# Reading for Lecture 3

***Release v10***

**Christopher Lee**

September 30, 2011

## 1 Essential Probability Concepts

### 1.1 Normalization

The probabilities of a set of disjoint subsets must sum to 1.

#### Formal Definition

Since probability is defined as normalized (divided) by the probability measure of the total set, it follows that the probability of a complete set of disjoint subsets $s$ must sum to 1

$$\sum_s p(s) = 1$$

since the denominator is the same in all terms, and the numerators sum to the probability measure of the union of the disjoint subsets, i.e. the entire set.

### 1.2 Normalization of a Sequence

Consider a sequence $\vec{X}^n$ of random variables $X_1, X_2, ...X_n$ all drawn independently from the same distribution.

- Since they are independent, $p(\vec{X}^{i+j}) = p(\vec{X}^i)p(\vec{X}^j)$. If the average value of $p(X) = c$ (which is less than 1, unless all the probability is confined to a single state), then we expect the average value of $p(\vec{X}^n) = c^n$ to decrease exponentially with increasing $n$.
- Now consider

$$p(\vec{X}^0) = \frac{p(\vec{X}^{n+0})}{p(\vec{X}^n)} = 1$$

  This implies that the probability of the empty sequence $\vec{X}^0$ is just unity (1). Since the joint probability of multiple variables is equivalent to the intersection of the constraints imposed by each variable, you can think of the empty sequence as "no constraints", i.e. just the whole Venn diagram.

## 1.3 Random variable

A *labeling* of the set of possible events emitted by a stochastic process, where each possible *label value* is one *state* of the random variable.

### Formal Definition

Consider the set of all possible events that the stochastic process can emit. A specific random variable is defined as a particular division of that set into a set of disjoint subsets that each map to a unique state (value) of the variable. Note that this associates with each possible state of the variable a *probability measure* i.e. the size of the event subset that maps to it. Since the union of event subsets for all states is simply the complete event set, this probability measure sums to 1 (obeys normalization).

### Comments

- If the variable is discrete, the set of subsets is countable; if it continuous, the set of subsets is uncountable.
- Note that this definition makes explicit how a stochastic process emits *multiple* random variables: they are just *different* ways of dividing and labeling the same event set.
- The possible values of a random variable are referred to as its possible *states*, or more formally a "random state" or "random variate".
- A random variable is usually written as a capital letter e.g. $X$, whereas its possible *states* are usually written as lower case letters e.g. $x$. Thus $p(X = x)$ is the probability that random variable $X$ will take the value $x$.
- Note that whereas the same probability syntax $p(\cdot)$ means something quite different for a random variable vs. a random state. Specifically $p(x)$ is a number (i.e. a probability between zero and one inclusive), whereas $p(X)$ is a *function* that maps each possible value of $X$ to its corresponding probability measure. If $X$ is discrete, this is a probability mass function; if continuous, a probability density function.

## 1.4 Random state

A particular outcome of a random variable, i.e. one of its possible values or *states*.

### Formal Definition

A specific subset of the set of all possible events, which for this random variable maps to a unique *value* of this variable.

## 1.5 Random Variable vs. State Notation Convention

By convention, we write random variables using capital letters e.g. $X$, whereas we write possible *values* of such a variable using lower-case letters e.g. $x$. For example, we might say that the possible values of variable $X$ are $X \in \{x_1, x_2, ...x_m\}$. By contrast, writing $X_1, X_2, ...X_n$ implies that each $X_i$ is a distinct random variable.

## 1.6 Probability Mass Function (pmf)

A function that maps each possible value of a discrete random variable to its associated probability. As such it must of course obey basic principles of probability such as normalization.

## 1.7 Probability Density Function (pdf)

A function that maps each possible interval of values of a continuous random variable to its associated probability, by mapping each possible *value* to its associated *probability density*. I.e. if $\rho(X)$ is the pdf of a continuous random variable $X$, then

$$p(X \in [x, x + \Delta x]) = \int_x^{x+\Delta x} \rho(X) dX$$

## 1.8 Defining "Zero Information" as Independence

When we look at the world, we see many things that appear to be be connected, but we don't know why. The task of inference is to formulate and test hypotheses about these apparent connections. To start with, we should think carefully about how different variables can be connected statistically. For example, we would like an operational definition of *information*, i.e. whether one variable $X$ contains any information about another variable $Y$. We can formulate this very simply as *prediction value*, specifically, whether knowing the value of $X$ tells us anything new about $Y$. If

$$p(Y|X) = p(Y)$$

then by definition $X$ gives no information about $Y$; the distribution of $Y$ is truly independent of $X$. We adopt this as the definition of "zero information".

## 1.9 Probability Factoring vs. Coupling

Note that we can also multiply both sides by $p(X)$

$$p(Y|X)p(X) = p(Y)p(X)$$

to obtain the symmetric form

$$p(Y, X) = p(Y)p(X)$$

You can picture this as a rectangular Venn diagram where the two variables are orthogonal slicings, i.e. the x-axis represents different states of the $X$ variable, and the y-axis represents different states of the $Y$ variable. Assuming that the probability of each cell is proportional to its area in the Venn diagram, its area (as a fraction of the whole rectangle) must be equal to the product of the fractional areas of the horizontal and vertical "slices" that intersect in that cell. Note that this must be true for *every* cell in the joint probability diagram; otherwise the variables are *not* independent.

This property is commonly referred to as "statistical independence": we say "$X$ and $Y$ are independent." Thus, by our definition, independence is equivalent to *zero information*.

It is convenient to recast this property by taking the log of the probability. Defining $L(X) \equiv \log p(X)$, the independence condition becomes

$$L(X, Y) = L(X) + L(Y)$$

i.e. that the log-probability can be expressed as the sum of completely separate functions $f(X)$ and $g(Y)$. The presence of any term $h(X, Y)$ (that cannot be thus separated) will break independence. We refer to such terms as *coupled*. By definition, information contained in one variable about another variable can only flow through such coupled terms. Thus these coupled terms represent the true structure of information flow in any problem, and are of the greatest interest for understanding a problem and how to solve it.

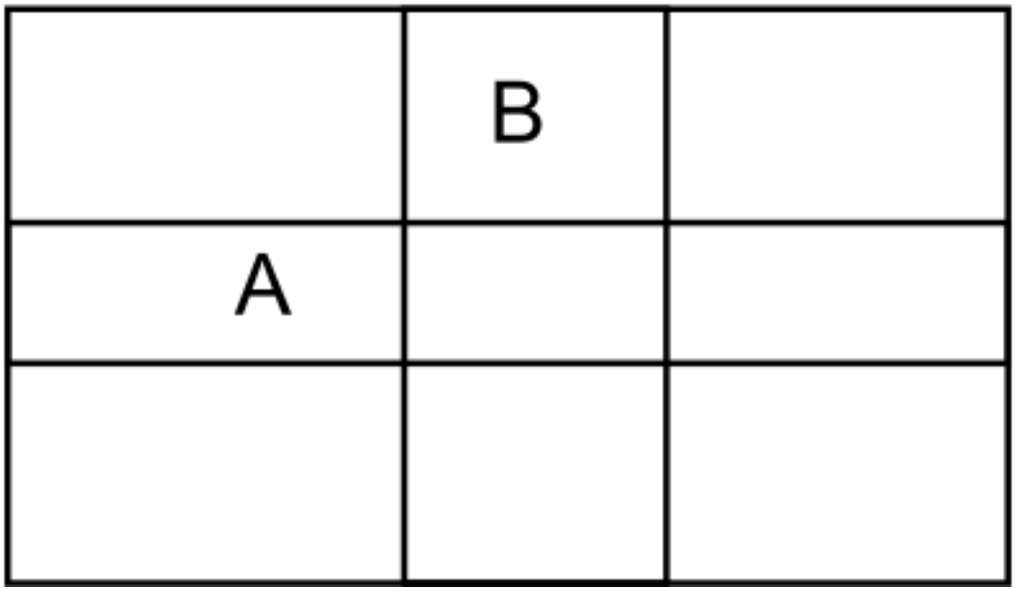


Figure 1: **A graphical illustration of independence of X and Y variables**
*The X variable is shown as being split into three discrete values on the x-axis. The Y variable is shown as being split into three discrete values on the y-axis. The probability of a given X value or (X,Y) pair is shown by the area of that region in the figure. For example, the cell labeled A represents the joint probability of the intersection* $(X = x_1, Y = y_2)$*, and B represents* $(X = x_2, Y = y_3)$*. Since the probability (area) of any given (X,Y) pair is simply proportional to the product of the probabilities (area fractions) of that value of X and of that value of Y, they are independent.*

# 2 Some Common Probability Misconceptions

## 2.1 Conditioning as causality

One commonsense view of conditional probability focuses on whether the *condition* could *cause* the subject state to change. If not, the condition should not alter the probability of the subject state. This is a variation on the general theme of confusing correlation and causation, common among both laymen and scientists with a pet theory.

### Formal Definition

The extreme form of this view asserts that $p(X|Y) = p(X)$ unless $Y$ actually *causes* $X$. This forgets that $X, Y$ may simply be correlated, for example if they are both "caused" by a third variable $Z \to X, Z \to Y$.

## 2.2 Normalization over conditions

Summing a conditional probability over its condition variable(s) produces an undefined quantity. Specifically, it does not sum to 1! Normalization only applies to the *subject* variable.

### Formal definition

Correct normalization always means summing over the *subject* variable

$$\sum_X p(X|Y) = 1$$

not the condition variable

$$\sum_Y p(X|Y) \neq 1$$

Note that this sum has no meaning as a probability, because it is *not* a probability. For example, it can be much greater than one. To turn this into a meaningful probability we'd have to multiply by $p(Y)$, which turns this into

$$\sum_Y p(X|Y)p(Y) = \sum_Y p(X,Y) = p(X)$$

## 2.3 Mixing up converse probabilities

Common-sense probability is imprecise about the exact relationship between converse conditional probabilities. It often treats them as *proportional*, i.e. if $p(O|H)$ is high, then $p(H|O)$ must be high; if $p(O|H)$ is low, then $p(O|H)$ must be low.

In the worst case, people simply mix up converses, i.e. they use one in place of the other, failing to distinguish their separate meanings. This may be an innocent error, but often it's a cynical attempt to justify a prior belief or marketing goal with any plausible-sounding statistics they can find.

### Formal definition

The common-sense viewpoint seems to spring from the fundamental fact that the converse "joint probability products" are indeed exactly equal

$$p(H|O)p(O) = p(H,O) = p(O|H)p(H)$$

"Other things being equal", $p(O|H)$ and $p(H|O)$ are indeed proportional. Specifically, if the prior $p(H)$ is constant (uninformative), and we vary $H$ while holding $O$ constant, then

$$p(H|O) = Cp(O|H)$$

where the proportionality constant is the marginal odds ratio $C = p(H)/p(O)$.

This would appear to be a great cognitive convenience for a common inference scenario:

- We are confronted with some situation where we must assess several possible interpretations $h_1, h_2, ...$ based upon a specific set of observations $obs$. In other words $O = obs$ is indeed held fixed.
- We have essentially no prior information about the occurence of $h_1, h_2, ....$ In other words, $p(H)$ is indeed uninformative.

In this case $p(obs|H)$ is indeed our sufficient statistic for assessing different values of $H$.

### Comments

Note that classical statistics in many ways adds to this confusion by giving "official standing" to this belief, in the form of p-value tests. P-value tests use $p(O \geq obs|h_0)$ as a proxy for testing $p(h_0|obs)$. Unfortunately, people become so used to doing this that they no longer think about the underlying chain-rule logic. They simply become accustomed to equating the converses. While classical statistics had its own logical reasons for taking this approach (extreme discomfort with the problem of how to determine priors), in real-world practice it often turns into simply drawing conclusions based on a fallacy.

# 3 A Recipe for Inference

## 3.1 Pure Inference

The projection operation is very useful for Bayesian inference when expressed in the following form:

$$\sum_{i=0}^{n} p(A|B_i)p(B_i) = p(A)$$

This enables us to rewrite Bayes Law in the form:

$$p(H|O) = \frac{p(O|H)p(H)}{p(O)} = \frac{p(O|H)p(H)}{\sum_{\forall h} p(O|h)p(h)}$$

where the hidden variable $H$ is summed over all possible disjoint values $h$. This can be considered a "pure" form of inference in that it replaces the somewhat mysterious term for the probability of the observations $p(O)$, with an explicit calculation that is entirely based on inference models. In other words, the entire calculation is supplied purely by models (specifically, their likelihood functions and priors). The denominator in this expression can be considered a "normalization" in the sense that it is simply a summation of the term in the numerator over all possible values of the hidden variable $H$. Whereas it's often not obvious how to calculate the "probability of the observations" $p(O)$ directly, we now have a simple expression that depends only on the likelihood model $p(O|H)$ and the prior $p(H)$, the same factors that we need for the numerator.

## 3.2 The Bayesian Recipe

We can use Bayes Law as a "recipe" whose parts give us a very clear list of the ingredients necessary for solving any inference problem:

$$p(H|O) = \frac{p(O|H)p(H)}{\sum_{\forall h} p(O|h)p(h)}$$

We can list these ingredients directly from the terms in Bayes Law:

- *What is hidden* ($H$)? The core of inference is distinguishing clearly between hidden variables vs. observed variables. In general, almost anything we really want to know is hidden; the real question is how to formulate what we want to know as a precise mathematical parameter. Choosing the right variables for describing the hidden state is often the key to solving the problem. When dealing with the hidden, discovering a clear statement of *what you want to know* means deciding which aspects of the outward appearance of a problem are extraneous and should be ignored, versus which part(s) are core.

- *What is observed* ($O$)? This step is usually more straightforward than choosing the hidden variables. We must be careful not to miscategorize as "observable" quantities that actually are hidden. In general, anything that has uncertainty cannot be considered to be "observable", and should instead be considered hidden. Next, we must ensure that the way we enumerate observations allows no possibility of double-counting; i.e. each "observation" must truly be independent of every other observation in the observation dataset. We will examine this issue carefully later.
- *What is the likelihood model* ($p(O|H)$)? In some sense, the decisions of what hidden variable(s) to use, and what likelihood model(s) to use are really the same decision. That is, a hidden variable only has meaning to the degree that it is associated with a *model* for how it affects observable variables. Choosing a question to ask about the "hidden" world corresponds to choosing a model of that hidden world, consisting of a *structure* (e.g. "independent binomial events"); a set of *hidden variables* associated with that model (e.g. "a binomial probability $\Theta$"); and a *likelihood equation* that shows how the hidden variables influence the probability of observable events.
- *What is the prior* ($p(H)$)? There are two types of priors: those derived from previous datasets (as posteriors); and *uninformative* priors. If a prior is derived from a posterior, that posterior must have been computed from a completely separate set of observations, and the likelihood model must be of a form in which separate observations *factor* (i.e. can be separated into multiplicative factors). The most common uninformative prior is just a *constant*; in this case, the prior simply cancels from numerator and denominator. However, it should be remembered that priors are important, and that they are one of the major differences between Bayesian inference and other approaches (e.g. maximum likelihood, which we will explore later).
- *What is the set of all possible models?* The summation in the denominator must be taken over all possible values of the hidden variable(s), which in turn must be *disjoint*. This summation is fundamentally the most challenging step of inference, because strictly speaking it implies consideration of all possible models. For any non-trivial multi-dimensional problem, the number of terms grows exponentially, and summation by exhaustive enumeration is impractical. There are two possible outcomes. If a large subset of terms contribute substantial probability, then by definition no one term will have high posterior probability. In such cases there may be no reason to complete this calculation, since no individual model will have confidence. Once we know that "the available data do not permit a confident conclusion", there is often little point in computing the uncertainty to the last decimal point. In the opposite case where only one term contributes any significant probability, we can ignore all the remaining terms with little loss of accuracy in any of our computations.
- *How do we find models with significant probability?* Usually, we want to ignore terms that do not contribute significantly, while finding and retaining those that do. This transforms the summation into a *search* problem: efficient methods for finding model terms that contribute significant probability, with low probability of missing significant terms. This is the real, algorithmic challenge of inference, and the source of different computational complexities for different inference problems.
- *What is the posterior* ($p(H|O)$)? With all of the above ingredients in hand, we can finally calculate the result, the posterior evidence for a specific model $H$ given the set of observations $O$.

## 3.3 Example: What is the probability the sun will rise?

How could you calculate the probability that the sun will rise tomorrow? Stated in a vacuum, this question may appear somewhat mysterious or even absurd. On what basis could you calculate a mathematical probability for such a strange real-world question? The mathematician Pierre-Simon Laplace worked out a clever solution to this problem in the 18th century. However, as we all know, the problem with clever solutions is that they require being clever. Fortunately, the "recipe" of Bayesian inference provides a systematic and straightforward way to solve this problem:

- *What is hidden?* The question presupposes that there is a "probability that the sun will rise", and this is evidently a hidden variable. Let's call it $\theta$.
- *What is observed?* This question sharpens our thinking about the problem considerably. Treating it purely as an empirical problem, what observations do we have that are relevant to the hidden variable? Plainly, this is not the first time the "sun rising" question has been tested; you have seen the sun rise many times. Each of these

"events" is relevant, and can be quantified as the number of times you have observed the sun rise $n$ out of the total number of days you observed $N$. For the moment, let's assume you live in a temperate latitude and $n = N$.

- *What is the likelihood model?* The concept of a single "probability $\theta$ that the sun will rise" corresponds to a model of independent, binomial events: in each event, the sun either rises (with probability $\theta$) or doesn't rise (with probability $1 - \theta$). Under this model, the likelihood is simply the probability of $N$ independent "successes":

$$p(n = N|\theta) = \theta^N$$

- *What is the prior?* In the absence of other, prior information about the process governing sunrise, we can use an uninformative prior $p(\theta) = 1$.
- *What is the set of all possible models?* $\theta$ is a continuous variable with a range $\theta \in [0, 1]$. Also, we can check that our prior is normalized, i.e. $\int_0^1 p(\theta)d\theta = 1$
- *What is the posterior :math:'p(theta|n=N)'?*

$$p(\theta|n = N) = \frac{p(n = N|\theta)p(\theta)}{\int_0^1 p(n = N|\theta)p(\theta)d\theta} = \frac{\theta^N}{\int_0^1 \theta^N d\theta} = (N + 1)\theta^N$$

For $N = 0$, this just gives us back the prior (as it should, in the absence of any observations), a flat probability distribution. For $N = 1$ it yields a linear increase in probability density from 0 at $\theta = 0$ to a maximum at $\theta = 1$. For higher values of $N$, it becomes a curve that is increasingly sharply peaked at $\theta = 1$. This takes into account the intuitive fact that few observations give low confidence, but more observations give stronger and stronger confidence that "the sun always rises".

From this we can compute the expectation value for $\theta$:

$$E(\theta) = \int_0^1 \theta p(\theta|n = N)d\theta = \int_0^1 \theta(N + 1)\theta^N d\theta = \frac{N + 1}{N + 2}$$

If the sun has risen every day of the 20 years during which you have been conscious (approximately 7300 days), your estimate of the probability the sun will rise tomorrow is about 99.986%.

- This example illustrates the "pseudocount principle": Bayesian inference often behaves as if a single additional observation count is added to each of the possible hidden states. The pseudocount principle addresses the problem of insufficient observations for a hidden state: just because there may be zero observations of a particular state in a given dataset, does not mean that our estimate of its hidden frequency is zero. The pseudocount principle deals with this very simply, by just adding a single "observation" to each hidden state. Thus all states are guaranteed to have non-zero observation counts. For example, in this case we have two possible states ("sun rises"; "sun doesn't rise"), so we add one observation count to each, yielding observation counts of $N + 1$ and 1 respectively. This simplistic analysis yields exactly the same maximum likelihood estimate ($\theta = \frac{N+1}{N+2}$) as the Bayesian posterior expectation value computed using an uninformative prior, above.

This example, while silly, illustrates the utility of the "inference recipe": it provides a systematic series of steps for solving any problem, taking advantage of whatever empirical data that are relevant to the problem, and giving an explicit and rigorous measure of the level of uncertaintly in our inference.