Bayes for Beginners: Probability and Likelihood

By **C. Randy Gallistel**

Some years ago, a postdoctoral fellow in my lab tried to publish a series of experiments with results that — to his surprise — supported a theoretically important but extremely counterintuitive null hypothesis. He got strong pushback from the reviewers. They said that all he had were insignificant results that could not be used to support his null hypothesis. I knew that Bayesian methods could provide support for null hypotheses, so I began to look into them. I ended up teaching a Bayesian-oriented graduate course in statistics and now use Bayesian methods in analyzing my own data.

When I look back on the formulation of the statistical inference problem I was taught and used for many years, I am astonished that I saw no problem with it: To test our own hypothesis, we test a different hypothesis — the null hypothesis. If it fails, we conclude that our hypothesis is correct — without testing it against the data and without formulating it with the same exactitude with which we formulated the hypothesis we did test (i.e., the null). Moreover, we understand a priori that the null hypothesis can never be accepted; the best it can do is not be rejected. ("You cannot prove the null.") The realization of these absurdities made me a Bayesian.

There is a sense these days that Bayesian data analysis is a coming thing, so colleagues often consult me about it. In these consultations, I am struck by how much misunderstanding there is about basics. In searching for something to write about that would be of general interest, I settled on presenting the basics of Bayesian data analysis in what I hope is an accessible form.

Distinguishing Likelihood From Probability

The distinction between probability and likelihood is fundamentally important: Probability attaches to possible results; likelihood attaches to hypotheses. Explaining this distinction is the purpose of this first column.

Possible results are mutually exclusive and exhaustive. Suppose we ask a subject to predict the outcome of each of 10 tosses of a coin. There are only 11 possible results (0 to 10 correct predictions). The actual result will always be one and only one of the possible results. Thus, the probabilities that attach to the possible results must sum to 1.

Hypotheses, unlike results, are neither mutually exclusive nor exhaustive. Suppose that the first subject we test predicts 7 of the 10 outcomes correctly. I might hypothesize that the subject just guessed, and you might hypothesize that the subject *may* be somewhat clairvoyant, by which you mean that the subject may be expected to correctly predict the results at slightly greater than chance rates over the long run. These are different hypotheses, but they are not mutually exclusive, because you hedged when you said "*may* be." You thereby allowed your hypothesis to include mine. In technical terminology, my hypothesis is nested within yours. Someone else might hypothesize that the subject is strongly clairvoyant and that the observed result underestimates the probability that her next prediction will be correct. Another person could hypothesize something else altogether. There is no limit to the hypotheses one might entertain.

The set of hypotheses to which we attach likelihoods is limited by our capacity to dream them up. In practice, we can rarely be confident that we have imagined all the possible hypotheses. Our concern is to estimate the extent to which the experimental results affect the relative likelihood of the hypotheses we and others currently entertain. Because we generally do not entertain the full set of alternative hypotheses and because some are nested within others, the likelihoods that we attach to our hypotheses do not have any meaning in and of themselves; only the relative likelihoods — that is, the ratios of two likelihoods — have meaning.

Using the Same Function 'Forwards' and 'Backwards'

The difference between probability and likelihood becomes clear when one uses the probability distribution function in general-purpose programming languages. In the present case, the function we want is the binomial distribution function. It is called *BINOM.DIST* in the most common spreadsheet software and *binopdf* in the language I use. It has three input arguments: the number of successes, the number of tries, and the probability of a success.

When one uses it to compute probabilities, one assumes that the latter two arguments (number of tries and the probability of success) are *given*. They are the parameters of the distribution. One varies the first argument (the different possible numbers of successes) in order to find the probabilities that attach to those different possible results (top panel of Figure 1). Regardless of the given parameter values, the probabilities always sum to 1.

By contrast, in computing a likelihood function, one is *given* the number of successes (7 in our example) and the number of tries (10). In other words, the given results are now treated as parameters of the function one is using. Instead of varying the possible results, one varies the probability of success (the third argument, not the first argument) in order to get the binomial likelihood function (bottom panel of Figure 1). One is running the function backwards, so to speak, which is why likelihood is sometimes called *reverse* probability.

The information that the binomial likelihood function conveys is extremely intuitive. It says that given that we have observed 7 successes in 10 tries, the probability parameter of the binomial distribution from which we are drawing (the distribution of successful predictions from this subject) is very unlikely to be 0.1; it is much more likely to be 0.7, but a value of 0.5 is by no means unlikely. The ratio of the
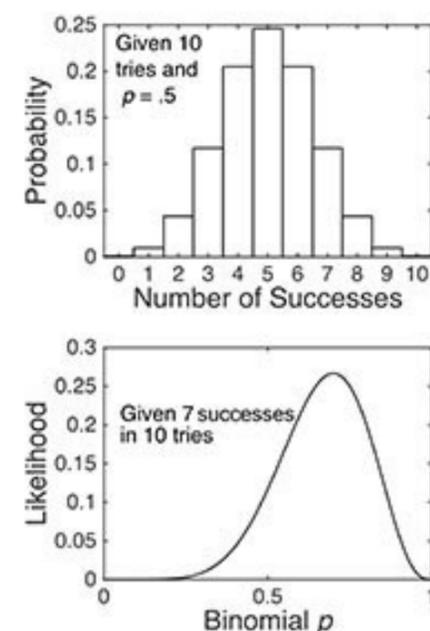
Figure 1. The binomial probability distribution function, given 10 tries at p = .5 (top panel), and the binomial likelihood function, given 7

likelihood at p = .7, which is .27, to the likelihood at p = .5, which is .12, is only 2.28. In other words, given these experimental results (7 successes in 10 tries), the hypothesis that the subject's long-term success rate is 0.7 is only a little more than twice as likely as the hypothesis that the subject's long-term success rate is 0.5.

In summary, the likelihood function is a Bayesian basic. To understand likelihood, you must be clear about the differences between probability and likelihood:

Probabilities attach to results; likelihoods attach to hypotheses. In data analysis, the "hypotheses" are most often *a* possible value or a *range* of possible values for the mean of a distribution, as in our example.

The results to which probabilities attach are mutually exclusive and exhaustive; the hypotheses to which likelihoods attach are often neither; the range in one hypothesis may include the point in another, as in our example.

To decide which of two hypotheses is more likely given an experimental result, we consider the ratios of their likelihoods. This ratio, the relative likelihood ratio, is called the "Bayes Factor."

successes in 10 tries (bottom panel). Both panels were computed using the *binopdf* function. In the upper panel, I varied the possible results; in the lower, I varied the values of the *p* parameter. The probability distribution function is discrete because there are only 11 possible experimental results (hence, a bar plot). By contrast, the likelihood function is continuous because the probability parameter *p* can take on any of the infinite values between 0 and 1. The probabilities in the top plot sum to 1, whereas the integral of the continuous likelihood function in the bottom panel is much less than 1; that is, the likelihoods do not sum to 1.

Tags: C. Randy Gallistel Columns, Data, Experimental Psychology, Methodology, Statistical Analysis