-handed
- (2) blue-eyed and right-handed
- (3) not blue-eyed but left-handed
- (4) not blue-eyed but right-handed

Bernoulli's law of large numbers says that the expected values of the fraction of kangeroos with characteristics (1)-(4) will be equal to the probabilities (p_1, p_2, p_3, p_4) we assign to each of these propositions.

This is represented by a 2x2 truth or contingency table as shown below:

	<i>Left - Handed True</i>	<i>Left - Handed False</i>
<i>Blue - eyed True</i>	p_1	p_2
<i>Blue - eyed False</i>	p_3	p_4

Although there are four possible combinations of eye-color and handedness to be considered, the related probabilities are not completely independent of each other. We have the standard normalization requirement

$$\sum_{i=1}^4 p_i = 1$$

In addition, we also have two conditions on the marginal probabilities

$$p_1 + p_2 = \text{prob}(\text{blue \& left} | I) + \text{prob}(\text{blue \& right}) = \text{prob}(\text{blue} | I) = 1/3$$

$$p_1 + p_3 = \text{prob}(\text{blue \& left} | I) + \text{prob}(\text{not - blue \& left}) = \text{prob}(\text{left} | I) = 1/3$$

Since any $p_i \geq 0$, these imply that $0 \leq p_1 \leq 1/3$. Using this result we can characterize the contingency table by a single variable $x = p_1$ as in the table below:

	<i>Left - Handed True</i>	<i>Left - Handed False</i>
<i>Blue - eyed True</i>	$0 \leq x \leq \frac{1}{3}$	$\frac{1}{3} - x$
<i>Blue - eyed False</i>	$\frac{1}{3} - x$	$\frac{2}{3} + x$

where we have used

$$x = p_1$$

$$p_1 + p_2 = \frac{1}{3} \rightarrow p_2 = \frac{1}{3} - x$$

$$p_1 + p_3 = \frac{1}{3} \rightarrow p_3 = \frac{1}{3} - x \quad , \quad p_1 + p_2 + p_3 + p_4 = 1 \rightarrow p_4 = \frac{2}{3} + x$$

All such solutions, where $0 \leq x \leq 1/3$, satisfy the constraints of the testable information that is available. Which one is best?

Common sense leads us towards the assignment based on independence of these two traits, that is, any other assignment would indicate a knowledge of kangaroo eye-color told us something about its handedness. Since we have no information to determine even the sign of any potential correlation, let alone its magnitude, any choice other than independence is not justified.

The independence choice says that

$$x = p_1 = \text{prob}(\text{blue \& left} | I) = \text{prob}(\text{blue} | I)\text{prob}(\text{left} | I) = \frac{1}{9} .$$

In this particular example it was straightforward to decide the most sensible pdf assignment in the face of the inadequate information.

We now ask whether there is some function of the $\{p_i\}$ which, when maximized subject to the known constraints, yields the "independence" solution. The importance of finding an answer to this question is that it would become a good candidate for a general **variational principle** that could be used in situations that were too complicated for our common sense.

Skilling has shown that the only function which gives $x=1/9$ is the entropy S in eq(3.39) or

$$S = -\sum_{i=1}^4 p_i \log_e(p_i) = -x \log_e(x) - 2\left(\frac{1}{3} - x\right) \log_e\left(\frac{1}{3} - x\right) - \left(\frac{1}{3} + x\right) \log_e\left(\frac{1}{3} + x\right)$$

The results of Skilling's investigations, including three proposed alternatives, is shown in the table below:

Variational Function	Optimal x	Implied Correlation
$-\sum p_i \log_e(p_i)$	0.1111	None
$-\sum p_i^2$	0.0833	Negative
$-\sum \log_e(p_i)$	0.1301	Positive
$-\sum \sqrt{p_i}$	0.1218	Positive

Clearly, only the MaxEnt assumption leads to an optimal value with no correlations as expected.

Let us look at another example that lends further support to the MaxEnt principle.

The Team of Monkeys

Suppose there are M distinct possibilities $\{X_i\}$ to be considered. How can we assign truth tables ($prob(X_i|I) = p_i$) to these possibilities given some testable information I . What is the most honest and fair procedure?

Imagine playing the following game.

The various propositions are represented by different boxes all of the same size into which pennies are thrown at random. The tossing job is often assigned to a team of monkeys under the assumption that this will not introduce any underlying bias into the process.

After a very large number of coins have been distributed into the boxes, the fraction found in each of the boxes gives a possible assignment of the probability for the corresponding $\{X_i\}$.

The resulting pdf may not be consistent with the constraints of I , of course, in which case it must be rejected as a potential candidate.

If it is in agreement, then it is a viable option.

The process is then repeated by the monkeys many times. After many such trials, some distributions will be found to come up more often than others. The one that occurs most frequently (and satisfies I) would be a sensible choice for $prob(\{X_i\}|I)$. This is so because the team of monkeys has no axe to grind (no underlying bias) and thus the most frequent solution can be regarded as the one that best represents our state of knowledge. It agrees with all the testable information available while being as indifferent as possible to everything else.

Does this correspond to the pdf with the greatest value of $-\sum p_i \log_e(p_i)$?

After the monkeys have tossed all the pennies given to them, suppose that we find n_1 in the first box, n_2 in the second box, and so on. We then have

$$N = \sum_{i=1}^M n_i = \text{total number of coins} \quad (3.44)$$

which will be assumed to be very large and also much greater than the number of boxes M .

This distribution gives rise to the candidate pdf $\{p_i\}$ for the possibilities $\{X_i\}$:

$$p_i = \frac{n_i}{M}, \quad i = 1, 2, \dots, M \quad (3.45)$$

Since every penny can land in any of the boxes there are M^N number of different ways of tossing the coins among the boxes. Each way, by assumption of randomness and no underlying bias by the monkeys, is equally likely to occur. All of the basic sequences, however, are not distinct, since many yield the same distribution $\{n_i\}$. The expected frequency F with which a $\{p_i\}$ will arise, is given by

$$F(\{p_i\}) = \frac{\text{number of ways of obtaining } \{n_i\}}{M^N} \quad (3.46)$$

The numerator is just the number of ways to distribute N coins in a distribution $\{n_i\}$ which is given by

$$\text{number of ways of obtaining } \{n_i\} = \frac{N!}{n_1! n_2! \dots n_M!} \quad (3.47)$$

Using (3.44) in (3.43) we get

$$\log(F) = -N \log(M) + \log(N!) - \sum_{i=1}^M n_i \log(n_i!) \quad (3.48)$$

Using Stirling's approximation $\log(n!) \approx n \log(n) - n$ for n large, we find

$$\begin{aligned}\log(F) &= -N\log(M) + N\log(N) - \sum_{i=1}^M n_i \log(n_i) - N + \sum_{i=1}^M n_i \\ &= -N\log(M) + N\log(N) - \sum_{i=1}^M n_i \log(n_i)\end{aligned}$$

Now using (3.42) we get

$$\begin{aligned}\log(F) &= -N\log(M) + N\log(N) - \sum_{i=1}^M p_i N \log(p_i N) \\ &= -N\log(M) + N\log(N) - \sum_{i=1}^M p_i N (\log(p_i) + \log(N)) \\ &= -N\log(M) + N\log(N) - N \sum_{i=1}^M p_i \log(p_i) + N\log(N) \sum_{i=1}^M p_i \quad (3.49) \\ &= -N\log(M) + N\log(N) - N \sum_{i=1}^M p_i \log(p_i) + N\log(N) \\ &= -N\log(M) - N \sum_{i=1}^M p_i \log(p_i)\end{aligned}$$

Maximizing the $\log(F)$ is equivalent to maximizing F , which is the expected frequency with which the monkeys will come up with the candidate pdf $\{p_i\}$, that is, maximizing $\log(F)$ will give us the assignment $\text{prob}\{X_i|I\}$ which best represents our state of knowledge consistent with the testable information I . Since M and N are constants, this is equivalent to the constrained maximization of the entropy function $S = -\sum p_i \log_e(p_i)$.

Discussion

In discussions of Bayesian methods, opponents often use the words 'subjective probabilities' to say that the methods are not as valid as normal "objective" probability theory.

These opponents are misguided.

The main point of concern centers around the choice of the prior pdf, that is, what should we do if it is not known?

This is actually a very strange question. It is usually posed this way by opponents of the Bayesian methods in an attempt to prove its subjective nature.

No probability, whether prior, likelihood or whatever, is ever "known. It is simply an assignment which reflects the relevant information that is available. Thus, $\text{prob}(x|I_1) \neq \text{prob}(x|I_2)$, in general, where the conditioning statements I_1 and I_2 are different.

Nevertheless, objectivity can, and must, be introduced by demanding the two people with the same information I should assign the same pdf. I think that this consistency requirement is the most important idea of all.

We have seen how invariance arguments, under transformation groups,

can be used to uniquely determine a pdf when given only the nature of the quantities involved. MaxEnt provides a powerful extension when we have testable constraints.

While we may yet be far from knowing how to convert every piece of vague information into a concrete probability assignment, we can deal with a wide variety of problems with these ideas.

The important point is that nowhere in our discussion have we explicitly differentiated between a prior and a likelihood. We have only considered how to assign $\text{prob}(X|I)$ for different types of I . If X pertains to data, then we call $\text{prob}(X|I)$ a likelihood. If neither X nor I refers to (new) measurements, then we may say it is a prior.

The distinction between the two cases is one of nomenclature and not of objectivity or subjectivity. If it appears otherwise, then this is because we are usually prepared to state conditioning assumptions for the likelihood function but shy away from doing likewise for the prior pdf.