PLSC 502: "Statistical Methods for Political Research"

Probability
September 20, 2016

Probability

Probability can be thought of as a particular kind of function (mapping), where the domain is a set of events or values of a variable, and the range is a set of values on the interval $[0, 1]$. Gill (2006) goes into a bit more detail about the broad mathematics behind probability; for now, let's concentrate on some simple examples.

Imagine observing the occurrence of some process. Formally, we refer to each possible event as an *outcome*, the result of a particular observation of a process as an *realization*, and the set of all possible events (outcomes) as the *sample space* (often denoted S). I'll generically refer to the variable in question as X, and a given outcome as x; I'll also refer to the number of possible outcomes as J, so that the sample space has J elements (indexed by j, with $J \in [0, \infty]$). Thus, generically, we have:

$$X \in S = \{x_1, x_2, ...x_J\} \tag{1}$$

In week one of the 2016 NFL season, for example, the St. Louis Rams might have scored any integer-valued number of points from zero to infinity (the sample space); they actually scored 0 points (the realization). So, the sample space for that variable X could be written:

$$X \in S = \{0, 1, 2, ...\}$$

and the realization is just

$$X_{\text{Rams}} = 0.$$

Like random variables, sample spaces can be *discrete* or *continuous*: while a phenomenon like the occurrence of a lightning strike or the number of points scored is an inherently discrete phenomenon, other things of interest (time, temperatures, etc.) are continuous.

Now, consider the *probability* of some outcome in the sample space. We'll talk about two general theoretical perspectives on probability: *frequentist* and *Bayesian*. *Frequentist* perspectives on probability view probability as a long-run relative frequency. That is, if we gather repeated realizations of the process under study, the fraction of times a particular outcome in the sample space is observed (as the number of repeated realizations grows without limit) is the probability of that outcome. Intuitively, this suggests that the basic probability of an event is equal to:

$$\Pr(\text{Event}) = \frac{\text{The number of times the \textit{event of interest} can or could occur}}{\text{The number of times \textit{any event} can or could occur}}.$$

More formally, we can write this as:

$$\Pr(X = x) = \lim_{N\to\infty} \left(\frac{\sum_N I\{X_i = x\}}{N} \right) \tag{2}$$

where $I\{\cdot\}$ is an *indicator function* for $X_i = x$ and N defines the number of realizations (often termed "trials").

From this definition, several important *characteristics* of probabilities follow that you'll need to remember:

1. Probabilities necessarily **range between zero and one**:

   $$\Pr(X = x) \in [0, 1]. \tag{3}$$

   [Think about Eq. (2), and you'll see why...].

2. The **sum of probabilities for all outcomes always equals one**:

   $$\sum_{j=1}^{J} \Pr(X = x_j) \equiv \Pr(S) = 1.0 \tag{4}$$

   That is, *something* is going to happen.

3. The **Multiplication Rule**:

   > The probability of obtaining a *combination* of independent, mutually exclusive outcomes (say, x_j and x_ℓ) is equal to the *product* of their separate probabilities.

   - That is, $\Pr(X = x_j \cap X = x_\ell) = \Pr(X = x_j) \times \Pr(X = x_\ell)$, $j \neq \ell$
   - Think about this rule in terms of the word "and"... $\Pr(X = x_j \ \underline{\text{and}}\ X = x_\ell)$.

4. The **Addition Rule**:

   > The probability of obtaining *any one* (or more) of several independent, mutually exclusive outcomes is equal to the *sum* of the probabilities for those events.

   - That is, $\Pr(X = x_j \cup X = x_\ell) = \Pr(X = x_j) + \Pr(X = x_\ell)$.
   - The addition rule also implies the rule (above) that the sum of all probabilities equals one.
   - Think about this rule in terms of the word "or"... that is, $\Pr(X = x_j \ \underline{\text{or}}\ X = x_\ell)$.

- Note as well that if the events are not mutually exclusive, we write:

  $$\Pr(X = x_j \cup X = x_\ell) = \Pr(X = x_j) + \Pr(X = x_\ell) - \Pr(X = x_j \cap X = x_\ell)$$

  This prevents "double counting" of events that "overlap." So, for example, if we want to know the probability of drawing (from a standard 52-card deck with no jokers) either a diamond card *or* a face-card, it would be equal to:

  $$\begin{aligned}
  \Pr(Z) &= \Pr(\text{Diamond}) + \Pr(\text{Face-Card}) - \Pr(\text{Diamond-Suited Face Card}) \\
  &= \frac{1}{4} + \frac{12}{52} - \frac{3}{52} \\
  &= 0.25 + 0.23 - 0.06 \\
  &= \mathbf{0.42}
  \end{aligned}$$

Independence and Conditional Probabilities

Next, consider two outcomes x_j and x_ℓ. We said at the beginning that the probability of two independent events both happening is equal to the product of their individual probabilities, provided that they were *independent* and *mutually exclusive.* In fact, we normally use the former to define the latter, rather than the other way around.

Consider the probability function $\Pr(X = x_j, X = x_\ell) = \Pr(X = x_j \ \cap \ X = x_\ell)$.

- We'll later call this a joint probability density function ("joint PDF").
- The two probabilities of which it is composed, $\Pr(X = x_j)$ and $\Pr(X = x_\ell)$, are the *marginal* PDFs.

Now suppose that x_j and x_ℓ are *independent...*

- Then, by our rule above, $\Pr(X = x_j, X = x_\ell) = \Pr(X = x_j) \times \Pr(X = x_\ell)$.
- That is, the joint PDF is equal to the product of the marginal PDFs...
- E.g. two dice: $\text{Prob}(1, 1) = \text{Prob}(1) \times \text{Prob}(1)$.

If the two events are not independent, then we need to consider their *conditional* probabilities.

- I.e., the probability of x_j given x_ℓ.
- We write this as $\Pr(X = x_j | X = x_\ell)$, or conversely as $\Pr(X = x_\ell | X = x_j)$.
- We say "The probability of x_j given x_ℓ," etc.

This leads us to the ***rule of conditional probability***:

$$\begin{aligned}\Pr(X = x_j | X = x_\ell) &= \frac{\Pr(X = x_j, X = x_\ell)}{\Pr(X = x_\ell)}, \text{ and} \\ \Pr(X = x_\ell | X = x_j) &= \frac{\Pr(X = x_j, X = x_\ell)}{\Pr(X = x_j)}\end{aligned}$$

What does this mean?

- If two variables are *independent*, it means that its conditional probability vis-à-vis the other is the same as its marginal probability.
- That is,

$$\begin{aligned}\Pr(X = x_j | X = x_\ell) &= \frac{\Pr(X = x_j, X = x_\ell)}{\Pr(X = x_\ell)} \\ &= \frac{\Pr(X = x_j) \times \Pr(X = x_\ell)}{\Pr(X = x_\ell)} \\ &= \Pr(X = x_j)\end{aligned}$$

In other words, the probability of x_j doesn't depend on the value of x_ℓ...

On the other hand, if the two are not independent, we need to consider their joint probabilities. In addition, Gill notes that *any* joint probability can be decomposed into a product of conditional probabilities, using the multiplication rule. So, for example, if x_j, x_ℓ, and x_k are not independent, we can write their joint probability as:

$$\Pr(X = x_j, X = x_\ell, X = x_k) = \Pr(X = x_j | X = x_\ell, X = x_k) \times \Pr(X = x_\ell | X = x_k) \times \Pr(X = x_k)$$

Bayes' Rule

The idea of conditional probabilities implies a number of things. Perhaps most usefully, it provides us with a mechanism for integrating information into our probability assessments. In particular, consider two outcomes x_j and x_ℓ. We know that

$$\Pr(X = x_j | X = x_\ell) = \frac{\Pr(X = x_j, X = x_\ell)}{\Pr(X = x_\ell)}$$

and

$$\Pr(X = x_\ell | X = x_j) = \frac{\Pr(X = x_\ell, X = x_j)}{\Pr(X = x_j)}$$

Since we know that $\Pr(X = x_j, X = x_\ell) = \Pr(X = x_\ell, X = x_j)$, we can write:

$$\Pr(X = x_j|X = x_\ell) \times \Pr(X = x_\ell) \quad = \quad \Pr(X = x_\ell|X = x_j) \times \Pr(X = x_j)$$

which in turn allows us to state that:

$$\Pr(X = x_j|X = x_\ell) = \frac{\Pr(X = x_\ell|X = x_j) \times \Pr(X = x_j)}{\Pr(X = x_\ell)}$$

This is **Bayes' Rule** (Gill calls it "Bayes Law"). In this setup,

- $\Pr(X = x_j)$ is often referred to as the "prior" probability of x_j – prior in the sense that it doesn't take into account any information about x_ℓ.
- $\Pr(X = x_j|X = x_\ell)$ is the *posterior probability* of x_j – the one that takes into account information about x_ℓ.
- $\Pr(X = x_\ell|X = x_j)$ is the conditional probability of x_ℓ.

All of this will become important later, when we discuss (among other things) estimation...

Probability and Odds

We can express the probability of an event in terms of its *odds*. In particular, if the probability of an event x_j is $\Pr(X = x_j)$, then the odds of x_j is just the ratio:

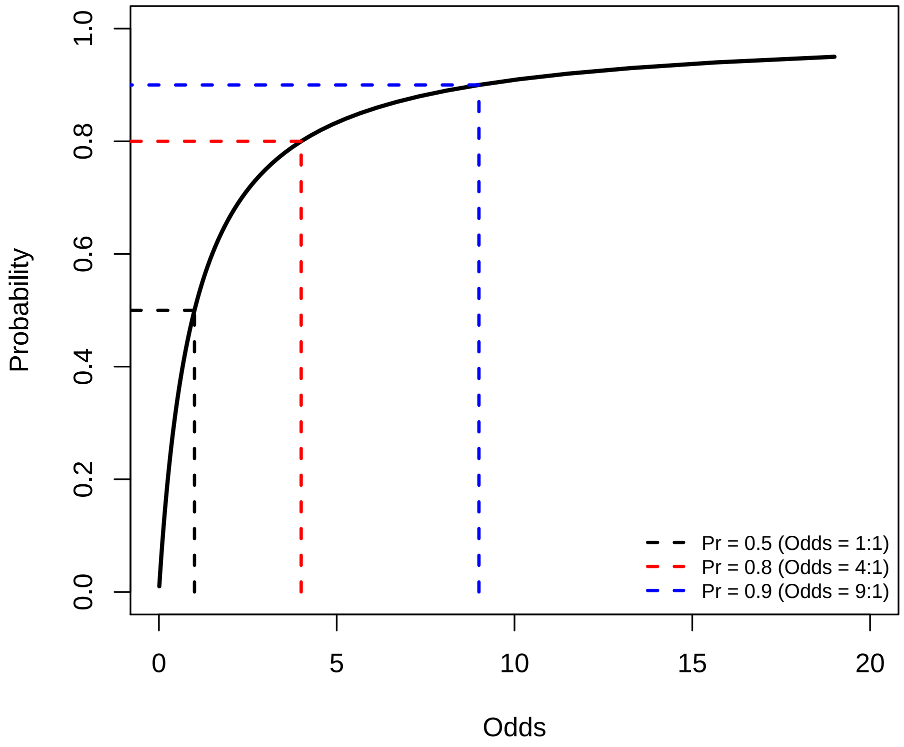
$$\begin{aligned} \text{Odds}(X = x_j) &= \frac{\Pr(X = x_j)}{\Pr(X \neq x_j)} \\ &= \frac{\Pr(X = x_j)}{1 - \Pr(X = x_j)} \end{aligned} \tag{5}$$

Note that, unlike probabilities, odds are unbounded from above (though they are still bounded at zero from below). The relationship between the odds of an event and its probability are shown in Figure 1.

Odds are often denoted with a colon: "The odds of x_j are 4:1 (in favor)." This means that:

- $\Pr(X = x_j) = \frac{4}{4+1} = 0.8$,
- $\Pr(X \neq x_j) = \frac{1}{4+1} = 0.2$

Figure 1: Probability and Odds

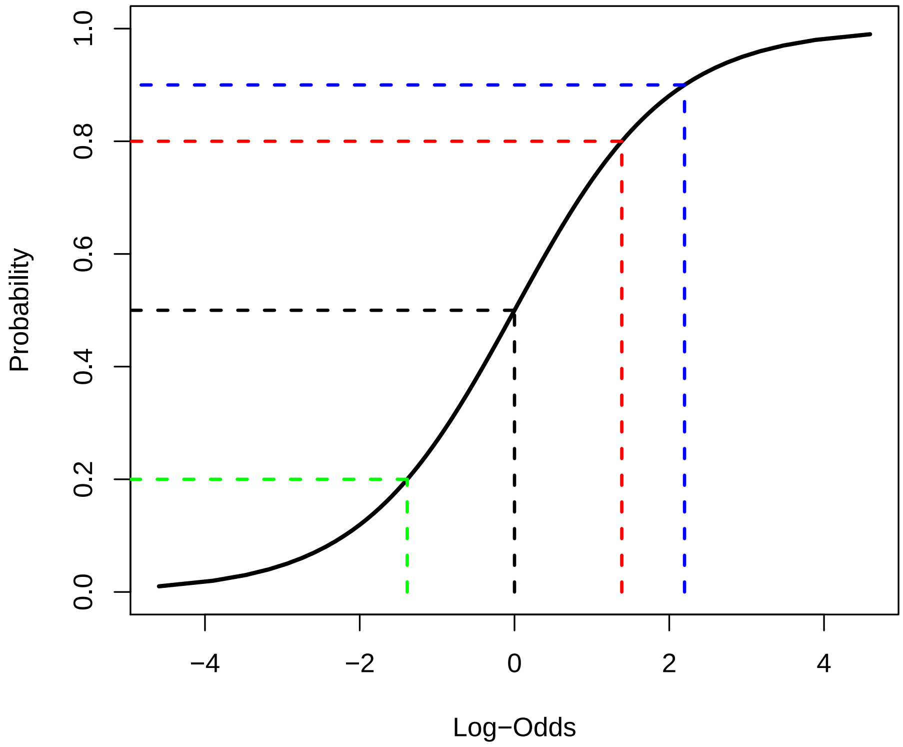


At times, odds can be a more intuitive way of thinking about probabilities than are raw probabilities themselves. Less intuitive – but also useful, for reasons we'll go into a bit more later – is the *log-odds*, the natural logarithm of the odds:

$$
\begin{aligned}
\ln[\text{Odds}(X = x_j)] &= \ln\left[\frac{\Pr(X = x_j)}{\Pr(X \neq x_j)}\right] \\
&= \ln\left[\frac{\Pr(X = x_j)}{1 - \Pr(X = x_j)}\right] \qquad (6)
\end{aligned}
$$

Unlike the odds (which range in $[0, \infty)$), the log-odds are unbounded (that is, they have a range of $(-\infty, \infty)$); this turns out to be mathematically useful in a number of circumstances. At the same time, log-odds are not very intuitive: if I told you that the log-odds of a horse winning a race were -2.4, that wouldn't tell you much (but if I said it was a 10:1 shot, you'd get it...).

Figure 2: Probability and Log-Odds



Likelihood

Finally, consider a bunch (say, N) realizations of some variable X. Data on X take the form of N separate realizations of outcomes in the sample space:

$$\begin{array}{rcl} X_1 & = & x_1 \\ X_2 & = & x_2 \\ X_3 & = & x_3 \\ \vdots & & \vdots \\ X_N & = & x_N \end{array}$$

Here x_i denotes the actual value of X observed for each data point. A useful quantity is the *likelihood*: the joint probability of all realizations of X in the data:

$$L(X) = \Pr(X_1 = x_1, X_2 = x_2, ...X_N = x_N). \tag{7}$$

Recall that, if the realizations across different X_is are independent, then we can write this joint probability as the product of the marginal probabilities:

$$\begin{aligned} L(X) &= \Pr(X_1 = x_1) \times \Pr(X_2 = x_2) \times ... \times \Pr(X_N = x_N) \\ &= \prod_{i=1}^{N} \Pr(X_i = x_i). \end{aligned} \tag{8}$$

You can think of this intuitively as the probability of observing what we did for the *entire* set of observed data.

Now, think for a minute about Eq. (9):

- Each individual probability is between zero and one, and
- We're multiplying N such terms together, so
- $L(X)$ will necessarily be both (a) between 0 and 1 as well, but (b) almost always *veeery* close to 0.

The fact that a likelihood is often a very, very small number can lead to practical problems, most often because it's easy to run out of precision when calculating it. As a result, we often consider the *log-likelihood*, defined as:

$$\begin{aligned} \ln L(X) &= \ln \left[\prod_{i=1}^{N} \Pr(X_i = x_i) \right] \\ &= \sum_{i=1}^{N} \ln[\Pr(X_i = x_i)] \end{aligned} \tag{9}$$

Now,

- Each of the terms on the right-hand side are a log of a number between 0 and 1 (that is, each is a number between 0 and $-\infty$).
- The sum of those numbers is just some relatively large (in magnitude) negative number.

This is easier to work with; moreover, it retains the same intuitive and mathematical properties of the original likelihood. Likelihoods are tremendously useful, as we'll see shortly...