Likelihood Ratios: A Tutorial on Applications to Research in Psychology

Scott Glover

Royal Holloway University of London

RUNNING HEAD: LIKELIHOOD RATIOS

Address correspondence to:

Dr. Scott Glover

Dept. of Psychology

Royal Holloway University of London

Egham, Surrey, TW20 0EX

scott.glover@rhul.ac.uk

Abstract

Many in psychology view their choice of statistical approaches as being between frequentist and Bayesian. However, a third approach, the use of likelihood ratios, provides several distinct advantages over both the frequentist and Bayesian options. A quick explanation of the basic logic of likelihood ratios is provided, followed by a comparison of the likelihood-based approach to frequentist and Bayesian methods. The bulk of the paper provides examples with formulas for computing likelihood ratios based on *t*-scores, ANOVA outputs, *chi-square* statistics, and binomial data, as well as examples of using likelihood ratios to test for models that make *a priori* predictions of effect sizes. Finally, advice on interpretation is offered.

Keywords: likelihood ratios, t-tests, ANOVA, chi-square, binomial

Introduction: What is a Likelihood Ratio?

A likelihood ratio is a statistic expressing the relative likelihood of the data given two competing models. The likelihood ratio, λ , can be written as

$$\lambda = \frac{f(X|\hat{\theta}_2)}{f(X|\hat{\theta}_1)} \quad , \quad (1)$$

where f is the probability density, X is the vector of observations, and $\hat{\theta}_1$ and $\hat{\theta}_2$ are the vectors of parameter estimates that maximize the likelihood under the two models. Often, likelihood ratios involve comparing the likelihood of the data given a model based on the point estimate (also known as the “maximum likelihood estimate” or “MLE”) relative to the likelihood of the data given no effect (the null hypothesis). A “raw” likelihood ratio is the expression of the relationship between the frequency densities of those two models, as illustrated in Figure 1. For example, a raw likelihood ratio of $\lambda = 5$ results when the density of the MLE is five times the density of the H_0 distribution at the same point. This indicates that the data are five times as likely to occur given an effect based on the maximum likelihood estimate than given no effect (Goodman & Royall, 1988; Royall, 1997).

Figure 1. The raw likelihood ratio based on the maximum likelihood estimate (MLE). The grey curve shows the distribution based on the observations which form the basis of the alternative hypothesis (H_a). The blank curve shows the distribution under the null hypothesis (H_o). The dotted and solid arrows show the frequency density of the distributions under the two hypotheses, and the raw likelihood ratio is the ratio of these two densities. In this example, the raw likelihood ratio is $\lambda = 5.0$ in favor of the alternative hypothesis over the null.

Adjusted Likelihood Ratios

In many circumstances, a raw likelihood ratio must be adjusted to reflect the different number of parameters in the models under consideration. In the typical case of determining whether an effect differs from zero, for example, the model based on the MLE will usually have an extra parameter(s) relative to the null, and will almost always provide a better fit to the data. Failure to adjust the likelihood ratio for unequal numbers of parameters would result in a bias towards the model with more parameters, a phenomenon known as “overfitting” (Burnham & Anderson, 2002). The result of applying this penalty to the model with more parameters is an “adjusted” likelihood ratio, expressed as λ_{adj} . This tutorial will include instructions for how

to calculate both raw (λ) and adjusted (λ_{adj}) likelihood ratios, and when it is appropriate to use them. For testing the null versus some unspecified alternative model, the adjusted likelihood ratio is the appropriate statistic.

A likelihood ratio may be used to compare the evidence for any two models, a property that gives this approach to data analysis great flexibility. For example, a likelihood ratio can be used to compare the fit of the null to a specific effect size predicted by a particular theory, or to compare two different-sized effects based on two different models' predictions, as will be described towards the end of this tutorial.

Relation to Other Approaches

Likelihoodism is one of three basic approaches to statistical analysis, the other two being frequentist and Bayesian. However, both frequentist and Bayesian approaches are based on likelihood, and so likelihoodism shares some features with both, while also having important differences. As one example of a difference, whereas a p -value is based on an analysis of the probability of the data occurring if the null is true, and thus ignores the alternative model, a likelihood ratio directly compares the relative evidence for two competing models. By adopting a statistically symmetrical approach, the likelihood ratio provides a clearer index of the strength of the evidence for or against an effect than does a p -value.

The Bayesian approach is similar to likelihoodism in that it also involves model comparison. Indeed, a Bayes Factor is nothing more than a likelihood ratio adjusted by some prior distribution of parameter values. However, in contrast to a Bayesian, a likelihoodist eschews the use of a prior distribution to inform their analyses, focusing solely on the evidence provided by the data. The respective philosophies of the Bayesian and likelihood-based approaches differ thus because the likelihoodist applies their subjectivity at the end of the analysis. That is, the likelihoodist decides what to believe based on the evidence in

conjunction with their own intuitions about what may or may not be true, whereas the Bayesian attempts to mathematically formalize these prior beliefs into their statistical model. The objections of likelihoodists to the formalization of prior belief are detailed elsewhere (Edwards, 1972; Royall, 1997), and the interested reader is invited to view these sources for a discussion of some of the conceptual and mathematical issues that make statistical modelling of one's prior beliefs unattractive to a likelihoodist.

As a parable comparing the three basic approaches to data analysis, imagine three detectives are asked to investigate a murder with two possible suspects, Mr. Null and Ms. Alternative, and report the outcome of their analysis. The first detective, a frequentist trained in null hypothesis significance testing, would only examine the evidence against Mr. Null, and if this evidence suggested it seemed quite improbable that Mr. N were guilty, the detective would infer that Ms. A must have committed the foul deed ($p < .05$).

A second detective trained in the Bayesian method would begin their investigation by first assigning a prior probability to each suspect's guilt. They would do this as a matter of procedure, regardless of how much or little information regarding the case they might have. If based on actual evidence, this prior probability might be weighted in favor of either Mr. Null or Ms. Alternative, and might under appropriate circumstances form a reasonable starting point. If based on no evidence, however (the "uninformed prior"), this prior probability might be neutral or biased, specific or vague. Regardless of how defensible their prior probability might be, the manner in which it is mathematically formalized will have an impact on how the Bayesian detective ultimately presents the evidence.

Finally, the detective trained in likelihoodism would begin with no prior probabilities, but simply describe the evidence against both Mr. Null and Ms. Alternative, and compare the relative probability (likelihood) of each one's guilt. By examining the evidence against *both*

suspects, without introducing any prior bias into their calculations, the likelihoodist detective would arguably give the most objective report of all three investigators regarding which suspect was more likely to be the culprit, based on the data alone.

This objectivity - the fair and even appraisal of the two “suspects” - is in my view the core advantage of using likelihood ratios over the frequentist and Bayesian methods. Of course, this same objectivity also applies when the appraisal of evidence is concerning two hypotheses or models.

Mathematical relation between Likelihood Ratios and p -values

Despite using a different approach to model testing, likelihood ratios are typically closely related to p -values. Thus, a data set that gives a large likelihood ratio will also return a small p -value, and vice-versa. In most prototypical hypothesis testing scenarios, an approximate transformation of a (two-tailed) p -value to an adjusted likelihood ratio is:

$$\lambda_{adj} \approx \frac{1}{7.4 p} \quad (2)$$

As such, $p = 0.05$ will normally correspond to $\lambda_{adj} \approx 2.7$, $p = 0.01$ will correspond to $\lambda_{adj} \approx 13.5$, and $p = 0.001$ will correspond to $\lambda_{adj} \approx 135$. Thus, p -values can also be viewed as describing the strength of the evidence, as noted by Fisher (1955), but do so only indirectly through their relation to likelihood (Dixon, 1998; Lew, 2013).

Computing Likelihood Ratios

A likelihood ratio can generally be computed from the same statistics used to compute a p -value. The remainder of this tutorial provides several examples of these calculations, including ones based on t -scores, ANOVA outputs, *chi-square* statistics, and binomial tests. A brief description of a model comparison application based on models that don't rely on the maximum likelihood estimate is also provided; this would commonly be used to test two competing models that make *a priori* more specific predictions about the data than simply the presence or absence of an effect. Finally, personal views on interpreting likelihood ratios, and on the importance of methodological and statistical rigor in data collection and analysis, are provided. From here on, I recommend that interested readers experiment with likelihood ratios as they go through the tutorial, to get a feel for the statistic and how it relates to their intuitive sense of the data, as well as how it relates to other statistics they may have more experience with.

Likelihood Ratios from t-scores

A t -test comparing two means can easily be converted into a “raw” likelihood ratio using the equation:

$$\lambda = \left(1 + \frac{t^2}{df}\right)^{\frac{n}{2}} \quad (3)$$

where df is the degrees of freedom of the test, and n is the total number of observations. This basic formula applies universally to t -scores obtained from independent samples, paired samples, and single-sample tests. However, note that this equation is only the raw likelihood

ratio, as it is based solely on the frequency distributions of the maximum likelihood estimate and the null. As the reader may recollect from earlier, one must often apply an adjustment to a raw likelihood ratio because the data will almost always fit the model with more parameters better than the model with fewer parameters, resulting in overfitting (Burnham & Anderson, 1998). Failure to adjust for overfitting will result in likelihood ratios biased towards the more complex model. In a t -test, for example, the null model includes two parameters: the variance, and a single value for the overall mean. In contrast, the alternative model includes three parameters: a separate mean for each of the two experimental groups, plus the variance. An adjustment for overfitting which works well for linear models is the Akaike Information Criterion, or AIC (Akaike, 1973):

$$AIC = 2k - 2\ln(\lambda), \quad (4)$$

where k is the number of parameters in the model. Transposing the equation, we get an AIC-based adjustment to the raw likelihood ratio:

$$\lambda_{adj} = \lambda [\exp(k_1 - k_2)] \quad (5)$$

where k_1 and k_2 are the number of parameters in the less and more complex models, respectively.

A more detailed correction exists that also adjusts for sample size was provided by Hurvich and Tsai (1989; cf. Glover & Dixon, 2004):

$$\lambda_{adj} = \lambda \left(\exp \left[k_1 \left(\frac{n}{n-k_1-1} \right) - k_2 \left(\frac{n}{n-k_2-1} \right) \right] \right) \quad (6)$$

where k_1 and k_2 are again the number of parameters in the less and more complex models, respectively.

The Hurvich and Tsai adjustment converges towards the AIC adjustment as n increases such that the differences grow continuously smaller as n rises from 25 upwards, and the adjustments provided by the two methods become quite similar once $n = 100$. The reader is encouraged to experiment with both adjustments, but in general I would recommend the Hurvich and Tsai adjustment when $n < 25$, and the computationally simpler AIC adjustment when n is 25 or higher. For ease of exposition, and as all of the exercises in this tutorial involve sample sizes of 25 or greater, I will be using the AIC throughout.

Applying the AIC to the likelihood ratio for a t -test (Eq. 3), we arrive at the adjusted likelihood ratio formula for the t -test, λ_{adj} :

$$\lambda_{adj} = \left(1 + \frac{t^2}{df} \right)^{\frac{n}{2}} [\exp(-1)] \quad (7)$$

Note here that the AIC adjustment for the t test reduces to $\exp(-1)$ because there is one fewer parameter in the null model than in the alternative.

For an example of how to compute an adjusted likelihood ratio based on the t statistic, imagine an experimenter interested in the effects of a distractor on reaction time. They

perform an experiment in which one group of participants responds to a target appearing alone in the visual field, whereas the other responds to that same target appearing amongst multiple distractors. Data from this imaginary experiment are presented in Figure 2 and Table 1.

Figure 2. Results of an imaginary experiment examining the effects of a distractor on reaction times. Error bars represent standard errors of the means.

Table 1. Data from Distractor Experiment

		Distractor (n = 25)	Control (n = 25)
Reaction time	mean	285	250
	std	82.11	76.84
	t (48)	2.30	
	p	< 0.05	

Inserting the relevant values into Eq. 7, we get:

$$\lambda_{adj} = \left(1 + \frac{2.30^2}{48}\right)^{\frac{50}{2}} [\exp(-1)]$$

$$= 5.02$$

Thus, the data are about five times as likely assuming distractors had an effect on reaction times than assuming distractors had no effect. If the above data had instead come from a repeated measures design with one group of $n = 25$, the formula would remain the same, but the df and n would be different:

$$\lambda_{adj} = \left(1 + \frac{2.30^2}{24}\right)^{\frac{25}{2}} [\exp(-1)]$$

$$= 4.44$$

Here, the adjusted likelihood ratio is marginally smaller than for the same t-score with an independent-samples design, but note this comes with a large savings in n due to it being repeated measures. (Also, variance will typically be lower in a repeated-measures design than in an independent samples design).

Likelihood Ratios from ANOVA Outputs

Likelihood ratios can also be calculated from the data obtained from an ANOVA. In these situations, a universally applicable approach is to use the following equation:

$$\lambda_{adj} = \left[\left(\frac{\text{unexplained variance}_{M_1}}{\text{unexplained variance}_{M_2}} \right)^{\frac{n}{2}} \right] \exp(k_1 - k_2) , \quad (8)$$

where M_1 and M_2 are the simpler and more complex models, respectively, the unexplained variance is the total sum of squares not accounted for by each model, and n is the total number of observations. Note also the presence of the AIC correction of $\exp(k_1 - k_2)$ for any extra parameter(s) in the more complex model.

To illustrate how to calculate likelihood ratios from ANOVA data, imagine the researcher follows up their distractor and reaction time study by conducting an experiment that includes a second independent variable, hours of sleep (Figure 3 and Table 2). Here, one group of participants is allowed a full night's sleep prior to the testing session whereas the other is limited to three hours sleep. As well as trying to replicate the effect of distractor on reaction time, the researcher is also interested in examining the main effect of sleep, as well as the interaction.

Figure 3. Data from the imaginary follow-up experiment combining the effects of a distractor and sleep deprivation on reaction times. Error bars represent mean standard error of the pairwise differences.

Table 2. ANOVA Output from Distractor/Sleep Experiment

Source	df	SS	MS	F	<i>p</i>
Distractor	1	240	240	10.43	< 0.01
Sleep	1	260	260	11.30	< 0.01
D X S	1	95	135	5.87	< 0.05
Error	21	483	23		
Total	24				

The ANOVA table provides all the information needed to compute the likelihood ratios for each of the main effects and the interaction. To begin with, we will consider the main effect of distractor. The unexplained variance for the (null) model not including the distractor effect is found by adding together the sum of squares for the distractor with the error term ($240 + 483 = 723$), whereas the unexplained variance for the model representing the interaction is simply the error (483). The value for *n* is the 25 observations on which the effect is based (1 per subject). Entering these values into Eq. 8, we get:

$$\begin{aligned}\lambda_{adj} &= \left(\frac{723}{483}\right)^{\frac{25}{2}} [\exp(-1)] \\ &= 56.95\end{aligned}$$

The analysis shows that the data are about 57 times as likely given an effect of the distractor than no such effect. The researcher next calculates the λ_{adj} for the main effect of sleep by

substituting the relevant values from the sum-of-squares table. For sleep, this is done by substituting the unexplained variance for sleep ($260 + 483 = 743$) into the numerator (the denominator remains the same as before).

$$\lambda_{adj} = \left(\frac{743}{483}\right)^{\frac{25}{2}} [\exp(-1)]$$

$$= 80.10$$

This shows the data are about 80 times as likely to occur under a model assuming an effect of sleep than under the null model that sleep had no effect.

Finally, the researcher calculates the λ_{adj} for the interaction. Again, this involves substituting the relevant values for the sum-of-squares of the interaction into the numerator, and leaving the denominator unchanged.

$$\lambda_{adj} = \left(\frac{578}{483}\right)^{\frac{25}{2}} [\exp(-1)]$$

$$= 3.47$$

This shows that the data are roughly 3.5 times as likely given the interaction exists than given no interaction.

From these calculations we see that the data are much more likely assuming an effect of distractor than for the null model, and the same is true for the effect of sleep. Further, there is

some evidence for the interaction between these variables, though it is not nearly as compelling as was the evidence for the main effects. The fact that one can simply substitute the appropriate values into the equation to examine different effects show how easily these calculations can be done.

Note also how easy it is to compare results across different analyses. For example, the evidence for the effect of distractor was $\lambda_{adj} = 5.02$ in the first experiment, and $\lambda_{adj} = 56.95$ in the second. It is plain to see from this that the evidence was more than ten times stronger in the second experiment. Indeed, from all the examples given so far, we can observe that the adjusted likelihood ratio gives a more straightforward, yet still nuanced description of the evidence for an effect than does simply reporting a p -value of say, $p < .05$ or $p < .01$, as is commonly done. Moreover, one will also note the association between larger likelihood ratios and smaller p -values.

These are of course very basic, “cookbook” approaches to computing likelihood ratios from ANOVA data, and are meant only to provide an introduction to the approach. More principled and sophisticated calculations of likelihood ratios from ANOVA outputs can be found elsewhere, including methods based on *a priori* predictions and mixed-model analyses (Bortolussi & Dixon, 2013), post hoc tests (Dixon, 2013), and analyzing contrasts and theoretically interesting effects (Glover & Dixon, 2004).

Likelihood Ratios from Chi-Squares

A χ^2 test for goodness of fit is applied to categorical data, wherein the values in each cell are compared across two (or more) conditions. For this tutorial we will examine a simple two condition case, although the equation for computing the likelihood ratio will apply to all tests using χ^2 . The λ_{adj} for the χ^2 goodness of fit test is:

$$\lambda_{adj} = [\exp(0.5 \chi^2)] [\exp(k1-k2)] \quad (9)$$

Note again the use of the AIC value of $\exp(k1-k2)$ to arrive at the adjusted likelihood ratio.

To provide a real-world example of how one might use a likelihood ratio, consider the data from Table 3, which describes the chance of replication success depending on whether the p -value of the original study was either $p < .005$, or $p < .05$ but $> .005$.

Table 3. Replication success of experiments with $p < .005$ versus $.005 < p < .05$ (taken from Benjamin et al., 2018, based on data from Open Science Project, 2015).

<i>Criterion</i>	Replicated	Failed to replicate
$p < .005$	23	24
$.005 < p < .05$	11	34

The χ^2 test comparing the two criteria in terms of replication success returns a value of $\chi^2(1) = 5.92, p = .015$. Ironically and somewhat bemusingly, this may or may not support the argument that lowering the threshold for statistical significance will improve replicability, depending on which criteria one adopts. By the $p < .005$ criterion, it fails to provide evidence that $p < .005$ is better, whereas by the $p < .05$ criterion, the obtained value implies that $p < .05$ is worse.

Whereas this paradoxical result nicely highlights the absurdity of null hypothesis significance testing, it is also interesting to examine what the adjusted likelihood ratio says about the data.

Inserting the χ^2 value of 5.92 into Eq. 9, we get:

$$\begin{aligned}\lambda_{adj} &= [\exp (.5 * 5.92)] [\exp (-1)] \\ &= 7.10\end{aligned}$$

In other words, the data are roughly seven times as likely on the assumption that experiments reporting $p < .005$ were more likely to replicate. Of course, this does not in itself explain the various factors that might result in more replications of smaller p -values (see, e.g., Lakens, Adolphi, Albers et al., 2018).

Likelihood Ratios from Binomial Tests

A binomial test is another means of analysing categorical data, but in which only one group is tested. The common example used to illustrate the logic of the binomial test is a series of coinflips: if the coin is fair, the outcome over a series of trials ought to follow a binomial distribution centred on $p(\text{heads}) = .5$. The standard NHST approach to a binomial test is to reject the hypothesis that the coin is fair if the outcome falls far enough into one or the other tail of this distribution ($p < .05$). A likelihood ratio, however, can be used to directly compare the relative strength of any two hypotheses about the probability of a particular outcome. The adjusted likelihood ratio for a binomial test is computed as:

$$\lambda_{adj} = \left(\frac{p(x_{obs})}{p(x_{null})} \right)^{n_x} \left(\frac{p(y_{obs})}{p(y_{null})} \right)^{n_y} [\exp(k_1 - k_2)] \quad (10)$$

where $p(x_{obs})$ and $p(y_{obs})$ refer to the probability of observing a “success” or “failure” result, respectively, $p(x_{null})$ and $p(y_{null})$ are the probability of a success or failure outcome under

the null model, and nx and ny are the actual number of observed successes and failures respectively.

To illustrate, suppose a researcher is interested in whether participants are able to subconsciously learn to categorise stimuli based on subliminal rule learning. Imagine the participants are first shown two classes of subliminal stimuli and instructed to press one of two buttons on each trial, associated with the category (C1 vs. C2) each stimulus falls into. Following this, they are shown visible images, and required to press the appropriate button depending on which category they believe the stimulus belongs in. Suppose that out of 500 trials, participants correctly identify the category 275 times. To test whether this is better than chance performance, we enter the relevant values into Eq. 10 and find:

$$\begin{aligned}\lambda_{adj} &= \left(\frac{.55}{.50}\right)^{275} \left(\frac{.45}{.50}\right)^{225} [\exp(-1)] \\ &= (1.1)^{275} (.9)^{225} [\exp(-1)] \\ &= 4.50\end{aligned}$$

Thus, the outcome of 275 successes out of 500 trials is 4.5 times as likely given performance exceeds chance levels than that it was at chance. This corresponds to a (one-tailed) $p = .014$ for the same outcome.

An adjusted likelihood ratio will sometimes provide evidence for the null model (this is a general property of likelihood ratios), something that is not possible when using NHST. For example, if only 255/500 successes were recorded, the adjusted likelihood ratio for the model assuming performance was better than chance becomes:

$$\begin{aligned}\lambda_{adj} &= \left(\frac{.51}{.50}\right)^{255} \left(\frac{.49}{.50}\right)^{245} [\exp(-1)] \\ &= .4067\end{aligned}$$

which corresponds to (one-tailed) $p = .344$. As this value is less than one, it actually favours the null model. The inverse of the λ_{adj} for the model that performance exceeded chance will be the likelihood ratio favouring the null model that performance was at chance levels:

$$\lambda_{adj}(null) = \frac{1}{\lambda_{adj}} \quad , \quad (11)$$

where $\lambda_{adj}(null)$ is used to denote that the adjusted likelihood ratio is in favour of the null model. Inserting the values from above we get:

$$\begin{aligned}\lambda_{adj}(null) &= \frac{1}{.4067} \\ &= 2.51\end{aligned}$$

Thus, the experiment in which only 255 successes were recorded out of 500 trials is about 2.5 times as likely given chance performance than given performance better than chance.

Testing for Theoretically Interesting Effects

The previous example showed how the adjusted likelihood ratio can provide some evidence for the null hypothesis compared to a model based on the maximum likelihood estimate.

Stronger evidence for the null can sometimes be obtained if one bases their alternative hypothesis on a value that does not coincide with the MLE, and in fact, the test can be done using the effects predicted by any two models (Figure 4). Here, I will provide some more examples to illustrate how these procedures can be applied using the t -test or the binomial test to examine the evidence for a theoretically interesting effect (the procedure using ANOVA output is described elsewhere, Glover & Dixon, 2004, p. 800-801; Glover, 2018).

Figure 4. Conceptual illustration of the procedure for testing two models that predict specific effect sizes other than the maximum likelihood estimate (MLE). The grey curve represents the distribution of the observed data, the two blank curves represent the respective distributions of the two models (M1 and M2) being tested. In this example, the likelihood ratio is $\lambda = 10.0$ in favor of M1 over M2.

First, let us reconsider the t -test data from Table 1, where a 35 msec effect of a distractor on reaction times was observed. The adjusted likelihood ratio for this effect was $\lambda_{adj} = 5.02$, meaning the data were about five times as likely given an effect of 35 msec than no effect. Now let's imagine that we are interested in comparing two competing theories, one of which predicts no effect, and one of which predicts an effect of 90 msec. Which of these two models do the data support?

To answer this question, we must consider the extent to which the observed data deviate from the effect predicted by each model. Naturally, if the observed effect is 35 msec, it is typically going to be more likely under the model in which the true effect is 0 msec than the one that predicted an effect of 90 msec. However, this alone does not tell us how *much* more likely the data are given one model vs. the other.

We can test these two models directly against each other by calculating a likelihood ratio based on what is referred to as a “theoretically interesting effect” (TIE, cf. Glover, 2018). This might be an effect that is predicted by a specific theoretical model, or simply one that is the minimum size to be considered noteworthy. For analysing the evidence for a theoretically interesting effect based on the results of a t -test, we first must determine the value of t that indexes the extent to which the data deviate from that effect. This can be calculated using the obtained t -score as follows:

$$t(tie) = t(obs) - t(obs) \left(\frac{TIE}{obs} \right) \quad (12)$$

where $t(obs)$ is the t -score obtained from the original analysis ($t = 2.30$ in this case), and TIE and obs are the size of the theoretically interesting effect and the observed effect, respectively. With a TIE of 90 msec and an obs of 35 msec, we get:

$$\begin{aligned} t(tie) &= 2.30 - (2.30) \left(\frac{90}{35} \right) \\ &= -3.61 \end{aligned}$$

We now have two separate t -scores: First, the $t(obs)$ of 2.30 indexes the extent to which the data deviate from the null model that predicted the effect was 0 msec (the original score from the analysis in Table 1, $t = 2.30$). Second, the $t(tie)$ of -3.61 indexes the deviation of the data from the TIE. Calculating the likelihood ratio for these two models involves algebraically incorporating both these scores into Eq. 3:

$$\lambda(tie \text{ vs. } null) = \left(\frac{[1 + \frac{t(obs)^2}{df}]}{[1 + \frac{t(TIE)^2}{df}]} \right)^{\frac{n}{2}} \quad (13)$$

where $\lambda(tie \text{ vs. } null)$ is the likelihood ratio in favor of the TIE model versus the null. In this case, the AIC adjustment reduces to $\exp(0) = 1$, as both models are fixed in terms of their means and so have an equal number of parameters. Thus, the AIC adjustment is superfluous and we can simply report the raw likelihood ratio, λ . Substituting the values in for $t(obs)$ and $t(TIE)$, we get:

$$\begin{aligned}\lambda(tie \text{ vs. } null) &= \left(\frac{[1 + \frac{2.30^2}{48}]}{[1 + \frac{-3.61^2}{48}]} \right)^{25} \\ &= .000592\end{aligned}$$

Or inversely, $\lambda(null \text{ vs. } tie) = 1668.7$. Here, the data are > 1600 times as likely given no effect than given an effect of 90 msec.

Of course, it also possible for the TIE procedure to find evidence for the TIE over the null.

For example, if the TIE in the above case were 45 msec rather than 90 msec the $t(tie)$ would be $-.657$, and the resulting likelihood ratio would be:

$$\begin{aligned}\lambda(tie \text{ vs. } null) &= \left(\frac{[1 + \frac{2.30^2}{48}]}{[1 + \frac{-.657^2}{48}]} \right)^{25}, \\ &= 10.9\end{aligned}$$

or about 11:1 in favor of the TIE over the null.

This general procedure has wider applications than simply testing a TIE versus a null model.

It can also be used to compare the fit of any two models that predict an effect of a different magnitude. Let us examine the idea of testing two different models using the binomial data

from the subliminal perception study described above, in which we observed 275/500

successes. Here, imagine that Model A predicted success on 57% of the trials, whereas Model

B predicted success on 52% of trials. The computation of the likelihood ratio in this case

involves comparing the relative likelihood of the observed 275/500 successes given either

model. This is done by re-arranging the formula for the binomial test (Eq. 10) to test the two

models directly against each other, as follows:

$$\lambda = (\text{Model A vs. Model B}) = \left(\frac{p(xA)}{p(xB)} \right)^{nx} \left(\frac{p(yA)}{p(yB)} \right)^{ny} \quad (14)$$

where $p(xA)$ and $p(xB)$ refer to the predicted probabilities of observing a success based on Models A and B; $p(yA)$ and $p(yB)$ are the predicted probabilities of a failure based on those same models; and nx and ny are the actual number of observed successes and failures. As before, because there is no difference in the number of parameters between Models A and B, the AIC adjustment reduces to 1 and can be dropped, leaving us with a raw likelihood ratio, λ . Note also that either model could be tested separately against the null model by simply substituting the null model's values into the equation in the place of the other model. Solving this equation for the Models A and B, we get:

$$\begin{aligned} \lambda (\text{Model A vs. Model B}) &= \left(\frac{.57}{.52} \right)^{275} \left(\frac{.43}{.48} \right)^{225} \\ &= 1.64 \end{aligned}$$

The result here is rather equivocal, showing that the outcome of 275/500 successes is only about 1.6 times as likely given Model A that predicted a 57% success rate versus Model B that predicted a 52% success rate.

Interpreting Likelihood Ratios

Many scientists appear to want a statistical analysis to give them a clear, “yes/no” answer, to be able to believe (or at least argue) strongly that an effect is either present or not present based on a p -value, Bayes Factor, or likelihood ratio. I feel this is an unproductive approach to scientific inference and represents a fundamental misapprehension of the information such

statistics contain. Any statistic used as an index of evidence will never provide an absolute black-or-white answer to the question of which of two possible interpretations of the data is correct, but can at best only offer an estimate in some shade of grey. Sometimes the shade is darker, sometimes lighter, and sometimes in the middle. It may be a hard truth to accept, but it is a truth nonetheless that when one deals with estimates one is dealing with uncertainty. Furthermore, the quality of the estimate itself can obviously be affected by issues of methodological rigour (Cohen, 1994; Gigerenzer, 2004; Greenland, Senn, Rothman, et al., 2016; Simmons, Nelson, & Simonsohn, 2011). Thus, regardless of whether one is frequentist, Bayesian, or likelihood-based in their approach, one should always interpret their statistics with a healthy amount of scepticism. Getting to the correct answer requires careful evaluation, replication, and sometimes intuition and common sense. These latter factors are important if often neglected elements of scientific inference.

Multiple Shades of Grey

With these caveats in mind I am unwilling to suggest assigning different likelihood ratios to categories such as “weak”, “moderate”, or “strong” evidence. Instead, I suggest researchers use their own judgement in interpreting likelihood ratios, and consider what kind of evidence they themselves require to be convinced, and whether that same evidence would also convince a sceptic. Further, I suggest researchers appreciate that no matter how much information they may have about the presence and/or size of an effect, having more information is always better, and can help to either darken or lighten the shades of grey inherent in statistics. Finally, careful parametrization and methodological rigor are fundamental to statistical analysis, and easily more important than which statistic you choose to analyze your data (cf. Wasserstein & Lazar, 2016).

Conclusions

In this tutorial, I outlined the logic of likelihood ratios and compared the likelihood-based approach to frequentist and Bayesian approaches, and argued for the intuitive appeal of likelihood ratios as an objective, clear index of the evidence for two statistical models based solely on the data at hand. I showed how to compute likelihood ratios from many common statistics, and also how to adapt likelihood ratios to test for models of different effect sizes than ones based on the maximum likelihood estimate, such as theoretically interesting effects. Finally, I offered advice on how to interpret likelihood ratios, encouraging scientists to use their reason and common sense, to employ methodological and statistical rigor, and to accept uncertainty as a part and parcel of scientific inference.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrove & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest: Academiai Kiado.
- Benjamin, D., Berger, J., Johannesson, M. et al. (2017). Redefine statistical significance. *Nature Human Behaviour*. In press.
- Bortolussi, M., & Dixon, P. (2003). *Psychonarratology: Foundations for the empirical study of literary response*. Cambridge University Press.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: A practical information-theoretic approach*. New York: Springer.
- Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- Dixon, P. (1998). Why scientists value p values. *Psychonomic Bulletin and Review*, 5, 390-396.
- Dixon, P. (2013). The effective number of parameters in post hoc models. *Behavior Research Methods*, 45, 604-612.
- Edwards, A. W. (1992). *Likelihood*. Johns Hopkins University: Baltimore.
- Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B*, 17, 69-78.
- Gigerenzer. (2004). Mindless statistics. *Journal of Socio-Economics*, 33, 587-606.
- Glover, S. (2018). Redefine statistical significance XIV: “Significant” does not necessarily mean “interesting.” <https://www.bayesianspectacles.org/redefine-statistical-significance-xiv-significant-does-not-necessarily-mean-interesting/>
- Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin and Review*, 11, 791-806.

- Goodman, S. N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health*, 78, 1568-1574.
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31, 337-350.
- Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.
- Lakens, D., Adolphi, F. G., Albers, C. A., et al. (2018). Justify your alpha. *Nature Human Behavior*, 2, 168-171.
- Lew, M. J. (2013). To P or not to P: on the evidential nature of P-values and their place in scientific inference. <https://arxiv.org/abs/1311.0081>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, 1-8.
- Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman and Hall.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science*, 22, 1359-1366.
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p -values: context, process, and purpose. *The American Statistician*, 70, 129-133.