p-Values and the Likelihood Principle

Andrew Neath*

Abstract
The p-value is widely used for quantifying evidence in a statistical hypothesis testing problem. A major criticism, however, is that the p-value does not satisfy the likelihood principle. In this paper, we show that a p-value assessment of evidence can indeed be defined within the likelihood inference framework. The connection between p-values and likelihood based measures of evidence broaden the use of the p-value and deepen our understanding of statistical hypothesis testing.

Key Words: hypothesis testing, statistical evidence, likelihood ratio, Wald statistic

1. Introduction

The p-value is a popular tool for statistical inference. Unfortunately, the p-value and its role in hypothesis testing is often misused in drawing scientific conclusions. Concern over the use, and misuse, of what is perhaps the most widely taught statistical practice has led the American Statistical Association to craft a statement on behalf of its members (Wasserstein and Lazar, 2016). For statistical practitioners, a deeper insight into the workings of the p-value is essential for an understanding of statistical hypothesis testing.

The purpose of this paper is to highlight the flexibility of the p-value as an assessment of statistical evidence. An alleged disadvantage of the p-value is its isolation from more rigorously defined likelihood based measures of evidence. However, this disconnect can be bridged. In this paper, we present a result establishing a p-value measure of evidence within the likelihood inferential framework.

In Section 2, we discuss the general idea of statistical evidence. In Section 3, we consider the likelihood principle and establish the aforementioned connection with the p-value. We close the paper in Section 4 with some concluding remarks on how the p-value plays a role in a broader class of hypothesis testing problems than may be currently appreciated.

2. The p-value and evidence

Before going any further, let's take a moment to think about what is meant by statistical evidence. Let's think of a researcher collecting data on some natural phenomenon in order to determine which of two (or more) scientific hypotheses is most valid. Data favors a hypothesis when that hypothesis provides a reasonable explanation for what has been observed. Conversely, data provides evidence against a hypothesis when what has been observed deviates from what would be expected. Scientific evidence is not equivalent to scientific belief. It is not until multiple sources of data evidence favor a hypothesis that a foundation of strong belief is built. Because belief arises from multiple researchers and multiple studies, the language for communicating an advancement of scientific knowledge is the language of evidence. Thus, quantification of evidence is a core principle in statistical science.

R.A. Fisher is credited with popularizing the p-value as an objective way for investigators to assess the compatibility between the null hypothesis and the observed data. The

*Southern Illinois University Edwardsville, Department of Mathematics and Statistics, Edwardsville, IL, aneath@siue.edu

Table 1: p-value scale of evidence

p	evidence against H_o
.10	borderline
.05	moderate
.025	substantial
.01	strong
.001	overwhelming

p-value is defined as the probability, computed under the null hypothesis, that the test statistic would be equal to or more extreme than its observed value. While the p-value definition is familiar to statistical practitioners, a simple example may help focus on the idea of quantifying evidence. Consider a scientist investigating a binomial probability θ. The goal is to test $H_o : \theta = 1/2$ against a lower tail alternative $H_1 : \theta < 1/2$. So, $X \sim B(n, 1/2)$ under the null hypothesis. In $n = 12$ trials, $x = 3$ successes are observed. Since small values of X support the alternative, the p-value is computed to be

$$\begin{aligned} p &= P_o\left(X \leq 3\right) \\ &= \sum_{i=0}^{3} \binom{12}{i} (1/2)^{12} = .0730 \end{aligned}$$

The null hypothesis is most compatible with data near the center of the null distribution. Data incompatible to the null distribution is then characterized by a small p-value. In this way, the p-value serves as an assessment of evidence against the null hypothesis.

The p-value is a probabilistic measure of evidence, but not a probabilistic measure of belief. The desire to interpret p as a probability on the null hypothesis must be resisted. But this leaves open the question of how to represent a p-value scale of evidence. Fisher recommended the scale displayed in Table 1 (Efron, 2013).

The Fisher scale seems to be consistent with common p-value interpretations. For our simple example, we can say there is moderate to borderline evidence against the null hypothesis. In the end, the choice of an appropriate evidence scale should depend on the underlying science, as well as an assessment of the costs and benefits for the application at hand (Gelman and Robert, 2013). Particularly troublesome to the goal of improving scientific discourse is a blind adherence to any threshold separating significant and non-significant results.

A perceived shortcoming of the p-value as an assessment of evidence can be illustrated from our simple example. Note that the p-value is not only a function of the data observed $(x = 3)$, but of more extreme data that has not been observed $(x < 3)$. The definition of the p-value as a tail probability implies that the computation of p depends on the sampling distribution of the test statistic. So, the p-value depends on the, perhaps irrelevant, intentions of the investigator, and not merely on the data observed. In this way, the p-value is in violation of the likelihood principle. We will see in the next section, however, that a p-value measure of evidence can be defined to satisfy the likelihood principle. With this result, a major criticism of the p-value is answered.

3. Likelihood inference

We will take a relatively informal approach in our introduction to likelihood inference. Readers interested in a more rigorous treatment are encouraged to consult Pawitan (2013) and Berger, Wolpert (1988). Simply put, the likelihood principle requires that an evidence measure satisfy two conditions: sufficiency and conditionality. The sufficiency condition states that evidence depend on the data only through a sufficient statistic. The p-value has no real issue in that regard. The conditionality condition states that evidence depend only on the experiment performed, and the data observed; not on the intention of the investigator. To see that the p-value fails in this regard, we return to the simple binomial example. Suppose instead of a predetermined sample size $n = 12$, the scientist's intention was to sample until $x = 3$ successes were observed. Under this scenario, the number of trials N is a random variable. Under the null hypothesis, $N \sim NB(3, 1/2)$. Since large values of N support the lower tail alternative, the p-value is computed to be

$$\begin{aligned} p &= P_o(N \geq 3) \\ &= \sum_{i=12}^{\infty} \binom{i-1}{2} (1/2)^i = .0327 \end{aligned}$$

Now, we have moderate to substantial evidence against the null. Equivalent hypotheses, tested from equivalent data, reach different levels of evidence. Computation of the p-value is not invariant to the sampling scheme, even though the plan to collect the data is unrelated to the evidence provided from what is actually observed. That an unambiguous p-value assessment does not seem to be available is a problem we will address.

The development of an evidence measure which does satisfy the likelihood principle proceeds as follows. Let $L(\theta)$ denote the likelihood as a function of an unknown parameter θ. (For simplicity, we take the single parameter case. Nuisance parameters and parameter vectors can be handled with slight adjustments to the development.) Let $\widehat{\theta}$ denote the maximum likelihood estimate. We consider the problem of testing the null hypothesis $H_o : \theta = \theta_o$ under the likelihood inference framework. Define the likelihood ratio as $LR(\theta_o) = L(\theta_o)/L\left(\widehat{\theta}\right)$. Then $0 < LR(\theta_o) < 1$. As $LR(\theta_o)$ decreases, data evidence against the null hypothesis increases. In this sense, $LR(\theta_o)$ provides a measure of evidence against the null hypothesis in the same spirit as a p-value.

We return once more to the binomial data. The likelihood ratio is invariant to sampling scheme. So, the measure of evidence is the same whether the data comes from a binomial or negative binomial. Write

$$LR(\theta) = \frac{\theta^x (1-\theta)^{n-x}}{\widehat{\theta}^x \left(1-\widehat{\theta}\right)^{n-x}}$$

where the sample proportion $\widehat{\theta} = x/n$ is the maximum likelihood estimate. For testing $\theta_o = 1/2$ with observed data $x = 3, n = 12$, we compute $\widehat{\theta} = 1/4$ and $LR(\theta_o) = .208$. We can say the data supports the null value at about 20% of the level of support to the maximum likelihood estimate. But while we are successful in creating a measure of evidence which satisfies the likelihood principle, we have lost the familiarity of working with a probability scale.

It would be desirable to calibrate a likelihood scale for evidence with the more familiar p-value scale. We can achieve this goal directly when the likelihood function is of the regular case. Let $l(\theta) = \ln L(\theta)$ denote the log-likelihood, and write its Taylor expansion

as

$$l(\theta) = l\left(\widehat{\theta}\right) + l'\left(\widehat{\theta}\right) \cdot \left(\theta - \widehat{\theta}\right) + \frac{l''\left(\widehat{\theta}\right)}{2} \cdot \left(\theta - \widehat{\theta}\right)^2 + \dots$$

The regular case occurs when the log-likelihood can be approximated by a quadratic function. Asymptotics for maximum likelihood estimators are derived under the conditions leading to the regular case. Since $l'\left(\widehat{\theta}\right) = 0$, then

$$l(\theta) \approx l\left(\widehat{\theta}\right) - \frac{1}{2}\frac{\left(\theta - \widehat{\theta}\right)^2}{\widehat{\sigma}^2}$$

where $\widehat{\sigma}^2 = -1/l''\left(\widehat{\theta}\right)$ is the reciprocal of the observed Fisher information $\widehat{FI} = -l''\left(\widehat{\theta}\right)$. We can then write the likelihood function at the null value θ_o as

$$L(\theta_o) \approx L\left(\widehat{\theta}\right) \cdot \exp\left\{-\frac{1}{2}z^2\right\}$$

where $z = \left(\widehat{\theta} - \theta\right)/\widehat{\sigma}$ is the Wald statistic for testing $H_o : \theta = \theta_o$. The likelihood ratio statistic becomes

$$LR(\theta_o) \approx \exp\left\{-\frac{1}{2}z^2\right\}. \tag{1}$$

Let's introduce a second example to demonstrate the approximation in (1). In a well-known example of data collection (MacKenzie, 2002), a statistics class experimented with spinning the newly minted Belgian Euro. Spinning instead of tossing a coin is more sensitive to unequal weighting of the sides. In $n = 250$ spins, $x = 140$ landed heads side up. Now, the intended sampling scheme is not at all clear from the summary provided. But quantifying evidence through the likelihood ratio statistic renders the question of experimenter intention unimportant. We have $\widehat{\theta} = .56$ and $\widehat{\sigma} = .0314$. For testing $H_o : \theta = .5$, we get $z = 1.91$. From (1), we compute the approximation $LR(\theta_o) \approx 0.161$. The exact value of the likelihood ratio statistic is computed as

$$LR(\theta_o) = \frac{(.5)^{140}(.5)^{110}}{(.56)^{140}(.44)^{110}} = 0.165$$

The use of z in approximating $LR(\theta_o)$ connects the Wald statistic to the likelihood ratio statistic. A z statistic also leads directly to the calculation of a p-value. Since $LR(\theta_o)$ depends on the data through test statistic z alone, then $LR(\theta_o)$ is a function of the corresponding p-value. Therefore, in the regular case, one can define a p-value which does satisfy the likelihood principle. No matter the intended sampling scheme in our example, the p-value for a two-sided alternative is seen from the computed Wald statistic to be $p = .056$.

We will extend the connection between a likelihood ratio statistic and a p-value to a more general case. Before that, let's think about some consequences of the regular case. We note that the development could proceed from the asymptotics of the likelihood ratio statistic directly. The Wald statistic z appears naturally in the regular case, so no extra difficulty is caused by its consideration. Since the likelihood function is invariant to sampling scheme, so is the Wald statistic. Specifically, the standard error $\widehat{\sigma}$ does not depend on the underlying sampling distribution. Let's demonstrate this by comparing the binomial and negative binomial sampling distributions. In both cases, $\widehat{\theta} = x/n$. In the binomial setting, X is the random variable with $Var\left(\widehat{\theta}\right) = \theta(1-\theta)/n$. The estimated variance becomes

$$\widehat{Var}\left(\widehat{\theta}\right) = \frac{\widehat{\theta}\left(1 - \widehat{\theta}\right)}{n} = \frac{x(n-x)}{n^3}.$$

Table 2: likelihood ratio scale of evidence

p (one-sided)	p (two-sided)	z	LR
.05	.10	1.645	.258
.025	.05	1.960	.146
.005	.01	2.326	.067
.0025	.005	2.576	.036

In the negative binomial setting, N is the random variable. Applying the delta method leads to the asymptotic variance $AVar\left(\widehat{\theta}\right) = \theta^2\left(1-\theta\right)/x$. The estimated variance here becomes

$$\widehat{AVar}\left(\widehat{\theta}\right) = \frac{\widehat{\theta}^2\left(1-\widehat{\theta}\right)}{x} = \frac{x\left(n-x\right)}{n^3}.$$

Thus, $\widehat{\sigma} = \sqrt{\widehat{Var}}$ is identical across sampling schemes. This property holds true whenever the likelihood belongs to the regular case. It is interesting to see that the variance parameter does depend on the sampling distribution. Test statistics based on evaluating the variance parameter at the null value are not invariant to the sampling scheme. An example of such a test statistic is the score statistic. Some prefer the score statistic in hypothesis testing because its error rate properties better approximate the stated levels (Agresti, 2013). However, a score statistic does not satisfy the likelihood principle. Under the Fisher viewpoint, the goal of hypothesis testing is to provide a statistical measure of evidence for the case at hand. Error rates for (hypothetical) repeated trials hold no sway under this philosophy (Hubbard and Bayarri, 2003). The Wald statistic would thus be preferable under the evidentiary view.

The arrangement which binds a p-value with the likelihood principle is beneficial to both schools of thought. As mentioned previously, the likelihood ratio scale for evidence lacks the familiarity of the p-value scale. The approximation in (1) allows one to more easily interpret a likelihood ratio. Translating z to both LR and p leads to an evidential equivalence displayed in Table 2.

A likelihood ratio near .15 is the evidential equivalent of a two-sided p-value near .05. The 1 in 20 rule applied to the likelihood ratio $(LR < .05)$ would translate to a more stringent rule than the $p < .05$ rule prevalent throughout much of statistical practice. Table 2 is our link between two seemingly disparate approaches to quantifying evidence.

We still need a way to unambiguously connect the p-value to the likelihood ratio for problems outside of the regular case. Evidence measured on the likelihood ratio scale is interpreted the same, whether from the regular case or not. Thus, we have an unambiguous measure of evidence against a null hypothesis $H_o : \theta = \theta_o$ on the likelihood ratio scale. We can read this in Table 2 as the right most column. The answer we are looking for can be found by reading Table 2 from right to left. For any likelihood ratio statistic, there exists a translated z statistic. Note that such a z statistic need not actually exist. We are only interested in the equivalence to some value on the evidence scale. We can then translate this z into a p-value measure of evidence. In other words, any likelihood ratio can be uniquely translated into a p-value. We thus have a p-value, or at least an evidential measure on the p-value scale, which satisfies the likelihood principle.

Let's demonstrate the computation of a likelihood based p-value by returning one last time to the simple binomial example. The likelihood function is not of the regular case, but that does not matter. Earlier in this section, we computed $LR\left(\theta_o\right) = .208$. We can connect

a likelihood ratio to a z statistic by solving (1) as

$$z = \sqrt{-2 \ln LR\left(\theta_o\right)}.$$

For our problem, we get $z = 1.77$. We can easily connect a z statistic to a p-value. Since the alternative hypothesis is one-sided, we can compute $LRp = .0384$. No matter the frequentist intention for the experiment, the calculations for LRp remain the same. The result is an unambiguous p-value calculation. One can use a p-value measure of evidence while adhering to the likelihood principle.

Any testing problem where evidence can be quantified through the likelihood function can also be quantified through a uniquely defined measure on the p-value scale. We can think of this measure as defining a p-value which does indeed satisfy the likelihood principle.

4. Concluding remarks

An understanding of what can be implied from hypothesis testing results is a necessary obligation for a conscientious scientist. There is much debate as to the role of the p-value in scientific reasoning and discussion. Criticism over the use of the p-value tends to focus on its deficiencies in comparison to more rigorously defined evidential measures. We have seen, however, that a p-value measure of evidence can be defined under the likelihood principle. The connection between p-values and likelihood based measures of evidence broaden the use of the p-value in statistical hypothesis testing. If one desires a quantification of evidence through the likelihood principle, a p-value can still be a useful instrument.

REFERENCES

Agresti, A. (2013), *Categorical Data Analysis*, Hoboken, NJ: Wiley.

Berger, J. and Wolpert, R. (1988), *The Likelihood Principle: A Review, Generalizations, and Statistical Implications*, Hayward, CA: Institute of Mathematical Statistics.

Efron, B. (2013), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge, UK: Cambridge University Press.

Gelman, A. and Robert, C. (2013), "Revised Evidence for Statistical Standards," *Proceedings of the National Academy of Sciences*, 111, 19.

Hubbard, R. and Bayarri, M. (2003), "p-Values are Not Error Probabilities," *Institute of Statistics and Decision Sciences, Working Paper*, No. 03-26.

MacKenzie, D. (2002), "Euro Coin Accused of Unfair Flipping," *New Scientist*, January 4.

Pawitan, Y. (2013), *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford, UK: Oxford University Press.

Wasserstein, R. and Lazar, N. (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129-133.