Readings for Lecture 2

Release 0.1

Christopher Lee

September 26, 2011

Contents

I think the most interesting question in the world is *how we think*. This is one of the basic questions of life – but how well do we understand it? Our difficulty is not a shortage of ideas, but rather that different fields give discordant answers. For example, mathematics and science provide two very different outlooks on how we think. On the one hand, mathematical logic is all about reasoning under absolute certainty – through mathematical proof. On the other hand, science always operates under *uncertainty*, both in its hypotheses and its empirical tests.

To see how these differ, let's imagine a scientist, Sonya, discussing this question with a mathematician, Matt. He says, "Mathematical logic is a rigorous theory of truth that is the *definition* of correct thinking. So the problem is already solved: *thinking* just means deriving a conclusion from some starting information according to the principles of mathematical proof."

Sonya says, "That's an interesting theory. How could we test it? Does it make any predictions?"

Matt chuckles. "Actually, my argument is that this is true *by definition*. After all, a thought is *correct* only if it can be proved mathematically. So correct thinking must follow the process of mathematical proof. Any departure from that process will be incorrect, and therefore lies outside our definition. Surely there is no need to dissect a hamster brain, or do some such experiment, to test a statement that is more a definition than a theorem."

"I see what you mean. Still, your idea seems to make specific predictions that are testable. For example, computers can perform billions of logical operations per second, far faster and more exactly than any human being. If thinking is just applying the rules of mathematical logic, we could program a computer to use mathematical logic to solve *math* problems, and it ought to be able to out-think you mathematicians."

"Oh no! People have tried that. You can't just give a computer the basic rules of logic and some axioms, and ask it to prove a desired statement. To get a program to generate even a simple proof, like the irrationality of $\sqrt{2}$, you have to input endless lists of hints and rules. I think the problem is that there's a combinatorially infinite number of ways to try to prove a given statement, and the computer doesn't know which one will work. So it has to try each possible path until it hits a dead-end, and then try the next path..."

"But wait. Doesn't a human mathematician face the same problem?"

"Sure. The only way to be logically certain that a given path will lead to a proof, would be if you already had the complete proof. That is an all-or-nothing proposition. But a mathematician doesn't just mindlessly try every possible path, like a computer. Instead he thinks about what strategy is *likely* to work, and uses his intuition."

"You seem to be saying that whereas the computer only has the all-or-nothing method of logical proof, a good mathematician is guided by a feeling for the probabilities, and can estimate the likelihood that a strategy will work."

"Yes. The mathematician George Polya called this process of reasoning under uncertainty 'plausible inference'. He wrote several books about this, but I doubt most mathematicians think about it a lot. The conundrum is that mathematical proof gets all the glory, but it gives only a *final confirmation* of ideas and strategies that you had to choose *before* you had a proof – using plausible inference."

"So mathematical proof deals with certainties, whereas plausible inference deals with uncertainty?"

"Something like that."

"But thinking *always* involves uncertainty, right? If I already know the answer to a question, it would be better to describe that as 'remembering' rather than 'thinking', don't you agree?"

"Yes..."

"So thinking is not mathematical logic, but instead plausible inference?"

"The two are connected."

If you've ever tried to construct a mathematical proof, you've probably experienced this paradox first-hand: you're trying to establish mathematical certainty, but *before* you succeed, you have to seek the path through uncertainty, using a combination of knowledge, intuition, and trial and error. You obviously must understand the rules of mathematical logic, but by themselves they are not enough – because they provide no guidance to "plausible inference" under uncertainty.

1 Inference under Uncertainty

Most of us have learned about mathematical logic, theorems and proofs in math classes. But what exactly is "plausible inference"? To begin with, let's define it as *reasoning under uncertainty*, in other words any situation where we try to estimate the *probability* of some assertion without having a proof that is *true* (100% probability) or *false* (0% probability). Inference deals with the gray zone between these two extremes – the place where most real-world thinking takes place. Inference is the foundation of the scientific method – because science seeks knowledge in the real world, where uncertainty is unavoidable.

We all have an intuitive sense of inference – the ability to *do* it even if we don't understand *how* we're doing it. Ironically, in school we learn a lot about mathematical logic, and very little about the mathematics of inference. So you may be surprised to hear that mathematics provides a simple and powerful theory of inference, which you will learn before this chapter is done! The best way to learn an idea is to see how it solves problems. So let's tackle a problem from the "real world"... of game shows!

1.1 A Classic Problem: Monty Hall

Imagine you're a contestant on the Monty Hall Show. There are three doors, one concealing a valuable prize, and if you pick the right door, you win. But there's a twist: after you say which door you've picked, Monty always opens another door and shows you it's empty (no prize). Then he asks you whether you want to switch your choice to the other closed door. Should you? Does it make any difference?

Luckily, you get to ask your two friends Toby and Rajiv for advice:

- *Toby*: I don't see how it could make any difference. After all, the prize was already placed behind a certain door before you made your pick, and neither your initial choice nor the host's revealing an empty door cause it to magically move from one door to another. Each door had an equal chance of being right before you chose, and the prize hasn't moved, so that should still be the case. So the probability for each of the two remaining doors should be equal (1/2).
- *Rajiv*: I think it could make a difference, if the host's opening an empty door contains any *information* about where the prize is hidden. For example, if the host had *no* information about the prize's location, could he always follow the game's rules (i.e. show you an empty door, so he can ask you whether you want to switch)? No, by random chance, 1/3 of the time he would open the door that contains the prize. This suggests that under the stated rules of the game (Monty always opens an *empty* door), he must know where the prize is actually hidden. His action might communicate some of his information to you; the question is *how much*?

Naturally, your friends gave you opposite advice!

1.2 Bayes Law

This is a classic example of *inference*: you are trying to draw conclusions about something that is *hidden* (for convenience, let's call it H) using some things that were *observed* (which we'll refer to as O). Fortunately, inference boils down to one simple identity defining a "conditional probability", which we write in the form

$$p(H|O) \equiv \text{ probability that event } H \text{ occurs in the subset of cases where event } O \text{ indeed did occur}$$

which differs from the conventional definition of an *unconditional probability*:

$$p(H) \equiv \text{ total probability of event } H \text{ (irrespective of whether } O \text{ occurred)}$$

Using the intuitive concept of probability as the fraction of some set of events that meet a particular condition, and indicating "the count of events where H occurred" as $|H|$, we can immediately express the "joint probability" that *both* H and O occur (in the language of set theory we write this as the "intersection" $H \cap O$):

$$p(H \cap O) = \frac{|H \cap O|}{|S|} = \frac{|H \cap O|}{|O|} \frac{|O|}{|S|} = p(H|O)p(O)$$

(where S means "the set of all possible events"). Since (by symmetry) it is equally true that

$$p(H \cap O) = p(O|H)p(H)$$

then

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

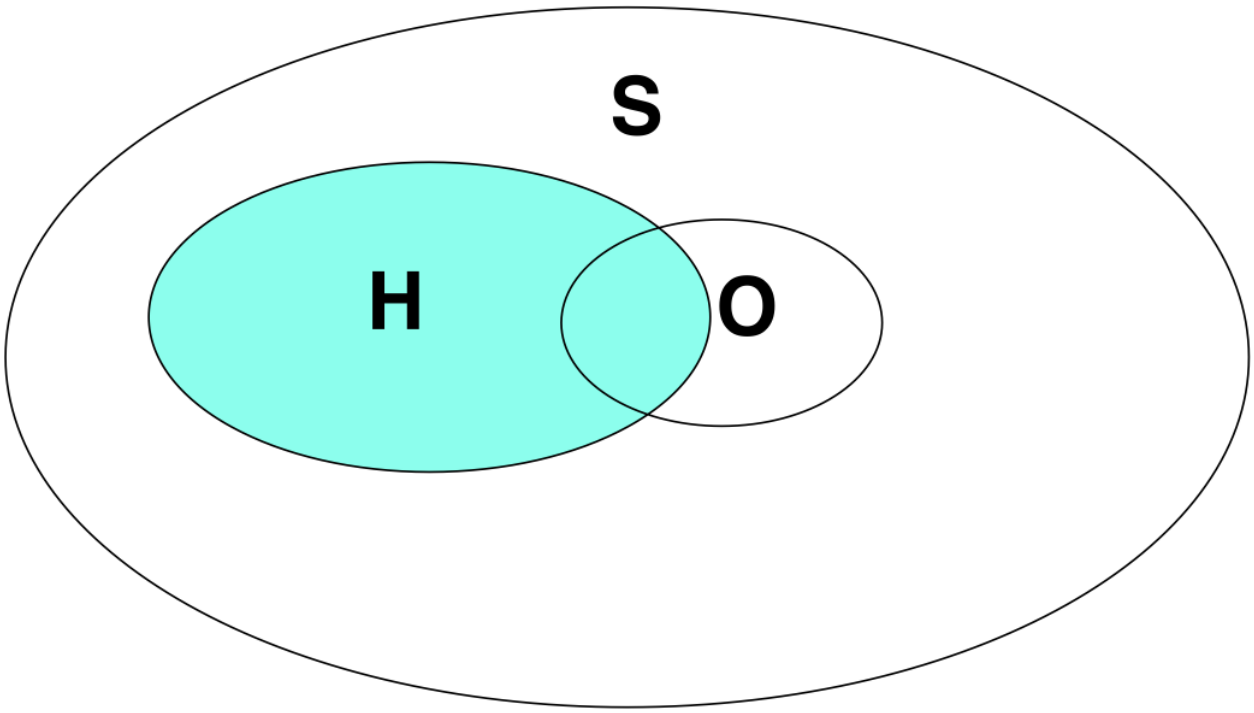

Figure 1: **A Venn diagram illustrating the conditional probability identity**

and

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

This is *Bayes Law*, and it is inference in a nutshell. It allows us to compute the probability of something *hidden* given that some observable event O has occurred, provided that we know the converse probability that observation O will occur *assuming* H has occurred.

We can immediately apply this to Monty Hall:

- *What is hidden*? Let's label the three doors A, B and C. Let's say you chose door A, and Monty opened door B to reveal that it was empty. Let's use the Greek letter δ for the hidden variable that indicates which door actually has the prize; i.e. $\delta = A$ means "door A has the prize". (In general in this text we will use Greek letters to designate hidden variables, on the principle that "it's Greek to me!").
- *What was actually observed*? Let's use a minus sign to indicate that a door was observed to be empty; i.e. B^- means "door B was observed to be empty".
- *What was the probability that your choice was correct, before Monty opened door* B? This is just the unconditional probability for the prize to be hidden behind any of the three doors (with equal probability), i.e.

$$p(\delta = A) = p(\delta = B) = p(\delta = C) = 1/3$$

- *How would the prize's location affect the probability of our observation*? This is where we perceive the "connection" between O and δ. Let's consider the three possible doors. If door A has the prize, Monty can choose either B or C to reveal as empty, with equal probability. So $p(B^-|\delta = A) = 1/2$. If door B had the prize, of course there's no way it could be observed as empty, so $p(B^-|\delta = B) = 0$. Finally, if door C had the prize,

Monty has no choice: he can't show A (you picked that one); he can't show C (it wouldn't be empty, and the game would be over), so $p(B^-|\delta = C) = 1$.

- *what is the total probability of our observation*? Either you picked the right door ($\delta = A$) or you didn't ($\delta \neq A$). In the first case, the two remaining doors are equally likely to be chosen by Monty to be revealed to be empty. In the second case, both of the remaining doors are equally likely to have the prize, or conversely, equally likely to be opened by Monty and shown as empty. Either way, Monty is equally likely to open either of the two doors B or C. Thus, intuitively, $p(B^-) = 1/2$. (We will show how to calculate $p(O)$ rigorously later).

So according to Bayes Law,

$$p(\delta = A|B^-) = \frac{p(B^-|\delta = A)p(\delta = A)}{p(B^-)} = \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

and

$$p(\delta = C|B^-) = \frac{p(B^-|\delta = C)p(\delta = C)}{p(B^-)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

So you are twice as likely to win by switching your choice to C, compared to staying with A! Monty's "observation" contains a lot of information about the prize location, and this is quantified directly by Bayes Law.

What can we learn by comparing this result to Toby and Rajiv's intuitive arguments?

- Toby's argument that "the probability of a past event (where the prize was placed) cannot change due to a *future* observation (when Monty later reveals door B to be empty)" is both wrong, and right. It's wrong in the sense that it mistakenly focuses on *causation* rather than *information*. That is, the question is not whether the future observation can somehow go back in time and alter the original probabilities, but simply whether it contains any useful information about the right answer. For example, say you and Monty pre-arrange to cheat by using code-phrases (e.g. Monty will say "Bob, do you want to change your choice?" if your original choice was correct, or "Robert, do you want to change your choice?" if it was wrong). This code-phrase is also an "observation" (in the same sense that showing door B to be empty is directly observable to you), and obviously it communicates to you perfect information for picking the right door. There is no metaphysical *how-can-that-be?* question here; it's merely a very "informative" observation.
- On the other hand, there is another sense in which the intuitive argument is absolutely correct. The *unconditional* probability (i.e. $p(\delta = A)$) *is* unchanged by the observation B^-; this is true by definition. It is the *conditional* probability $p(\delta = A|O)$ that is affected by O. On a fundamental level, the error of the intuitive argument is to consider probability to be a unitary, single concept ("$p(\delta = A)$ is what it is, regardless of anything else"), and failing to distinguish it from conditional probability. In other words, there is no point even talking about $p(\delta = A)$; once we observe B^- the probability that we are now talking about is a different probability, $p(\delta = A|B^-)$. It's correct that $p(\delta = A)$ is unchanged, but that's irrelevant to our question. The error is in thinking of the problem in terms of unconditional probability, rather than conditional probability.

2 What exactly do we mean by "Hidden" and "Observed"?

As you have probably already noticed, these questions require care in how we define words. Our goal is not to start a deep philosophical argument about what "thinking" means, but to choose definitions that capture the intuitive meaning in a form that different people will use consistently and unambiguously. To do this, we need to introduce a crucial concept that we will use frequently throughout this book.

2.1 Operational Definitions

An *operational definition* defines a variable or property in terms of an unambiguous, repeatable process that anyone can perform. This has several aspects:

- The operational definition must be *sufficient*. That is, following the steps specified in the operational definition must be all that is required to obtain the defined measurement or property.
- An operational definition must be *repeatable*. The utility of an operational definition would be lost if it were not repeatable by different persons, at different times.
- An operational definition need not be deterministic. That is, even when performed under apparently identical conditions, it need not yield the same measurement value. Operational definitions of random variables are perfectly valid.
- The purpose of an operational definition is to define a measurement or property in a way that is consistent and robust versus irrelevant sources of variation, e.g. who performs the measurement; whether they are tall or short; etc. The ideal operational definition would be insensitive to everything except the thing being measured. In reality, no operational definition can achieve this perfectly. However, our choice of which specific operational definition to use for a given kind of measurement (e.g. weight) should be guided by this principle. As we shall see later, the entire concept of an operational definition flows automatically from the simple goal of improving *prediction power*.
- An operational definition shifts the focus away from claims about what something *is*, towards a practical method for how to measure it.

2.2 Observed vs. Hidden Variables

We have defined inference intuitively as the *estimation of hidden states based on observed states*. It therefore behooves us to ensure that we have a clear, operational definition of "hidden" vs. "observed" states:

An *observed parameter* is an operationally defined measurement that yields a set of mutually exclusive states with no uncertainty. That is, the rules of the measurement process guarantee that each measurement yields exactly one of its possible states. Such a measurement by definition has no uncertainty, and therefore can be considered "observed". Any parameter that fails this definition is defined as "hidden". In plain English, *observed* just means "what you know" (facts you are already sure about), and anything that "you want to know" (implying that you do not yet know it with certainty) is *hidden*.

This distinction is so fundamental to the process of inference that we will enforce a strict rule on our nomenclature for hidden vs. observable variables:

- we will always write a hidden variable as a Greek letter, e.g. Θ.
- we will always write an observable variable as a Roman letter, e.g. X.

In this way we can always see at a glance whether a given conditional probability is a prediction from theory (e.g. $p(X|\Theta)$) or an inference from experimental data (e.g. $p(\Theta|X)$).

Quick Questions

1. For each of the following items, state one observable and one hidden parameter associated with it: a six-sided die; a photon; your friend Toby; a completed paper ballot; an email you received; a cryptographic signature. Explain precisely how each "observed parameter" fulfills the operational definition.
2. Make a list of 10 important observables and associated hidden variables in different areas of science, and try to state exactly what connects them, with reference to Bayes Law.

3 The Meaning of Bayes Law

Bayes Law is so central to inference that we will find ourselves referring to its various parts in nearly every sentence we utter when discussing an inference problem. So let's give each part a well-defined name, to eliminate any possible confusion about terms. Thus we can rewrite the symbolic form

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

in words:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Average Likelihood}}$$

- The *prior* $p(H)$ is our estimate of the probability of the hidden state H *before* we take any observations.
- The *posterior* $p(H|O)$ is our estimate of the probability of the hidden state H *after* we have observed event O.
- The *likelihood* $p(O|H)$ gives the probability of the observation *assuming* hidden state H.
- The *average likelihood* $p(O)$ is simply the total, unconditional, probability of observing O. This is sometimes also called the "marginal probability" of O.

3.1 Some Simple Likelihood Models for Bioinformatics

We now give two concrete examples of likelihood models commonly used in bioinformatics. In each case they give the conditional probability of an observed event count given a hidden parameter that reflects its true frequency.

The Binomial Distribution

Say a given observable variable has only two possible states, e.g. a gene either has a specific mutation a or it does not. If the probability of observing a on a given random draw is Θ, then the probability of observing a exactly m times in n independent draws is given by the *binomial distribution*

$$p(m|n, \Theta) = \binom{n}{m} \Theta^m (1-\Theta)^{n-m}$$

For example, since humans have two copies of each gene (one from each parent), the probability that you would have m copies with the a mutation is

$$p(m|n=2, \Theta) = \binom{2}{m} \Theta^m (1-\Theta)^{2-m}$$

which is just $\Theta^2, 2\Theta(1-\Theta), (1-\Theta)^2$ for $m = 0, 1, 2$ respectively.

The Poisson Distribution

Say an event occurs with uniform frequency μ over some continuous interval such as time or distance. Then the probability that the event happens exactly m times in an interval of length L is given by the Poisson distribution for

$\lambda = L\mu$:

$$p(m|\lambda) = e^{-\lambda}\frac{\lambda^m}{m!}$$

For example, if mutations occur uniformly over the genome at a density of μ mutations per kb, the probability of observing exactly m mutations in an interval of length L kb is given by the Poisson distribution with parameter $\lambda = L\mu$.

4 The Hidden Life of Conditional Probability

Popular culture regards what it calls "statistics" with mixed feelings bordering on schizophrenia. On the one hand, when a reporter or politician wants to sound knowledgeable about a subject, quoting probability numbers is the easiest way to impress. After all, probabilities are easy to remember (or fake – pick a number from 0 to 100!) and have no pesky units or magnitudes to forget (by contrast, swapping "million" and "billion" will tumble you instantly from expert to buffoon). The average listener may not be able to tell whether you have a deep understanding of political economy, but everyone will remember if you spouted numbers like a computer! On the other hand, ever since Mark Twain the old saw about "lies, damned lies, and statistics" has been the ready reply for anyone who wants to dispute such expertise. Many people have a strong feeling that probability numbers tell only the superficial story, and fail to capture the "deeper reality". And unfortunately, a lot of the time this suspicion is amply confirmed, by probability claims that do not mean what they appear to.

As students of statistics, we should wonder why our subject is often viewed as *worse* than "damned lies"! Presumably, Twain was referring to the double-barrelled problem of apparent objectivity combined with exceptional opportunities for deception. For my part, I think much of the problem is simple confusion about *conditional probability*, which most people have never heard of. Because it arises constantly in real-world problems, yet people lack the basic vocabulary for thinking about it, it is constantly misused, abused, and manipulated.

4.1 Why do we need "conditions"?

Consider the kinds of statements about probability we often hear in the media, such as "the probability of rain is 80%", or "The company's new AIDS diagnostic test is 97% accurate". These are *unconditional* probability statements. Even research articles in experimental science typically state probabilities in unconditional terms, and rarely give explicit conditional probability expressions. So it is logical to ask: why should we bother with conditional probability?

Table 1: A diagnostic disease test: 1000 patients were given a diagnostic test that gives either a positive () or negative () result, and independently assessed for whether they have the disease () or not () by rigorous clinical criteria.

	T^-	T^+	**total**
D^+	1	9	10
D^-	960	30	990
total	961	39	1000

To answer this, let's look at a simple example. A company reports that their new test for a disease is 97% accurate. The table above shows the raw data, which appear to support this claim. Among patients who do not have disease, the test gives the right answer 960/990=97% of the time, and among patients who have disease (a much rarer case), it gives the right answer 9/10=90% of the time.

There is just one catch here: these are not the conditional probabilities that a doctor (or patient) cares about! The whole point of the test result (T) is to give information about whether the patient has disease (D); i.e. we want to use *observed* variable T to learn about *hidden* variable D. Thus the probabilities above ($p(T^-|D^-)$ and $p(T^+|D^+)$) are irrelevant and useless. What we really care about is the converse, the probability that a patient has disease given a positive test result, $p(D^+|T^+)$. And there's the rub: $p(D^+|T^+) = 9/39 = 23\%$. More than three-quarters of the patients with positive test results do not actually have the disease!

This example illustrates several lessons:

- the "perfect lie": as this example shows, an unconditional probability statement can be both completely misleading and at the same time "factually correct"! The problem with an unconditional probability is that it doesn't tell you what conditions were used to obtain it. What assumptions (sensible or insane) gave rise to this number? You don't know. By choosing different conditions, I can obtain the number I want. As the example demonstrates, even within the strict limits of the correct data, freedom to pick our conditions gives us enough latitude to turn the conclusion upside down! The purpose of conditional probability is to make assumptions explicit.
- Strictly speaking, *every* probability calculation has at least some assumptions. So an unconditional probability statement is really a conditional probability travelling *incognito*–without telling you what its conditions were.
- It is crucial to distinguish the two possible directions of probability (posterior vs. likelihood). Usually what *matters* to us is the posterior (confidence about a hidden state that we care about). If someone instead bases their argument on quoting a likelihood ($p(O|H)$, the probability of an observation *assuming that we already know the true hidden state*), you should ask yourself whether this is fully justified, or whether this might be hiding important information.
- It is a fatal mistake to confuse one conditional probability with its converse (i.e. $p(X|Y)$ vs. $p(Y|X)$). They are quite different! Once you're aware of this distinction, you will find that people mix up converse probabilities all the time, sometimes due to poor thinking, and sometimes deceptively. When you listen to a politician, newspaper article, advertisement, or anyone else with "something to sell", see if you can catalog all the sins they commit against conditional probability. Remember that "97% test accuracy" may be completely irrelevant to the question that matters – especially if they don't even tell you what conditional probability it represents!
- Above all, we need conditional probability because it is the only way to connect hidden variables to observed variables. In practice, nearly everything we want to know is *hidden*. The only way we can learn anything about such hidden variables is by calculating their conditional probabilities based on our *observed* variables.

4.2 Conditional Probability as a Venn Diagram

Intersection vs. Conditioning: the Chain Rule

What exactly do we mean by "conditional probability"? We simply mean the *fraction of one set that intersects another set* (e.g. $p(dog|pet)$ means the subset of *pets* that are *dogs*). That means we can answer any question we have about conditional probability by drawing the Venn diagram that represents that conditional probability as an intersection of sets. For example, we draw a circle representing the set of *all pets*. Next we draw a circle representing *all dogs*. Some dogs are indeed pets, so the circles overlap. (Of course, there are also dogs who are not pets, and pets who are not dogs).

Let's see how to apply this simple set theory logic to computing conditional probabilities:

- First, let's emphasize that *unconditional probability* is fundamentally no different: $p(C)$ means the subset $S \cap C$ of a set S that satisfies some constraint C. The only difference is that we did not explicitly write down S as our condition, and thus the reader has no way to know what the set S is. For example, for $p(dog)$, is S "all pets" or "all animals", or what? We don't know.

- As an operational definition, we compute a probability as the fraction of one set within another:

$$p(C) \equiv \frac{|S \cap C|}{|S|}$$

  where $|S|$ means the "size of the set S".

- Substituting $A \cap B$ for C yields an expression for the "joint probability" of A and B:

$$p(A \cap B) \equiv \frac{|S \cap A \cap B|}{|S|}$$

- What if we want to express $p(A \cap B)$ in terms of $p(A)$? We can combine the two definitions above, yielding:

$$p(A \cap B) = \frac{|S \cap A \cap B|}{|S \cap A|} \frac{|S \cap A|}{|S|} = \frac{|S \cap A \cap B|}{|S \cap A|} p(A)$$

  This suggests a definition of conditional probability exactly in line with our verbal definition above:

$$p(B|A) \equiv \frac{p(A \cap B)}{p(A)} = \frac{|S \cap A \cap B|}{|S \cap A|}$$

  and the identity upon which Bayes Law is based:

$$p(A \cap B) = p(B|A)p(A)$$

- In a conditional probability expression $p(B|A)$, we distinguish B as the *subject* variable, and A as the *condition* variable.

- By itself, this equation may not seem like a powerful tool. But if you scrutinize it closely, you will see that this little seed can grow to handle any number of variables we want. Say we substitute in place of A a group of n variables $X_1 \cap X_2 \cap ... \cap X_n$. On the right-hand side we have a joint-probability of n variables, but on the left-hand side we have a joint-probability of $n + 1$ variables. So our expression for 2 variables gives us an expression for 3 variables, which gives us an expression for 4, etc. in a process of *induction*. For example, if we define $B = C \cap D$, then the above identity becomes

$$p(A \cap C \cap D) = p(A|C \cap D)p(C \cap D) = p(A|C \cap D)p(C|D)p(D)$$

  Note: when dealing with arbitrary numbers of variables, it is often convenient to leave the intersection symbol $\cap$ implicit whenever that does not create ambiguity, and just replace it by a comma-separated list of variables, so that "$p(A, B)$" really means $p(A \cap B)$.

- Generalizing this induction, we obtain the *conditional probability chain rule*

$$p(X_1, X_2, ...X_n) = \prod_{i=1}^{n} p(X_i|X_1, X_2...X_{i-1})$$

  which is our main powertool for manipulating conditional probability expressions.

- Since the joint probability $p(X_1, X_2, ...X_n)$ is symmetric to swapping any pair of variables X_i, X_j, this chain rule can be applied in $n!$ possible orders, e.g.

$$p(X_1, X_2, ...X_n) = p(X_1)p(X_2|X_1)p(X_3|X_1, X_2)...$$

but equally well

$$p(X_1, X_2, ...X_n) = p(X_n)p(X_{n-1}|X_n)p(X_{n-2}|X_{n-1}, X_n)...$$

Note that this is general: it applies to any joint probability.

- So far, we've talked about "states" s and "variables" X interchangeably. By "state" we simply mean a subset of the total set S. By "variable" we mean a specific way of dividing S into non-overlapping subsets, i.e. mutually exclusive states, each of which is assigned a unique label. In general in this text we will symbolize a variable with an uppercase letter X, and possible "values" of that variable (i.e. the labels of its allowed states) as lowercase letters e.g. $x_1, x_2, ...$ Thus $p(x_1)$ is a number (the probability of state x_1), whereas $p(X)$ implies the function that maps every possible value of X to its corresponding probability value. By "mutually exclusive", we mean that the intersection of any two distinct states of X is empty, i.e. $x_i \cap x_j = \emptyset$. Another way of saying this is that x_i and x_j are *disjoint*. Note also that the union of all possible states of a variable equals the entire set S.
- This also permits us to define a *random variable*, as simply a variable whose value is sampled randomly in the following way. We randomly select a member of the set S, and assign X the value (i.e. label) of the state that contains that member, which is guaranteed to be unique.
- Note that multiple variables X, Y, Z simply mean *different* ways of subdividing S. Each time we sample S we obtain a value of *all* the variables $X, Y, Z, ...$ that we have defined on S. Depending on how we choose these subdivisions, there can be a lot, a little, or no correlation between two random variables. (Obviously, if we chose the same set of states for both variables, they would be perfectly correlated).

Union and Projection Operations

Another important conditional probability operation is the union of several conditional probabilities, which follows from the equivalent set operation:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

where $p(A \cap B)$ takes into account the overlap between S_A and S_B. We define two subsets A and B as *disjoint* iff $p(A \cap B) = 0$ (i.e. their intersection is a null set), in which case $p(A \cup B) = p(A) + p(B)$. We can also apply the union operation on condition variables:

$$p(A|C_1 \cup C_2) = \frac{p(A \cap (C_1 \cup C_2))}{p(C_1 \cup C_2)} = \frac{p(A \cap C_1) + p(A \cap C_2) - p(A \cap C_1 \cap C_2)}{p(C_1) + p(C_2) - p(C_1 \cap C_2)}$$

Thus, for a set of disjoint constraints $b_1, b_2, ...b_n$ that sum to the entire set S, i.e.

$$b_1 \cup b_2 \cup ... \cup b_n = S$$

and

$$b_i \cap b_j = \emptyset | \forall i \neq j$$

we have

$$\sum_{i=1}^{n} p(A \cap b_i) = p(A)$$

This enables us to eliminate a subject variable B from a probability term by summing over all possible disjoint values of B. This is expressed in functional terms as

$$p(A, \cdot) = \sum_{B} p(A \cap B)$$

for a discrete variable B, or

$$p(A, \cdot) = \int p(A \cap B) dB$$

for a continuous variable B. $p(A \cap B)$ (or, equivalently, $p(A, B)$) is commonly referred to as the "joint probability of A and B", and $p(A, \cdot)$ (or, equivalently $p(A)$) is called the "marginal probability of A", because it is typically written in the margin of a joint probability table, as the sum of a given column or row.

This operation is a *projection*; that is, it represents the projection of a multidimensional function ($p(A, B)$) onto a single dimension (A). As we will see in the next chapter, this projection operation is of fundamental importance throughout information theory. From this projection operation we can see that there is a basic asymmetry in the relationship between the joint vs. marginal probability functions. We can derive the marginal from the joint probability ($p(A, B) \rightarrow p(A, \cdot)$) but not vice versa; we cannot obtain the joint probability from the marginal probabilities $p(A), p(B)$. Thus, projection "destroys information" in the sense that the marginals $p(A), p(B)$ contain only a subset of the information present in $p(A, B)$. (Note: computing the joint probability from the *conditional probability* (i.e. $p(A, \cdot), p(B|A) \rightarrow p(A, B)$) is another matter; the conditional probability can be considered to contain the "same information" as the joint probability.) We will analyze this information loss rigorously in Chapter 3. This asymmetry will turn out to be the basis for defining the *mutual information* shared between two random variables A and B.