Basic Probabilities Primer for DIG5876

R. Paul Wiegand

March 21, 2013

1 Introduction

Comprehensive discussion of probabilities is well outside the scope of this course; however, those students who aren't familiar with at least some basics regarding probabilities will face some challenges. This document is meant to address this gap by providing an easy-to-read primer of a handful of very basic topics in probability leading up to Bayes rule. It's meant for those who haven't been exposed to much probability and those who would like a quick refresher after having been away from it for a while. Those comfortable with the topics discussed in the document are free to skim it or omit reading it entirely.

While a comprehensive and extensive knowledge of probability is not strictly necessary for this course, students who are interested in pursuing in-depth study of modern machine learning techniques would do well to have had a course that covers probabilities. In lieu of such a course, self-study using any number of books on the subject would be useful. I suggest the freely available text by Grimstead and Snell, *Introduction to Probability*.

1.1 Some Terms: Experiments, Outcomes, & Events (Oh My!)

It's helpful to start by being clear with some common terminology used when discussing probability. Perhaps the four most basic terms are *experiment*, *outcome*, *events*, and *probability* itself. Let's deal with the first three first since they are the simplest.

- *Experiment* – A scenario that involves chance and produces outcomes
- *Outcomes* – The potential results of an experimental trial
- *Event* – One or more specific outcome of a particular experiment

Experiment 1 Flip a Coin. *Since the coin may land head-side up or tail-side up (assuming it cannot land any other way), the outcomes of that experiment are* heads *and* tails. *If I flip the coin and it lands with the head-side up, then I have a* heads *event.*

Experiment 2 Roll a Die. *The outcomes of rolling a fair, six-sided die are* $1, 2, 3, 4, 5,$ *and* 6. *Getting a* 6 *on a particular roll of the die is an event.*

1.2 Views of Probability

There are many views of what *probability* really is, ranging from slight philosophical distinctions to differences that have practical implications for the mathematics. By and large, these differences will have little impact on our discussions in this class, but it's still worth considering several common views of probability:

- *Frequency* – Probability is the relative frequency with which events occur in the long run;
- *Subjective* – Probability reflects a subjective degree of belief. For example, what we're willing to bet on given what we know;
- *Propensity* – Probability reflects an inherent uncertainty in some experiment or process;
- *Logical* – Probability is an extension of formal logic that incorporates rational / objective degree of belief.

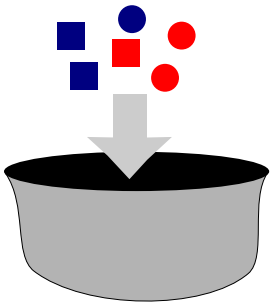
In many cases, these views are fairly consistent with one another; however, there are some interesting philosophical differences. For example, what is the probability of the outcome of an experiment that can only ever be performed one time? A *frequentist* might wrestle with this quite differently from someone with a *propensity* view. Another philosophical distinction is exposed when you consider the following notion. Suppose I flip a coin and observe the event that it lands heads-side up. Further, suppose you *observe* me flip the coin but do not know the result of the flip. What is the probability of the *heads* event? Is it fair to say that *to me*, it is 1.0, but *to you* it is 0.5? Is probability objective or subjective? Etc.

The philosophy of probability is all fun and games until someone might have lost a fractional body part that we are reasonably certain was an eye, but for this document let's take a *frequentist* approach and wash our hands of the philosophy. Taking the above examples and a frequentist approach, we simply enumerate the possible outcomes and compute probability by determining how frequently we expect them to occur in the future. In both of the coin-flipping and die-throwing examples above, if the coin and die are fair then we don't expect any outcome to appear with more frequency than another. In the first case, the probability of the event *heads* is the same as the probability of the event *tails* — since there are two outcomes, each occurs with probability $\frac{1}{2}$. Likewise, with a fair die, each number occurs with probability $\frac{1}{6}$.

For reference, I use the notation $\Pr\{A\}$ to mean the "*probability of event* A". Sometimes it's important to make the experiment and result clear in the notation. For example, in class I may sometimes write the "*probability that the outcome of experimental trial* X *is* 6" as $\Pr\{X = 6\}$. This can also be read is the "*probability that random variable* X *is 6*", but don't worry about what a *random variable* is just yet.

2 Fun with Shapes

Coin-flipping and dice-throwing are simple enough when we're discussing only single throws, or when the outcomes are very simply described. In many cases, though, we need to talk about more complicated combinations of different kinds of outcomes. Let's setup a simple running example that we can use for discussion.



Suppose we place six shapes in a basket, three squares and three circles. Suppose further than two of three squares are blue, while the third is red and that the colors of the circles are precisely the opposite. Our experiments will involve drawing a shape from the basket (with replacement — meaning we'll always throw the shape back in when we're done and shake the basket up again). With this, we can clearly elicit some obvious probabilities:

- What is the probability that a drawn shape is red? $\Pr\{red\} = \frac{1}{2}$
- What is the probability that a drawn shape is blue? $\Pr\{blue\} = \frac{1}{2}$
- What is the probability that a drawn shape is a circle? $\Pr\{circle\} = \frac{1}{2}$
- What is the probability that a drawn shape is a square? $\Pr\{square\} = \frac{1}{2}$

2.1 Joint Probability vs. Conditional Probability

Among the easiest things to get confused about in probability is the difference between a *joint probability* and a *conditional probability*. The way I like to make the difference clear in my mind is by trying to remember that the joint probability is considering joint events over some outcome space, while the conditional probability is considering events within a *subset* of the outcome space (the subset that meet some "*condition*"). This often translates to differences in how some denominator is expressed, as we'll see here.

The probability for joint events A and B is sometimes written $\Pr\{A \wedge B\}$, and sometimes $\Pr\{A, B\}$; I will use the latter. In the shapes example $\Pr\{square, blue\}$ refers to the "*probability of drawing a shape that is both a square and blue*". Continuing with our example:

- $\Pr\{square, blue\} = \frac{2}{6} = \frac{1}{3}$
- $\Pr\{circle, blue\} = \frac{1}{6}$
- $\Pr\{square, red\} = \frac{1}{6}$
- $\Pr\{circle, red\} = \frac{2}{6} = \frac{1}{3}$

Notice that the denominator was always six (before we reduce the fraction) because there are six possible outcomes in the set.

For conditional probability it's typical to use notation such as $\Pr\{A|B\}$, translating the funny vertical bar as "*given*". So we would read $\Pr\{square|blue\}$ as the "*probability that we draw a square, given that we are told beforehand the shape is blue*". Again, from our example:

- $\Pr\{square|blue\} = \frac{2}{3}$
- $\Pr\{circle|blue\} = \frac{1}{3}$
- $\Pr\{square|red\} = \frac{1}{3}$
- $\Pr\{circle|red\} = \frac{2}{3}$

Here the denominator is always 3 because in this case, whether red or blue, if we first select a color then there are three shapes remaining.

2.2 The Fundamental Rule

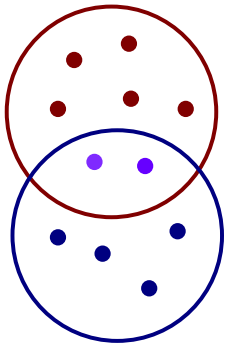
Joint probabilities and conditional probabilities are related by the *fundamental rule* (sometimes called the *product rule*). The fundamental rule says that the probability of two events occurring together is the probability of one event *given* that the other is true times the probability that the other is true:

$$\Pr\{A, B\} = \Pr\{A|B\} \cdot \Pr\{B\}$$

It's easy to get some intuition for this based on our shapes example. Suppose I *first* select the color and remove all shapes not of those colors from the basket, *then* I select the shape. The chance of getting *blue* is $\frac{1}{2}$, but if I *had* chosen blue the chance of getting a square from the subsequent draw would be $\frac{2}{3}$. This is the same as if I'd just selected a shape and asked what is the probability that it is both a *square* and *blue*:

$$\begin{aligned} \Pr\{square, blue\} &= \Pr\{square|blue\} \cdot \Pr\{blue\} \\ \frac{2}{6} &= \left[\frac{2}{3} \cdot \frac{1}{2}\right] = \frac{1}{3} \end{aligned}$$

Indeed, we can use this relationship to help understand the notion of *independence*. Event A is *independent* of event B if conditioning on B doesn't change the probability — that is, if you first remove all non-B events then the probability for A doesn't change. Formally, we say that A is independent of B iff $\Pr\{A|B\} = \Pr\{A\}$. It makes intuitive sense to think about it this way, but it's more common to deal with independence in terms of *joint probability*: When A is independent of B, the $\Pr\{A, B\} = \Pr\{A\} \cdot \Pr\{B\}$. But with the fundamental rule, we can see why this is the same thing now!



$$\Pr\{A, B\} = \Pr\{A|B\} \cdot \Pr\{B\} = \Pr\{A\} \cdot \Pr\{B\}$$

Notice that, in general if we want the probability that we get event A *or* event B, we cannot simply add $\Pr\{A\}$ and $\Pr\{B\}$ since any outcomes they have in common will be double-counted. So we must *subtract* the joint probability: $\Pr\{A \vee B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{A, B\}$. When A is independent of B, we can write this as: $\Pr\{A\} + \Pr\{B\} - \Pr\{A\} \Pr\{B\}$.

2.3 Marginal Probabilities

In the coming chapters, you'll also see terms like *marginal probability*, *prior probability*, or *posterior probability*, as well. We'll discuss the latter two below, but first let's get a handle on what a marginal probability is. It's easier to first explain *marginalization*, sometimes called the *summation rule*. The term comes from the idea that one might enumerate all the possible values of something on (say) some kind of ledger, writing the probabilities for each thing in the "margin" of the ledger, then add them all up.

Suppose we have some event X that may result in many outcomes, including x, and we have another event Y that can take on values $y_1, y_2, \ldots y_n$. Then one way of getting the probability of the event x is to *add up* all the *joint* probabilities of x and each y_i. Because of the fundamental rule, we can get a similar result using conditional probabilities.

$$\Pr\{x\} = \sum_{i=1}^{n} \Pr\{x, y_i\} = \sum_{i=1}^{n} \Pr\{x|y_i\} \Pr\{y_i\}$$

In this case, we are *marginalizing* over the outcomes for Y to get the *marginal probability* $\Pr\{x\}$. From our shape example:

$$\begin{aligned} \Pr\{square\} &= \Pr\{square, red\} + \Pr\{square, blue\} \\ &= \Pr\{square|red\} \Pr\{red\} + \Pr\{square|blue\} \Pr\{blue\} \\ \frac{1}{2} &= \left[\frac{1}{3} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{1}{2}\right] = \left[\frac{1}{6} + \frac{2}{6} = \frac{3}{6}\right] = \frac{1}{2} \end{aligned}$$

3 Bayes Rule

Notice something interesting about the fundamental rule: Since the *order* doesn't matter when computing a joint probability, the fundamental rule really gives us *two* expressions:

$$\begin{aligned} \Pr\{A, B\} &= \Pr\{A|B\} \Pr\{B\} \\ \Pr\{A, B\} &= \Pr\{B|A\} \Pr\{A\} \end{aligned}$$

Noting both of these expressions, it doesn't take much algebra to realize the fact of the enormously useful equation known as *Bayes Rule*[1]:

$$\Pr\{A|B\} = \frac{\Pr\{B|A\} \Pr\{A\}}{\Pr\{B\}}$$

Bayes essentially gives us a way to flip our conditional probabilities around when we don't have everything we might want. The basic idea of Bayes is easy to remember, but students sometimes have a hard time remembering which term goes where in the expression. If you get confused, just remember that you can easily work it out from the fundamental rule.

[1]Note that many mathematicians actually accredit Laplace with determining this in spite of the popular use of the phrase "*Bayes Rule*".

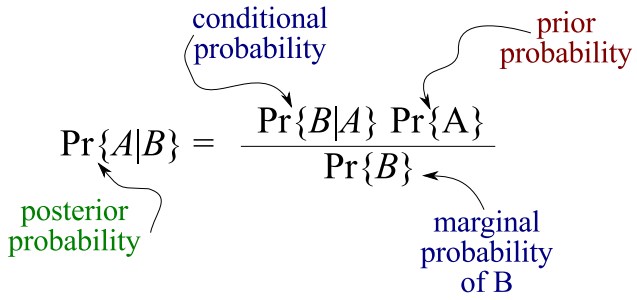
3.1 Bayes Anatomy

Often when Bayes is applied (particularly in machine learning), we are *updating* some probability after having been given some new piece of information — that is, we have some estimate for $\Pr\{A\}$, then we're given some new information, and we'd like to see how this affects the probability. Perhaps the following very slight reformulation will make the intuition for this clearer:

$$\Pr\{A \text{ after new knowledge}\} = \Pr\{A\} \cdot (\text{some update factor based on knowledge})$$

$$\Pr\{A|B\} = \Pr\{A\} \cdot \left(\frac{\Pr\{B|A\}}{\Pr\{B\}}\right) = \frac{\Pr\{B|A\}\Pr\{A\}}{\Pr\{B\}}$$

In that sense, we can think of $\Pr\{A\}$ as the *prior probability* for A (what we knew about A *before* we learn B), and $\Pr\{A|B\}$ as the *posterior probability* of A (what we know about A *after* we learn B). Here is a diagram to bring it all together:



Sometimes we don't know $\Pr\{B\}$ directly. But if we can calculate $\Pr\{B|a_i\} \cdot \Pr\{a_i\}$ for all a_i, then we can use marginalization to help us:

$$\Pr\{a_j|B\} = \frac{\Pr\{B|a_j\}\Pr\{a_j\}}{\sum_{a_i \in A} \Pr\{B|a_i\}\Pr\{a_i\}}$$

4 Confusion & Probabilities

The ideas and mathematics behind marginal and conditional probabilities are simple, but people don't always think naturally about probabilities. As a result of our (perhaps) poor intuition, it's sometimes easy to jump to the wrong conclusions. This section provides several examples where the most obvious / common answer tends to be the wrong answer, precisely because of our counter-intuition. See if you can work out why the results are the way they are.

4.1 Bertrand's Boxes

Suppose there are three boxes, each with two coins in them. One box has two gold coins, one has two silver coins, and the third has one of each. Suppose further that I select a box uniformly at random (i.e., each of the three boxes have the same probability of being selected) and select a coin uniformly at random from the box without looking at the other coin.

Question: If the first coin is gold, what is the probability that the *other* coin in that box is gold?

Answer: The probability is $\frac{2}{3}$.

4.2 Monty Hall

On a game show, you are presented with three doors, behind one of which is an iPad and behind the other two are rolled up newspapers from two years ago. You are asked to pick one of the doors but before the door is opened one of the *other* two doors is opened to reveal a rolled up newspaper. You are given the option to keep your choice or switch to the remaining unopened door. Whichever door you open, you win the prize behind it.

Question: Is the probability of winning the iPad higher, lower, or the same, if you choose to switch versus if you stick with your original choice?

Answer: It is better to switch. The probability of winning if you switch is $\frac{2}{3}$.

Hint: This is essentially the same question as *Bertrand's Boxes*.

4.3 Am I Diseased?

You are given a test for a disease that shows a positive result. You know that approximately 8 in 1,000 people get the disease, and you know that the test is pretty accurate: Only about 2 in 100 patients with the disease test negatively, and only about 3 in 100 patients without the disease test positively.

Question: Is it more likely than not that you have the disease?

Answer: No. Knowing the test results, the probability that you have the disease is still only about 0.21.

Hint 1: You are told Pr $\{testresult|diseasestate\}$, but you really want know Pr $\{diseasestate|testresult\}$. Do you see why?

Hint 2: Remember the marginalization we did at the bottom of the previous page.