Published in *Advances in Methods and Practices in Psychological Science*. Slight differences exist between this preprint and the version of record: https://doi.org/10.1177/2515245917744314.

Introduction to the concept of likelihood and its applications

Alexander Etz
University of California, Irvine

Abstract

We introduce the statistical concept known as likelihood and discuss how it underlies common Frequentist and Bayesian statistical methods. This article is suitable for researchers interested in understanding the basis of their statistical tools, and is also ideal for teachers to use in their classrooms to introduce the topic to students at a conceptual level.

Introduction

Likelihood is a concept that underlies most common statistical methods used in psychology. It is the basis of classical maximum likelihood estimation methods, and it plays a key role in Bayesian inference. Despite the ubiquity of likelihood in modern statistical methods, few basic introductions are available to the practicing psychological researcher. This tutorial aims to explain the concept of likelihood and illustrate in an accessible way how it enables some of the most used classical and Bayesian statistical analyses; in doing so many finer details are skipped over, but interested readers can consult Pawitan (2001) for a complete mathematical treatment of the topic (Edwards, 1974, provides a historical review). This article is aimed at applied researchers interested in understanding the basis of their statistical tools, and is also ideal for research methods instructors who want to introduce the topic to students at a conceptual level.

Likelihood is a strange concept, in that it is not a probability, but it is proportional to a probability. The likelihood of a hypothesis (H) given some data (D) is proportional to the probability of obtaining D given that H is true, multiplied by an arbitrary positive constant K. In other words, $L(H) = K \times P(D|H)$. In most cases hypotheses represent different values of a parameter in a statistical model, such as the mean of a normal distribution. Since a likelihood is not actually a probability it doesn't obey various rules of probability; for example, likelihoods need not sum to 1.

A critical difference between probability and likelihood is in the interpretation of what is fixed and what can vary. In the case of a conditional probability, $P(D|H)$, the hypothesis is fixed and the data are free to vary. Likelihood, however, is the opposite. The likelihood of

The author was supported by grant #1534472 from the National Science Foundation's Methods, Measurements, and Statistics panel, as well as the National Science Foundation Graduate Research Fellowship Program #DGE1321846. A portion of this material previously appeared on the author's personal blog. The author is very grateful to Quentin Gronau and J. P. de Ruiter for helpful comments.

a hypothesis, $L(H)$, conditions on the data as if they are fixed while allowing the hypotheses to vary. The distinction is subtle, so it is worth repeating: For conditional probability, the hypothesis is treated as a given and the data are free to vary. For likelihood, the data are treated as a given and the hypotheses vary.

The Likelihood Axiom

Edwards (1992, p. 30) defines the Likelihood Axiom as a natural combination of the Law of Likelihood and the Likelihood Principle. The Law of Likelihood states that "within the framework of a statistical model, a particular set of data supports one statistical hypothesis better than another if the likelihood of the first hypothesis, on the data, exceeds the likelihood of the second hypothesis" (Edwards, 1992, p. 30). In other words, there is evidence for H_1 over H_2 if and only if the probability of the data under H_1 is greater than the probability of the data under H_2. That is, D is evidence for H_1 over H_2 if $P(D|H_1) > P(D|H_2)$. If these two probabilities are equivalent, then there is no evidence for either hypothesis over the other. Furthermore, the strength of the statistical evidence for H_1 over H_2 is quantified by the ratio of their likelihoods. We write the likelihood ratio as $LR(H_1, H_2) = L(H_1)/L(H_2)$ — which is equal to $P(D|H_1)/P(D|H_2)$ since the arbitrary constants cancel out of the fraction.

The following brief example will help illustrate the main idea underlying the Law of Likelihood.[1] Consider the case of Earl, who is visiting a foreign country that has a mix of women-only and mixed-gender saunas (known to be visited equally often by both genders). After a leisurely jog through the city he decides to stop by a nearby sauna to try to relax. Unfortunately, Earl does not know the local language so he cannot determine from the posted signs whether this sauna is women-only or mixed-gender. While Earl is attempting to decipher the signs he observes three women independently exit the sauna. If the sauna is women-only, the probability that all three exiting patrons would be women is 1.0; if the sauna is mixed-gender, this probability is $.5^3 = .125$. With this information Earl can compute the likelihood ratio between the women-only hypothesis and the mixed-gender hypothesis to be $1.0/.125 = 8$, or eight-to-one evidence in favor of the sauna being women-only.

The Likelihood Principle states that the likelihood function contains all of the information relevant to the evaluation of statistical evidence. Other facets of the data that do not factor into the likelihood function (e.g., the cost of collecting each observation or the stopping rule used when collecting the data) are irrelevant to the evaluation of the strength of the statistical evidence (Edwards, 1992, p. 30; Royall, 1997, p. 22). They can be meaningful for planning studies or for decision analysis, but they are separate from the strength of the statistical evidence.

Likelihoods are inherently comparative

Unlike a probability, a likelihood has no real meaning *per se* due to the arbitrary constant K. Only by comparing likelihoods do they become interpretable, because the

[1] This light-hearted example was first suggested to the author by J. P. de Ruiter, who graciously permitted its inclusion in this manuscript.

constant in each likelihood cancels the other one out. A simple way to explain this aspect of likelihood is to give an example using the Binomial distribution.

Suppose a coin is flipped n times and we observe x heads and $n - x$ tails. The probability of getting x heads in n flips is defined by the Binomial distribution,

$$P(X = x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{1}$$

where p is the probability of heads, and

$$\binom{n}{x} = \frac{n!}{x!(n-x!)}$$

is the number of ways to get x heads in n flips. For example, if $x = 2$ and $n = 3$, this value is $3!/(2! \times 1!) = 6/2 = 3$, since there are three ways to get two heads in three flips (i.e., Head-Head-Tail, Head-Tail-Head, Tail-Head-Head). Thus, the probability of getting two heads in three flips if p is .50 would be $3 \times .50^2 \times (1 - .50)^1 = .375$, or three out of eight.

If the coin were fair, so that $p = .50$, and we flip it ten times, the probability it comes up six heads and four tails is

$$P(X = 6 \mid p = .50) = \frac{10!}{6! \times 4!} (.50)^6 (1 - .50)^4 \approx .21.$$

If the coin were a trick coin, so that $p = .75$, the probability of six heads in ten tosses is

$$P(X = 6 \mid p = .75) = \frac{10!}{6! \times 4!} (.75)^6 (1 - .75)^4 \approx .15.$$

To quantify the statistical evidence for the first hypothesis against the second, we simply divide one probability by the other. This ratio tells us everything we need to know about the support the data lends to one hypothesis *vis-a-vis* the other. In the case of 6 heads in 10 tosses, the likelihood ratio for a fair coin versus our trick coin, denoted $LR(.50, .75)$, is:

$$LR(.50, .75) = \left(\frac{10!}{6! \times 4!} (.50)^6 (1 - .50)^4 \right) \div \left(\frac{10!}{6! \times 4!} (.75)^6 (1 - .75)^4 \right) \approx .21/.15 = 1.4.$$

In other words, the data are 1.4 times more probable under a fair coin hypothesis than under the trick coin hypothesis. Notice how the first terms in each of the equations above, $10!/(6! \times 4!)$, are equivalent and completely cancel each other out in the likelihood ratio.

Same data. Same constant. Cancel out.

The first term in the equations above, $10!/(6! \times 4!)$, details our journey to obtaining six heads out of ten flips. If we change our journey (i.e., different sampling plan) then this changes the term's value, but crucially, since it is the same term in both the numerator and denominator it always cancels itself out. For example, if we were to change our sampling scheme from flipping the coin ten times and counting the number of heads, to one where we flip our coin *until six heads arise*, this first term changes to $9!/(5! \times 4!)$ (Lindley, 1993).

But, crucially, since this term is in both the numerator and denominator the information contained in the way the data are obtained disappears from the likelihood ratio. This result leads to the conclusion that the sampling plan should be irrelevant to the evaluation of statistical evidence, which is something that makes likelihood and Bayesian methods particularly flexible (Gronau & Wagenmakers, in press; Rouder, 2014).

Consider if we had left out the first term in the above calculations, so that our numerator is $P(X = 6 \mid p = .50) = (.50)^6(1 - .50)^4 = 0.000976$ and our denominator is $P(X = 6 \mid p = .75) = (.75)^6(1 - .75)^4 = 0.000695$. Using these values to form the likelihood ratio we get $LR(.50, .75) = 0.000976/0.000695 = 1.4$, confirming our initial result since the other terms simply canceled out before. Again it is worth repeating that the value of a single likelihood is meaningless in isolation; only in comparing likelihoods do we find meaning.

Inference using the likelihood function

Visual inspection

So far, likelihoods may seem overly restrictive because we have only compared two simple statistical hypotheses in a single likelihood ratio. But what if we are interested in comparing all possible hypotheses at once? By plotting the entire likelihood function we compare all possible hypotheses simultaneously, and this lets us 'see' the evidence the data provide in its entirety. Birnbaum (1962) remarked that "the 'evidential meaning' of experimental results is characterized fully by the likelihood function" (p. 269), so now we will look at some examples of likelihood functions and see what insights we can glean from them.[2]

The top panel of Figure 1 shows the likelihood function for six heads in ten flips. The fair coin and trick coin hypotheses are marked on the likelihood curve with dots. Since the likelihood function is meaningful only up to an arbitrary constant, the graph is scaled by convention so that the best supported value (i.e., the maximum) corresponds to a likelihood of 1. The likelihood ratio of any two hypotheses is simply the ratio of their heights on this curve. We can see from the plot that the fair coin has a higher likelihood than our trick coin, which we saw previously to be roughly a factor of 1.4.

The middle panel of Figure 1 shows how the likelihood function changes if instead of 6 heads out of 10 tosses, we tossed 100 times and obtained 60 heads: the curve gets much narrower. The strength of evidence favoring the fair coin over the trick coin has also changed, with the new likelihood ratio being 29.9. This is much stronger evidence, but due to the narrowing of the likelihood function neither of these hypothesized values are very high up on the curve anymore. It might be more informative to compare each of our hypotheses against the best supported hypothesis. This gives us two likelihood ratios: $LR(.60, .50) = 7.5$ and $LR(.60, .75) = 224$.

The bottom panel in Figure 1 shows the likelihood function for the case of 300 heads in 500 coin flips. Notice that both the fair coin and trick coin hypotheses appear to be very near the minimum of the graph; yet their likelihood ratio is much stronger than before. For this data the likelihood ratio is $LR(.50, .75) = 23912304$, or nearly twenty-four million. The inherent relativity of evidence is made clear here: The fair coin was supported when compared to one particular trick coin. But this should not be interpreted as absolute evidence

[2] An R script is available to reproduce these plots (https://osf.io/t2ukm/).

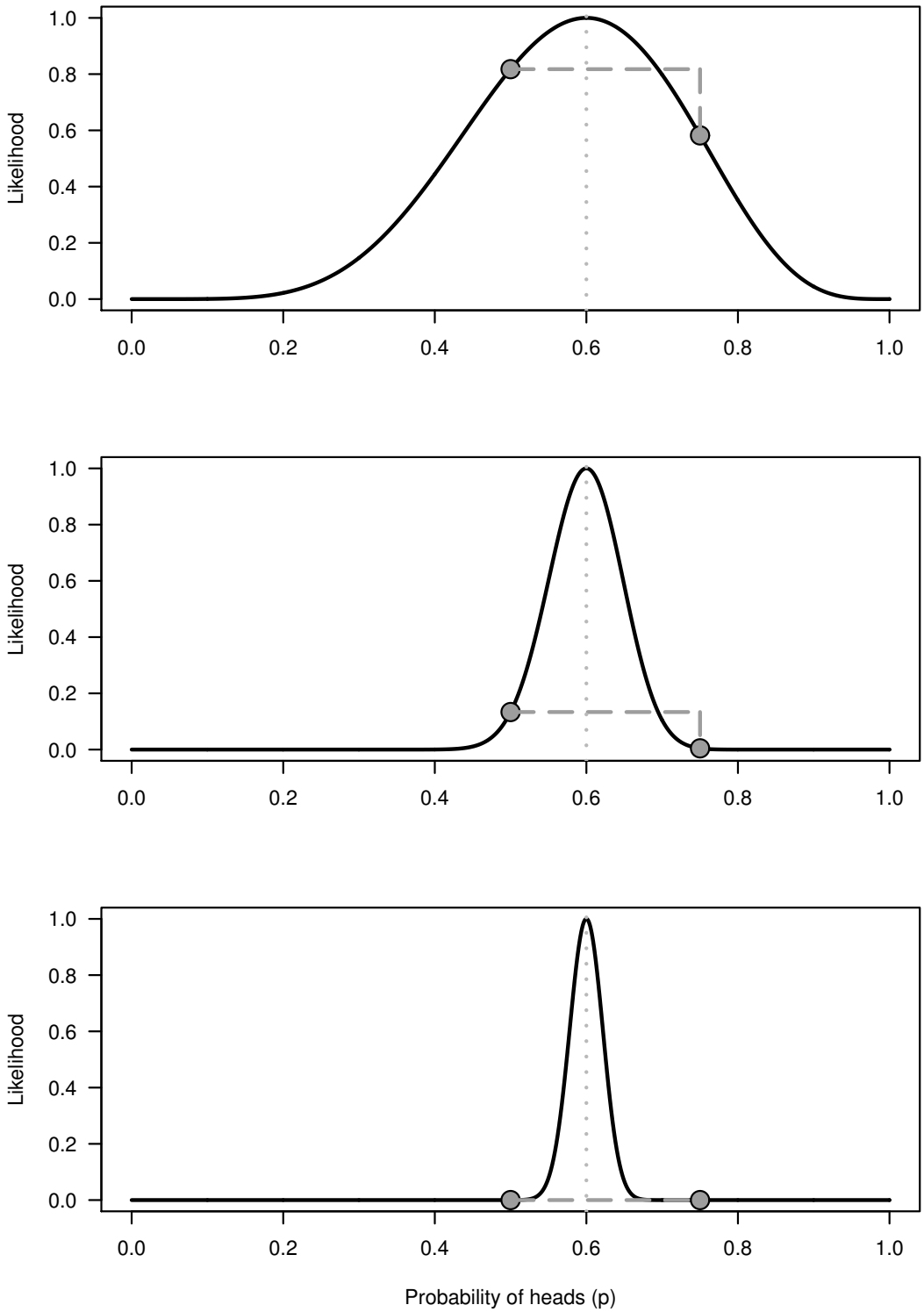

Figure 1. **Top.** The likelihood function for the case of 6 heads in 10 flips. **Middle.** The likelihood function for 60 heads in 100 flips. **Bottom.** The likelihood function for 300 heads in 500 flips.

for the fair coin, because the likelihood ratio for the maximally supported hypothesis vs the fair coin, $LR(.60, .50)$, is nearly twenty-four thousand.

We need to be careful not to make blanket statements about absolute support, such as claiming that the maximum is "strongly supported by the data". Always ask, "Compared to what?" The best supported hypothesis will usually only be weakly supported against any hypothesis a little smaller or a little larger.[3] For example, in the case of 60 heads in 100 flips the likelihood ratio between $p = .60$ and $p = .75$ is very large at $LR(.60, .75) = 224$, whereas the likelihood ratio between $p = .60$ and $p = .65$ is only $LR(.60, .65) = 1.7$, which is barely any support one way or the other. Consider the following common real-world research setting: We have run a study with a relatively small sample size, and the estimate of the effect of primary scientific interest is considered "large" by some criteria (e.g., Cohen's $d > .75$). We may find that the estimated effect size from the sample has a relatively large likelihood ratio compared to a hypothetical null value (large enough to "reject the null hypothesis", see below), while at the same time it has a much less forceful likelihood ratio compared to a "medium" or even "small" effect size. Without relatively large sample sizes we are often precluded from saying anything precise about the size of the effect because the likelihood function is not very peaked in small samples.

Maximum likelihood estimation

A natural question for a researcher to ask is, what is the hypothesis that is most supported by the data? This question is answered by using a method called maximum likelihood estimation (Fisher, 1922; see also Ly et al., in press and Myung, 2003). Looking again at the plots in Figure 1, the vertical dotted lines mark the value of p that is has the highest likelihood, known as the *maximum likelihood estimate*. We interpret this value of p as being the hypothesis that makes the observed data the most probable. Since the likelihood is proportional to the probability of the data given the hypothesis, the hypothesis that maximizes $P(D \mid H)$ will also maximize $L(H) = K \times P(D \mid H)$. In simple problems, plotting the likelihood function will reveal an obvious maximum. For example, the maximum of the Binomial likelihood function will be located at the sample proportion of successes. In Box 1 we show this to be true using a little bit of elementary calculus.[4]

With the maximum likelihood estimate in hand there are a few possible ways to proceed with classical inference. First, one can perform a *likelihood ratio test* comparing two versions of the proposed statistical model: One where we set a parameter to a hypothesized null value, and one where we estimate the parameter from the data (these are called *nested* models). In practice this amounts to a comparison between the value of the likelihood at the maximum likelihood estimate against the value of the likelihood at the proposed null

[3] The amount that the likelihood ratio changes with small deviations from the maximum is fundamentally captured by its peakedness. Formally, the peakedness (or curvature) of a function at a given point is found by taking the second derivative at that point. If that function happens to be the logarithm of the likelihood function for some parameter θ (theta), and the point of interest is its maximum point, the negative of the second derivative is called the *observed Fisher Information*, or sometimes simply the *observed Information*, written as $I(\theta)$ (taking the negative is a convention to make the Information a positive quantity, since the second derivative of a function at its maximum will be negative). See Ly, Marsman, Verhagen, Grasman, and Wagenmakers (in press) for more technical details.

[4] In more complicated scenarios with many parameters there are usually not simple equations one can directly solve to find the maximum, so we must turn to numerical approximation methods.

value. Likelihood ratio tests are commonly used when drawing inferences with structural equation models. In the Binomial example from earlier, this would mean comparing the probability of the data if p were set to .50 (the fair coin hypothesis) to the probability of the data x given the value of p estimated from the data (the maximum likelihood estimate). In general we call the parameter of interest θ (theta), and it can be shown that when the null hypothesis is true, and as the sample size gets large, twice the logarithm of this likelihood ratio approximately follows a chi-squared distribution with a single degree of freedom (Casella & Berger, 2002, p.489; Wilks, 1938),

$$2 \log \left[\frac{P(X \mid \theta_{\text{max}})}{P(X \mid \theta_{\text{null}})} \right] \;\dot{\sim}\; \chi^2(1), \tag{2}$$

where $\dot{\sim}$ means "is approximately distributed as". If the value of the quantity on the left-hand side of Equation 2 lies far enough out in the tail of the chi-squared distribution, such that the p-value is lower than a prespecified cutoff α (alpha, often chosen to be .05), then one would make the decision to reject the hypothesized null value.[5]

Secondly, one can perform a *Wald test*, where the maximum likelihood estimate is compared to a hypothesized null value and this difference is divided by the estimated standard error. Essentially, we are determining how many standard errors separate the null value and the maximum likelihood estimate. The t-test and Z-test are arguably the most common examples of the Wald test, which are used for inference about parameters ranging from simple comparisons of means to complex multilevel regression models. For many common statistical models it can be shown (e.g., Casella & Berger, 2002, p. 493) that if the null hypothesis is true, and as the sample size gets large, this ratio approximately follows a normal distribution with a mean of zero and standard deviation of one,

$$\frac{\theta_{\text{max}} - \theta_{\text{null}}}{\text{se}(\theta_{\text{max}})} \;\dot{\sim}\; N(0, 1). \tag{3}$$

Analogously to the likelihood ratio test, if the value of this ratio falls far out in the tails of the normal distribution, such that the p-value is less than the prespecified α, then one would make the decision to reject the hypothesized null value. This result also allows for easy construction of 95% confidence interval bounds by computing $\theta_{\text{max}} \pm 1.96 \times \text{se}(\theta_{\text{max}})$. It is important to note that these confidence intervals are based on large-sample approximations, whose performance can be suboptimal in relatively small samples (Ghosh, 1979).

Figure 2 depicts the relationship between the two types of tests by showing the log of the likelihood function for 60 heads in 100 flips. The likelihood ratio test looks at the difference in the height of the likelihood at its maximum compared to the height at the null, and rejects the null if the difference is large enough. In contrast, the Wald test looks at how many standard errors the maximum is from the null value and rejects the null if the estimate is sufficiently far away. In other words, the likelihood ratio test evaluates

[5] We saw in the previous section that the value of the likelihood ratio itself does not depend on the sampling plan, but now we see that the likelihood ratio test does depend on the sampling plan since it requires the sampling distribution of twice the logarithm of the likelihood ratio to be chi-squared. Royall (1997) resolves this potential inconsistency by pointing out that the likelihood ratio and its test answer different questions: The former answers the question, "How should I interpret the data as evidence?" The latter answers the question, "What should I do with this evidence?"

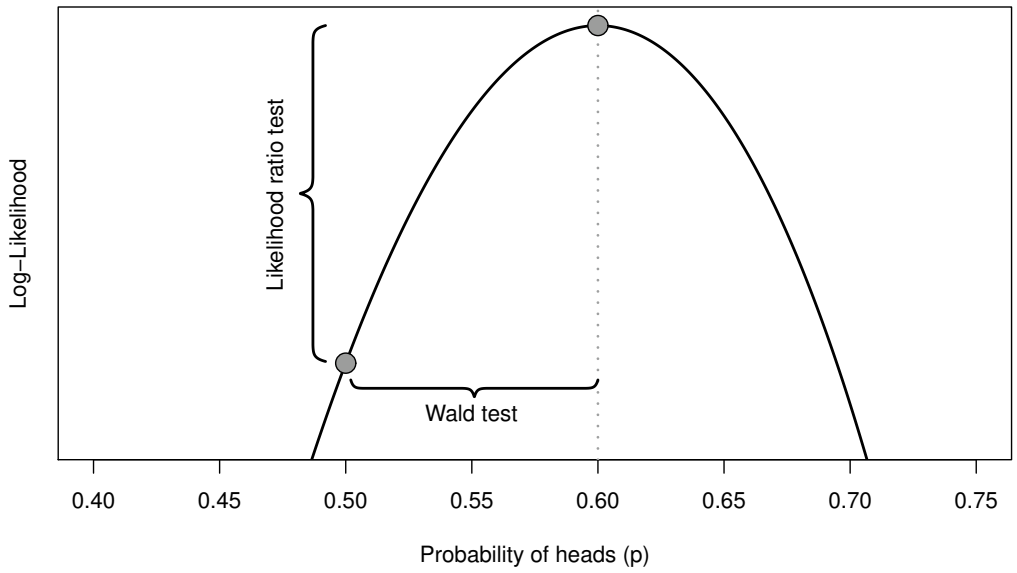

Figure 2. The logarithm of the Binomial likelihood function for 60 heads in 100 flips. The x-axis is restricted to the values of p that have appreciable log-likelihood values. The likelihood ratio test evaluates the points' vertical discrepancy, whereas the Wald test evaluates their horizontal discrepancy.

the vertical discrepancy of two values on the likelihood function (y-axis) and the Wald test evaluates the horizontal discrepancy between two values of the parameter (x-axis). As sample size grows very large the results from these methods converge (Engle, 1984), but in practice each has its advantages. An advantage of the Wald test is its simplicity when testing a single parameter at a time; all one needs is a point estimate and its standard error to easily perform hypothesis tests and compute confidence intervals. An advantage of the likelihood ratio test is that it is easily extended to a simultaneous test of multiple parameters by increasing the degrees of freedom of the chi-squared distribution in Equation 2 to be equal to the number of parameters being tested (Wald tests can be extended to the multiparameter case but they are not as elegant as the likelihood ratio test).

Bayesian updating via the likelihood

As we have seen, likelihoods form the basis for much of classical statistics in the form of maximum likelihood estimation. Likelihoods are also a key component of Bayesian inference. The Bayesian approach to statistics is fundamentally about making use of all available information when drawing inferences in the face of uncertainty. This information may come in the form of results from previous studies, or it may come in the form of newly collected data, or, as is usually the case, from both. Bayesian inference allows us to synthesize these two forms of information to make the best possible inference.

We quantify our previous information using what is known as a *prior distribution*.

We write the prior distribution of θ, our parameter of interest, as $P(\theta)$; this is a function that specifies which values of θ are more or less likely, based on our interpretation of our previous relevant information. The information gained from our new data is represented by the likelihood function, proportional to $P(D \mid \theta)$, which is then multiplied by the prior distribution (and rescaled) to yield the posterior distribution, $P(\theta \mid D)$, with which we can perform our desired inference. Thus, we can say that the likelihood function acts to update our prior distribution to a posterior distribution. A detailed technical introduction to Bayesian inference can be found in Etz and Vandekerckhove (in press), and interested readers can find an annotated list of useful Bayesian references in Etz, Gronau, Dablander, Edelsbrunner, and Baribault (in press).

Mathematically, we use Bayes' rule to obtain the posterior distribution of our parameter of interest θ,

$$P(\theta \mid D) = K \times P(\theta) \times P(D \mid \theta),$$

where in this context $K = 1/P(D)$ is merely a rescaling constant. We often write this more simply as

$$P(\theta \mid D) \propto P(\theta) \times P(D \mid \theta),$$

where $\propto$ means "is proportional to". In words, we say the posterior distribution is proportional to the prior distribution multiplied by the likelihood function.

In the example below we will see how to use the likelihood to update a prior into a posterior. The simplest way to illustrate how likelihoods act as an updating factor is to use conjugate distribution families (Raiffa & Schlaifer, 1961). A prior and likelihood are said to be conjugate when the resulting posterior distribution is the same type of distribution as the prior. For example, if we have Binomial data we can use a Beta prior to obtain a Beta posterior (see Box 2). Conjugate priors are by no means required for doing Bayesian updating, but they reduce the mathematics involved and so are ideal for illustrative purposes.

Consider the coin example from before, where we had 60 heads in 100 flips. Imagine that going in to this experiment we had some reason to believe the coin's bias was within .2 of being fair in either direction, so that it is likely p is within the range of .30 to .70. We could choose to represent this information using the Beta(25,25) distribution[6] shown as the dotted line in Figure 3. The likelihood function for the 60 flips is shown as the dot-and-dashed line in Figure 3, and is identical to that shown in the middle panel from Figure 1. Using the result from Box 2, we know that the outcome of this experiment will be the Beta(85,65) posterior distribution shown as the solid line in Figure 3.

The entire posterior distribution represents the solution to our Bayesian estimation problem, but often we report summary measures to simplify the communication of the results. For instance, we could point to the maximum of the posterior distribution – known as the maximum a posteriori (MAP) estimate – as our best guess for the value of p, which in this case is $p_{\text{map}} = .568$. Notice that this is slightly different from the estimate we would obtain if we computed the maximum likelihood estimate, $p_{\text{mle}} = .60$. This discrepancy is due

[6] The Beta(a, b) distribution spans from 0 to 1, and its two arguments a and b determine its form: when $a = b$ the distribution is symmetric around .50; as a special case, when $a = b = 1$ the distribution is uniform (flat) between 0 and 1; when $a > b$ the distribution puts more mass on values above .50, and vice-versa.

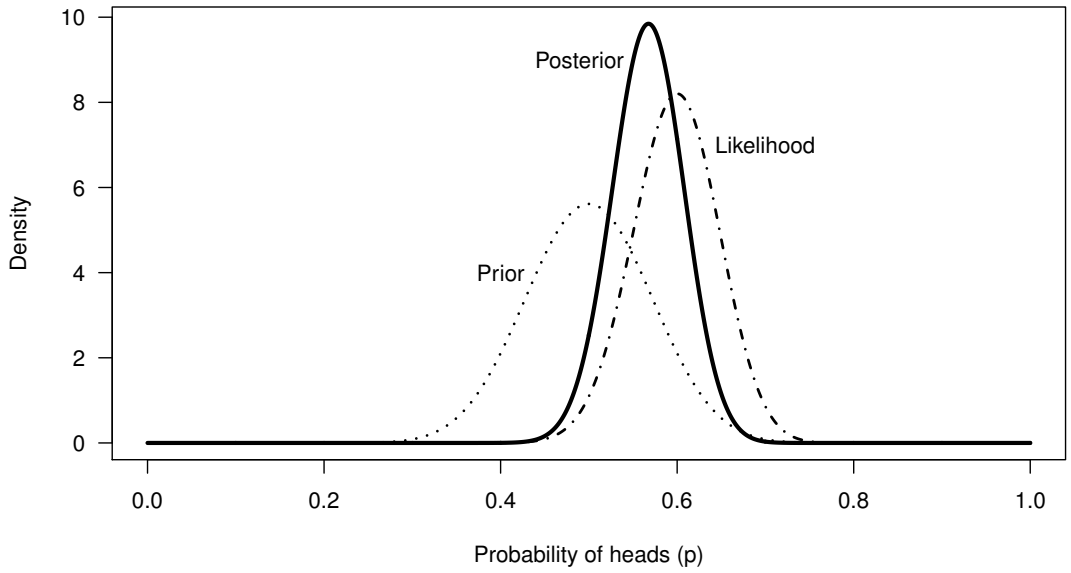

Figure 3. Bayesian updating of the prior distribution to posterior distribution. The Posterior distribution is a compromise between the information brought by the prior and the information brought by the likelihood.

to the extra information about p provided by the prior distribution; as shown in Box 2, the prior distribution effectively adds a number of previous successes and failures to our sample data. Thus, the posterior distribution represents a compromise between the information we had regarding p before the experiment and the information gained about p by doing the experiment. Since we had previous information suggesting p is probably close to .50, our posterior estimate is said to be "shrunk" towards .50. This and other regularization methods tend to lead to more accurate estimates and better empirical predictions (Efron & Morris, 1977), especially when we have a relatively small sample size.

The likelihood also forms the basis of the *Bayes factor*, a tool for conducting Bayesian hypothesis tests first proposed by Wrinch and Jeffreys (Jeffreys, 1935; Wrinch & Jeffreys, 1921), and independently developed by Haldane (Haldane, 1932; although see Etz & Wagenmakers, 2017). An important advantage of the Bayes factor is that it can be used to compare any two models regardless of their form, whereas the frequentist likelihood ratio test can only compare two hypotheses that are nested. Nevertheless, when comparing nested hypotheses Bayes factors can be seen as simple extensions of likelihood ratios. In contrast to the frequentist likelihood ratio test outlined above, which evaluates the size of the likelihood ratio comparing θ_{null} to θ_{max}, the Bayes factor takes a weighted average of the likelihood ratio across *all* possible values of θ; the likelihood ratio is evaluated at each value of θ and weighted by the prior probability density we assign that value, and then these products are added up to obtain the Bayes factor (mathematically, this is done by

integrating the likelihood ratio with respect to the prior distribution; see the Appendix).

Conclusion

The aim of this tutorial has been to provide an accessible introduction to the concept of likelihood. We have seen how the likelihood is defined, what a likelihood function looks like, and how this function forms the basis of two common inferential procedures: maximum likelihood estimation and Bayesian inference. The examples used were intentionally simple and artificial to keep the mathematical burden light, but hopefully they can give some insight into the fundamental statistical concept known as likelihood.

Author contributions

All aspects regarding this manuscript are the work of AE.

Box 1. The Binomial likelihood function is given in Equation 1 above, and our goal is to find the value of p that makes the probability of the outcome x the largest. Recall that to find possible maximums or minimums of a function, we take the derivative of the function, set it equal to zero, and solve. We can find the maximum of the likelihood function by first taking the logarithm of the function and then taking the derivative, since maximizing $\log[f(x)]$ will also maximize $f(x)$. This will make our life easier, since taking the logarithm changes multiplication to addition and the derivative of $\log[x]$ is simply $1/x$. Moreover, since log-likelihood functions are generally unimodal concave functions (they have a single peak and open downward), if we can do this calculus and solve our equation then we will have found our desired maximum.

We begin by taking the logarithm of Equation 1,

$$\log\Big[P(X = x \mid p)\Big] = \log\left[\binom{n}{x}\right] + \log\Big[p^x\Big] + \log\Big[(1-p)^{n-x}\Big]. \tag{4}$$

Remembering the rules of logarithms and exponents, we can rewrite this as

$$\log\Big[P(X = x \mid p)\Big] = \log\left[\binom{n}{x}\right] + x\log[p] + (n-x)\log[1-p]. \tag{5}$$

Now we can take the derivative of Equation 5 as follows:

$$\begin{aligned}
\frac{d}{dp}\left(\log\Big[P(X = x \mid p)\Big]\right) &= \frac{d}{dp}\left(\log\left[\binom{n}{x}\right] + x\log[p] + (n-x)\log[1-p]\right) \\
&= 0 + x\left(\frac{1}{p}\right) + (n-x)(-1)\left(\frac{1}{1-p}\right) \\
&= \frac{x}{p} - \frac{n-x}{1-p},
\end{aligned}$$

(where the -1 in the last term of the second line comes from using the chain rule of derivatives on $\log[1-p]$). Now we can set the above equal to zero and a few algebraic

steps will lead us to the solution for p,

$$\begin{aligned} 0 &= \frac{x}{p} - \frac{n-x}{1-p} \\ \frac{n-x}{1-p} &= \frac{x}{p} \\ np - xp &= x - xp \\ np &= x \\ p &= \frac{x}{n}. \end{aligned}$$

In other words, the maximum of the Binomial likelihood function is found at the sample proportion, namely, the number of successes x divided by the total number of trials n.

Box 2. Conjugate distributions are convenient in that they reduce Bayesian updating to some simple algebra. We take the formula for the Binomial likelihood function,

$$\text{Likelihood} \propto p^x(1-p)^{n-x}, \tag{6}$$

(notice we dropped the leading term from before) and then multiply it by the formula for the Beta prior with a and b shape parameters,

$$\text{Prior} \propto p^{a-1}(1-p)^{b-1}, \tag{7}$$

to obtain the following formula for the posterior:

$$\text{Posterior} \propto \underbrace{p^{a-1}(1-p)^{b-1}}_{\text{Prior}} \times \underbrace{p^x(1-p)^{n-x}}_{\text{Likelihood}}. \tag{8}$$

The terms in Equation 8 can be regrouped as follows:

$$\text{Posterior} \propto \underbrace{p^{a-1}p^x}_{\text{Successes}} \times \underbrace{(1-p)^{b-1}(1-p)^{n-x}}_{\text{Failures}}. \tag{9}$$

which suggests that we can interpret the information contained in the prior as adding a certain amount of previous data (i.e., $a-1$ past successes and $b-1$ past failures) to the data from our current experiment. Since we are multiplying together terms with the same base, the exponents can be added together in a final simplification step,

$$\text{Posterior} \propto p^{x+a-1}(1-p)^{n-x+b-1}.$$

This final formula looks like our original Beta distribution but with new shape parameters equal to $x+a$ and $n-x+b$. In other words, we start with the prior, Beta(a,b), and add the successes from the data, x, to a and the failures, $n-x$, to b, and our posterior is a Beta($x+a$, $n-x+b$) distribution.

Appendix

Note that for a generic random variable ω (omega), the expected value (i.e., average) of the function $g(\omega)$ with respect to a probability distribution $P(\omega)$ can be written as

$$E_\omega \left[g(\omega) \right] = \int_\Omega g(\omega) P(\omega) d\omega. \tag{10}$$

We will use this result below to show that the Bayes factor between nested models can be written as the expected value of the likelihood ratio with respect to the specified prior distribution.

The Bayes factor comparing H_1 to H_0 is written similarly to the likelihood ratio,

$$BF_{10} = \frac{P(D \mid H_1)}{P(D \mid H_0)}. \tag{11}$$

In the context of comparing nested models, H_0 specifies that $\theta = \theta_{\text{null}}$, so $P(D \mid H_0) = P(D \mid \theta_{\text{null}})$; H_1 assigns θ a prior distribution, $\theta \sim P(\theta)$, so that $P(D|H_1) = \int_\Theta P(D \mid \theta) P(\theta) d\theta$. Thus, we can rewrite Equation 11 as

$$BF_{10} = \frac{\int_\Theta P(D \mid \theta) P(\theta) d\theta}{P(D \mid \theta_{\text{null}})}. \tag{12}$$

Since the denominator of Equation 12 is a fixed number we can bring it inside the integral, and we can see this has the form of the expected value of the likelihood ratio between θ and θ_{null} with respect to the prior distribution of θ,

$$\begin{aligned} BF_{10} &= \int_\Theta \overbrace{\frac{P(D|\theta)}{P(D \mid \theta_{\text{null}})}}^{LR(\theta, \theta_{\text{null}})} P(\theta) d\theta \\ &= E_\theta \left[LR(\theta, \theta_{\text{null}}) \right]. \end{aligned}$$

References

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, *53*, 259–326.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Duxbury Pacific Grove, CA.

Edwards, A. W. F. (1974). The history of likelihood. *International Statistical Review/Revue Internationale de Statistique*, 9–15.

Edwards, A. W. F. (1992). *Likelihood*. Baltimore, MD: The Johns Hopkins University Press.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

Engle, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of econometrics*, *2*, 775–826.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (in press). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin and Review*.

Etz, A., & Vandekerckhove, J. (in press). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*.

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*(2), 313–329.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*, 309–368.

Ghosh, B. (1979). A comparison of some approximate confidence intervals for the binomial parameter. *Journal of the American Statistical Association*, *74*(368), 894–900.

Gronau, Q. F., & Wagenmakers, E.-J. (in press). Bayesian evidence accumulation in experimental mathematics: A case study of four irrational numbers. *Experimental Mathematics*.

Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *28*, 55–61.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophy Society*, *31*, 203–222.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*, 22–25.

Ly, A., Marsman, M., Verhagen, A., Grasman, R. P., & Wagenmakers, E.-J. (in press). A tutorial on Fisher information. *Journal of Mathematical Psychology*.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100.

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. Oxford University Press.

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory*. Cambridge (MA): The MIT Press.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.