Hail the Impossible:

$p$-values, Evidence, and Likelihood
Abstract

Significance testing based on $p$-values is standard in psychological research and teaching. Typically, research articles and textbooks present and use $p$ as a measure of statistical evidence against the null hypothesis (the Fisherian interpretation), although using concepts and tools based on a completely different usage of $p$ as a tool for controlling long-term decision errors (the Neyman-Pearson interpretation). There are four major problems with using $p$ as a measure of evidence and these problems are often overlooked in the domain of psychology. First, $p$ is uniformly distributed under the null hypothesis and can therefore never indicate evidence for the null. Second, $p$ is conditioned solely on the null hypothesis and is therefore unsuited to quantify evidence, because evidence is always relative. Third, $p$ designates probability of obtaining evidence (given the null), rather than strength of evidence. Fourth, $p$ depends on unobserved data and subjective intentions and therefore implies, given the evidential interpretation, that the evidential strength of observed data depends on things that did not happen and subjective intentions. In sum, using $p$ in the Fisherian sense as a measure of statistical evidence is deeply problematic, both statistically and conceptually, while the Neyman-Pearson interpretation is not about evidence at all. In contrast, the likelihood ratio escapes the above problems and is recommended as a tool for psychologists to represent the statistical evidence conveyed by obtained data relative to two hypotheses.

Keywords: $p$-value, significance testing, evidence, error control, subjectivity, likelihood
Hail the Impossible:

*p*-values, Evidence, and Likelihood

“Alice laughed. 'There's no use trying,' she said. 'One can't believe impossible things.' 'I daresay you haven't had much practice,' said the Queen. 'When I was your age, I always did it half an hour a day. Why, sometimes, I've believed as many as six impossible things before breakfast.”

*Alice in Wonderland*. Lewis Caroll.

Every psychology researcher is familiar with significance testing based on *p*-values. Despite the controversies surrounding significance testing, *p*-values are still hailed in psychology, in the sense that they abound and flourish. In this article, I emphasize problems with *p*-values that are often overlooked within psychology. In psychological research *p*-values are typically taken as an index of statistical evidence, but this results in several major problems, not easily disposed of. Commonly suggested additions to *p*, such as effect size and confidence intervals, are not up to the task either. If the objective is to quantify statistical evidence (and judging from research articles, it often is) it is recommended that researchers use likelihood ratios (cf. Royall, 1997), rendering *p*-values more or less practically obsolete.

Significance testing based on *p*-values, and whether using it as an index of evidence hails the impossible, has been a subject of controversy for a long time (Berger & Sellke, 1987; Cornfield, 1976; Edwards, 1972; Krantz, 1999; Neyman & Pearson, 1933). Unfortunately, psychologists frequently misunderstand what *p*-values are and what they may consistently be used for (Gigerenzer, 1993; 2004; Haller & Krauss, 2002; Hubbard & Armstrong, 2006; Wagenmakers, 2007). When peculiarities about *p* are noted in the psychological literature, focus is often directed to the less interesting and less devastating properties or problems of significance testing, such that .05
is an arbitrary threshold (Rosnow & Rosenthal, 1989), or that \( p \) does not indicate effect size or the posterior probability of a hypothesis (Cohen, 1994).

It is surprising that textbooks of statistics in psychology do not incorporate more fundamental controversies of and alternatives to significance testing, such as likelihood theory and Bayesian statistics. If they did, students would be in a better position to understand what \( p \) is, what it is not, what it is based on, and how it could be used consistently. The exclusion of alternative fundamental statistical perspectives while relying overly on the classical ones is difficult to motivate rationally. Actual research is affected by this, because the eager young minds of today are the settled old minds of tomorrow. In psychology, we have practiced a lot to believe impossible things. Perhaps the time is ripe to be done with it. Perhaps the Queen should be overthrown.

There are three main parts in this article. First, significance testing and p-values are introduced and discussed along with the classic evidential interpretation of p-values. In this context, several serious problems with the p-value as an index of statistical evidence are discussed. Second, the relation between p-values, long-term decision error control, and statistical evidence is discussed. In this context, it is emphasized that decision error control is distinct from the strength of evidence in obtained data. Third and finally, likelihood ratios are introduced and illustrated as an alternative to the p-value in order to quantify statistical evidence.

**Significance Testing**

Significance testing involves the act of computing a p-value, evaluate this, and then things tend to get messy. Let us withhold the mess for a second and simply consider the act of computing a p-value.
Formally, a $p$-value is simply a certain part of the sampling distribution associated with a test statistic for a set of sampling assumptions. An example will help.

Suppose we conduct a memory test on 50 punk rock musicians who have not had a beer in five weeks. Their mean ($M$) score is 1.40. For simplicity (without loss of generality), assume that we know that among punk rock musicians in general (the ones who drink beer most of the time), scores on the test are distributed normally with a mean ($\mu$) of 0.00 and a standard deviation ($\sigma$) of 6.00.

Imagine that we repeatedly and infinitely sample 50 scores from the punk-rock-in-general-distribution and compute a mean score for each sample. By the central limit theorem, the resulting distribution of means will be normally distributed, have a mean $\mu_M = \mu = 0.00$, and a standard deviation $\sigma_M = \sigma/\sqrt{N} = 6.00/\sqrt{50} \approx 0.85$. The distribution of means is called the sampling distribution of the mean and its standard deviation is called the standard error of the mean.

We can compute, in terms of the standard error, how much our obtained mean $M = 1.40$ differs from the mean of the sampling distribution by calculating a $z$-score, $z = (M-\mu_M)/\sigma_M = (1.40-0.00)/0.85 \approx 1.65$. For a normal distribution we can determine the proportion of scores that fall at or beyond the obtained $z$-score of 1.65, which defines our $p$-value, $p = .049$ (one-tailed, in this case). This is shown in Figure 1.

So, if our beer-deprived punk rock musicians really were a sample from a normally distributed population with the same mean and standard deviation as the beer-drinking punk rock musicians, then the probability of observing a mean score of 1.40 or more positively extreme than 1.40 is .049. This may or may not make us inclined to think that punk rock musicians would nail the lyrics to a larger extent if they deferred their Heineken until after the show.

**What about $p$?**

What do we do with our $p = .049$? The standard consequence is to reject the null hypothesis which in our example, according to standard procedures, would be the statistical hypothesis that the
population mean for punk rockers who drink beer is the same as the population mean for punk rockers who do not drink beer. Why does this rejection of the null take place?

There are two standard answers to the Why question. One answer (the Fisherian) is that we reject the null because we have obtained evidence against the null. It would be unlikely to observe results such as ours \( (M = 1.40) \) or more extreme results \( (M > 1.40) \) if the null is true and therefore we have evidence against the null. The other answer (the Neyman-Pearsonian) is that we reject the null because our test statistic falls within a particular specified region of the sampling distribution which allows us to control the extent to which we make the wrong decision in the long run in an specific optimal way, provided that we have specified an alternative hypothesis as well (that we would have rejected had the test statistic not fallen within the specified region). Let us have a look at both answers.

\textbf{p and Evidence}

First, consider the Fisherian interpretation, associated with the work of sir Ronald Fisher. Although Fisher vacillated between somewhat different interpretations of \( p \) in the context of significance testing over the years (Gigerenzer, 1993), it is clear that most of the time Fisher regarded \( p \) as providing a measure of evidence against the null hypothesis. For example, we are told that “…the probability integral [i.e., the \( p \)-value]…would give, with any degree of accuracy required, the probability, on that hypothesis, that a greater discrepancy should occur than that actually observed. If the value of \( P \) so calculated turned out to be a small quantity such as 0.01, we should conclude with some confidence that they hypothesis was not in fact true of the population actually sampled.” (Fisher, 1925, p. 41). We are also told that “It is open to the experiment to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result” (Fisher, 1935, p. 13).

Fisher expressed the idea that it is the smallness of the probability of the data and more extreme
data under a given hypothesis (the null) that constitutes evidence against the hypothesis and is the *reason* for our potentially rejecting it. It is difficult not to consider Fisher’s view as an expression of *p* as a measure of evidence against the conditioned hypothesis and this view became even more outspoken later in his career when he suggested that researchers provide the exact *p*-value (rather than *p* < .05) in their reports (Fisher, 1956). Or, in the words of more recent scholars: “Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of *p*?” (Rosnow & Rosenthal, 1989).

For Fisher, the *exact* value of *p* matters to a great deal, because it is supposed to be a *property of the data* (and more extreme data) that carries *evidential value*. Although Fisher did consider a level of significance of .05 for *p* early on (Fisher, 1935), this was more of a conventional standard for a heuristic evidential cut-off and definitely not a mechanism to control decision error rates. Also note that the Bayesian route of arriving at posterior subjective probabilities (after the data comes in) of hypotheses by updating prior probabilities (before the data comes in) using the probability of the (obtained and not more extreme) data given a hypothesis was not open to Fisher, because he explicitly rejected Bayes theorem (Fisher, 1925).

Problems with interpreting and using *p* as a measure of evidence abound, almost to the extent to which such interpretation and usage abound in psychology. Let us consider four of them, as listed in Table 1.

**DEFCON 4: *p* and the null**

One problem which psychologists often do note, although not always for the right reasons, is that *p* cannot be used as evidence *for* the null hypothesis. Fisher (1935) was very much aware of this and the reason is that, for *p*-values to function the way they do, they have to be uniformly distributed when the null hypothesis is true (Rouder, Speckman, Sun, Morey, & Iversson, 2009). If the null is true, then *p* ≤ .05 is expected 5% of the time, *p* ≤ .10 is expected 10% of the time, *p* ≤ .15 is
expected 15% of the time and so on. In effect, if the null is true then all $p$-values (and all equally sized ranges of $p$-values) are equally probable, implying that you cannot use the largeness of $p$ as evidence for the null. The fact that $p$ is uniformly distributed under the null also implies that if you keep collecting data when the null is true and calculate (unadjusted) new $p$-values as the data accumulates, then you can find as small $p$-values as you wish for. All you need is patience, but sooner or later you will end up with, say, $p < .00001$.

Fisher rejected the concept of statistical power, so for Fisher the motivation for not using $p$-values as evidence for the null had nothing to do with statistical power. The motivation that many psychologists have for not accepting the null is, in my view, often based on a suspected lack of statistical power. But, if you buy the concept of statistical power, then you are no longer in the realm of Fisherian evidence, but in the Neyman-Pearsonian realm of decision error control (discussed further ahead in this article). Statistical power can in principle be achieved to the point where one may “safely” accept the null (in the sense of acting as if it was true) over an alternative hypothesis specifying a very small effect, in the sense that the long-term risk of making the wrong decision when the effect is there is very small, say, 1%. However, that safety is granted through properties of the decision procedure, not properties of the data. Fisherian evidential $p$-values, irrespective of how large they are, do not constitute evidence for the null because their distribution does not change with increasing sample size when the null is true and the $p$-value itself is conditioned only on the truth of the null.

**DEFCON 3: $p$, the null, and the null only**

The fact that $p$-values are conditioned on the truth of the null and the null only introduces a second problem, if one is inclined to interpret $p$-values as evidence. The problem is that because evidence is a relative concept (Goodman & Royall, 1988; Royall, 1997) there is no way for $p$-values to constitute evidence for anything, because computing them does not involve any other hypothesis
than the null.

Suppose I randomly draw a card from a deck. It is the king of hearts. On the hypothesis that the deck was a regular deck the probability of the outcome is about .02. For most people, $p = .02$ signifies a quite improbable outcome. Do you consider the king of hearts evidence against the hypothesis that the deck is regular? Chances are that you do not, which shows that the occurrence of an improbable outcome given a hypothesis does not constitute evidence against the hypothesis. If there is no additional hypothesis on the table then there is no way for us to consider the draw of a king of hearts as evidence against a regular deck. A draw of any particular card is neither evidence for nor evidence against a regular deck in general. Only if an additional hypothesis is given can we consider the draw as evidence. For example, a draw of a king of hearts is evidence for a deck consisting only of kings of hearts relative to a regular deck (cf. Goodman & Royall, 1988), because the draw is much more likely under the trick deck hypothesis (but if your belief in that hypothesis was very weak before the draw, it may still be very weak after the draw, after being updated through Bayes theorem).

The important point here is that $p$-values correspond exactly to the deck scenario, because the calculation of $p$-values only and always involves one hypothesis, the null. The evidential interpretation of p-values is thus equally problematic as the evidential interpretation of the deck scenario. A quick retort to the classic alternative hypothesis as specified in the Neyman-Pearson framework does not per se solve this problem, because the $p$-value itself is always conditioned solely on the null. (As discussed further ahead, the Neyman-Pearson framework escapes the problem by not being about evidence, embedding the use of $p$ in a formal decision procedure.)

**DEFCON 2: $p$, the null, the null only, and probability, not strength**

The previous paragraph points toward a general third problem concerning the relation between $p$ and the concept of evidence. When we say that these data provide some degree of evidence for
hypothesis H1 (as compared to that of, say, hypothesis H2), we presumably take this to mean that the degree designates a certain level of strength. The evidence could have been weaker or it could have been stronger and what we have obtained falls somewhere on the scale. Using $p$ as an indicator of evidential strength confuses the relative strength of evidence (for one hypothesis compared to another) with the probability of obtaining that (or more extreme) evidence (given only one hypothesis) (Dienes, 2008; Royall, 1997). Identical $p$-values, if consistent as a measure of evidence, should always convey the same evidential strength, but they do not (Hubbard & Lindsay, 2008; Royall, 1997; Wagenmakers, 2007). As noted in the previous paragraph, the probability of obtaining evidence given a hypothesis will not translate into strength of evidence if an alternative probability is absent, and for $p$-values the alternative is always absent.

**DEFCON 1: $p$, the null, what did not happen, and the universal chart of intentions**

A fourth problem with $p$-values as a measure of evidence is that they depend on unobserved data. This problem is rarely noted in the psychological literature, but is well known in the statistical literature (Wagenmakers, 2007). Because the $p$-value is defined in terms of the integral of an extreme region in the sampling distribution where the hypothesis is fixed and the data varies, the unavoidable consequence is that $p$-values are defined in terms of data that were not observed. For example, in Figure 1, the $p$-value of .049 was obtained by looking, not only at the point where $z = 1.65$, but at the entire area below the curve beyond $z = 1.65$ as well. This region contains data that we could have observed, but did not observe, such as $z = 1.69, z = 2.2, z = 3.45$, and so on.

This property of $p$-values, that they go beyond the observed data, is virtually never justified nor discussed in introductory statistics and methodology books in psychology, but is simply treated as one of several steps in arriving at $p$-values (for a refreshing exception, see Dienes, 2008). But this property of $p$-values is devastating for interpreting $p$-values as measures of evidence, an
interpretation which is common in introductory books and among researchers, although often mixed up with a Neyman-Pearsonian stance toward decision error control (Gigerenzer, 1993; Hubbard & Armstrong, 2006). If $p$-values depend on unobserved data, then taking $p$-values as a measure of evidence implies that the evidence obtained from an experiment depends on things that did not happen (cf. Kruschke, 2009). This is, for many, a counter-intuitive and ultimately undesirable property for any measure of evidence, and a property that violates the likelihood principle (Birnbaum, 1962) which states that all inferential information contained in a sample is provided by the likelihood function which does not incorporate unobserved data.

It is not in any way obvious why things that could have happened, but did not happen, should influence the evidence indicated by a statistic. No one has put it better than Jeffreys (1998): “What the use of $P$ implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure.” (p. 385). The fact that $p$-values are defined in terms of unobserved data in a sampling distribution brings with it some peculiar consequences. Despite the objective aura surrounding $p$ in many introductory books and in many results sections, $p$ is ultimately subjective in a way that is completely unknowable (Dienes, 2008; Royall, 1997) as illustrated by the following example.

It was actually Joe the punk rock research assistant who helped me conduct the punk rock memory test that reportedly generated a $p$-value of .049. I designed the study but Joe handled all the experimentation and data analysis. I remember the morning that Joe entered my office to report the results. I asked him what the results were and he said “It looks real interesting, we have a $p$-value of .049 indicating an effect in the predicted direction. Why, no more beer for me. No sir, at least not before any show. Or, not before the important shows. We have to brush up our act as punk rockers and nail the lyrics. By the way, I looked at the data after 20 punk rockers you know. We had a $p$-value of .20 at that point. Good thing we ran 50 punk rockers.”. Somewhat perplexed I yell “What? You looked at the data after 20 punk rockers?! Houston, we have a problem. Our $p$-value may or
may not be .049 anymore.” Joe looks at me with suspicion: “Hey! I know how to calculate a $p$-value. I’m not Einstein but just because I’m a punk rocker it doesn’t mean I screw up every time I get the chance. Do it yourself if you don’t trust me.”

Joe the punk rock research assistant is sincere and good with numbers, but what he fails to take into account is the connection between the sampling distribution and hypothetical actions. We need to ask ourselves, what would have happened under alternative scenarios when Joe peeked at the data after 20 punk rockers? What if the results had been significant then? Would we have stopped the experiment and be done? What if he had told me that the results were $p = .20$ after 20 punk rockers? Would we have stopped or continued toward the planned 50 anyhow? What would Joe say to me and would he be able to convince me in any direction? Would I be able to convince him? Answering these questions matter a great deal, because the $p$-value is defined via the sampling distribution, which in turn is only defined given a particular sampling procedure. If we do not know how we repeatedly are to imagine sampling from a population, then we do not have a definite sampling distribution.

Note that Joe the punk rocker is only included in this story in order to make the point crystal clear, but the same problem arises with or without Joe peeking at the data. What if I conducted the entire experiment alone? What would I have done had the results not been significant after the planned 50 punk rockers? Would I have run additional punk rockers? Of course, I could say “no, I would have stopped”, but who really knows? The general problem of what anybody would have done had things been different is an ancient philosophical problem connected to the nature of free will, consciousness, agency, determinism, quantum wave functions, and what not, and it is not at all clear what the answer should be.

The requirement to answer the question of what one would have done had things turned out differently in order to even generate a $p$-value is no trivial requirement and it is ultimately counter-intuitive. Why should things that did not happen matter? The dependence of $p$-values on
unobserved data leads to a set of procedures that students and researchers often take for granted, such as stopping rules, corrections for multiple testing and post-hoc reasoning. While such procedures are well motivated within the realm of *p*-values it is a useful exercise to ask oneself whether such procedures should be part of a measure of evidence (cf. Cornfield, 1976). Do you think that what did not happen should influence the evidential value of your data? Do you think that whether you thought of an idea on Monday or Tuesday should influence the evidential value of your data? If the data remains the same, but the mental state of a researcher in a Swedish university changes slightly, does the evidence indicated by the data suddenly change? Ultimately, *p*-values are inherently subjective because they depend on your answering what you would have done under alternative scenarios. Not just what you think you would have done, but what you *actually would* have done (Royall, 1997).

Can we avoid the above difficulties by sticking to the Neyman-Pearson framework? The answer is: to some extent, but not completely, and maybe not in a way that will satisfy what most researchers want. The good news is that the Neyman-Pearson framework is much more coherent than the Fisherian. The bad news is that the former does not provide researchers with what they often want: a measure of evidence. In addition, some of the counter-intuitive consequences of *p*-values remain in the Neyman-Pearson framework.

**p and Decision Error Control**

Jerzy Neyman and Egon Pearson did not follow Ronald Fisher in considering *p*-values as measures of evidence against a hypothesis. Instead, the Neyman-Pearson framework is all about controlling decision error rates. Neyman and Pearson (1933) express the underlying motive succinctly: “We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis. But we may look at the purpose of tests from another view-point. Without hoping to know whether
each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.” (p. 290-291, italics added).

Briefly, in the Neyman-Pearson framework a single hypothesis (such as the null) is not enough, but an alternative hypothesis is required. Depending on the outcome of our particular test statistic we reject one of the two hypotheses. We know that, in doing so, we will be wrong to some extent in the long run. The probability of rejecting the null when it is true (Type I error) is \( \alpha \) and is fixed by specifying a critical region of the sampling distribution. If \( p < \alpha \) then the null is rejected. The probability of rejecting the alternative hypothesis when it is true (Type II error) is \( \beta \) and is a function of \( \alpha \), \( N \) (sample size), whether the test is one- or two-tailed, and effect size. Statistical power is defined as \( 1 - \beta \) and gives the probability of accepting the alternative if it is true. Note that each of the two hypotheses either is true or not. The probabilities apply to the long-term relative frequency of our decisions (reject or accept) given certain sampling assumptions, a specified test statistic, whether a given hypothesis is true or not, and specific decision rules. Also note that \( \alpha \), not \( p \), indicates Type I error rate. In the Neyman-Pearson framework \( \alpha \), the probability of Type I errors, is a fixed property of the test procedure and \( p \) is a random variable whose exact value is irrelevant. All we need to know about \( p \) is whether \( p < \alpha \).

The four problematic points concerning \( p \) used as a measure of evidence discussed previously were 1) \( p \) cannot indicate evidence for the null, 2) \( p \) does not respect the relativity of the concept of evidence, 3) \( p \) confuses strength of evidence with the probability of obtaining evidence, and 4) \( p \) depends on unobserved data (see Table 1). If the Fisherian interpretation cannot handle these problems, perhaps they Neyman-Pearson interpretation can?

The Neyman-Pearson framework directly dodges the first three problems because in that framework there is simply no use of \( p \) as evidence at all. What the Neyman-Pearson framework does is to specify a set of decisions and a set of decision rules. Depending on the outcome a
decision is taken and long-term error rates are under control. The Neyman-Pearson framework does not involve the objective of providing measures of evidence, so the first three problems listed above can hardly be considered problems for that framework, if the framework is judged by its own standards. Avoiding the problems while retaining $p$ may seem virtuous, but it requires detaching and trading in the concept of evidence from $p$. For instance, if $\alpha = .05$ then $p = .00001$ is not more evidence against the null than $p = .049$, or even $p = .99$, because neither of them indicate evidence. Informally and intuitively, one may still be stuck with the thought that “surely my data that gave $p = .00001$ indicates more evidence against the null than $p = .99$ for equal sample sizes?” Well, true, but it is not of much more help than simply looking at a graph and $p$ alone is not enough to settle the score. In the Neyman-Pearson framework $p$ does not indicate evidence, but one may be tempted by a compelling bending of the mind back into the Fisherian interpretation (it really is difficult to withstand those gravitational forces). In that case, however, one is back on square one, once again facing the problems outlined previously (e.g., that $p$ is independent from any alternative hypothesis). One could investigate how $p$ fares against more consistent measures of evidence and use $p$ on that basis. This has been done (e.g., Berger & Sellke, 1987; Wagenmakers, 2007) and the answer is that $p$ often overstates the evidence against the null. Quite regardless of being biased, if $p$ does not embody the philosophical and statistical principles of the concept of evidence, then why not stick to those measures that do embody consistent principles of evidence, such as likelihood ratios (Glover & Dixon, 2004), given that evidence is the wanted property? Neither Fisher $p$ usage nor Neyman-Pearson $p$ usage provides such a property in a sensible way. Within the Neyman-Pearson framework error rates may (ideally) be consistently controlled, which is a wonderful thing if that is the wanted property. It comes at the price of evidence and Neyman and Pearson (1933) were of course well aware of this. By design (or by the Neyman-Pearson lemma), the Neyman-Pearson procedure may very well indicate that a hypothesis should be rejected even though the evidence does not speak against it (Royall, 1997)
What about the fourth problem concerning the dependence of $p$ on unobserved data? This problem remains in the Neyman-Pearson framework, but the severity of the problem is a function of philosophical outlook. When a measure of evidence is on the line it is difficult to avoid the sting from the fact that what did not happen and unknowable intentions affects the evidence provided by a set of data. For many, the sting may very well appear equally daunting in the Neyman-Pearson context of pure action aimed to control decision error rate, but the problem is, in my opinion, less severe in the Neyman-Pearson context.

Someone might critically ask “but why should unobserved data affect the accept/reject verdict that my data affords?” It is somewhat misleading to suggest that your data “affords” a conclusion in the Neyman-Pearson context. The decision you make, the verdict given, the sentence assigned is a function of the region in which the data lands, but to say that your data “affords” a conclusion invites the interpretation that one rejects a hypothesis because the data speak against it. This is not the motive underlying the Neyman-Pearson approach. The objective is to as consistently as possible control decision error rates in the long run and, obviously, if we are to control this in the long run, we have to know what we are supposed to imagine that we do in the long run, which may involve a variety of alternative outcome scenarios. In this context, it is for example natural that two experiments differing only in their sampling plan and leading to the same data result in different $p$-values (or adjusted alpha levels). Unobserved data have a natural role to play here, but the most crucial insight that sometimes is forgotten is that in the context of Neyman-Pearson and decision error control, evidence is not the central part of the inferential process.

The general problem of not being able to ultimately know what one would have done had things turned out differently to some extent remains in the Neyman-Pearson approach. But, this is precisely why the Neyman-Pearson approach is surrounded by strict decision and sampling rules that cannot be tampered with. Fisher objected against such rules, which he found inappropriate for science and considered on a par with Russian collectively organized 5-year plans (Fisher, 1955).
What Fisher did not appreciate is exactly what makes the Neyman-Pearson approach as consistent as possible (although one can of course question whether the approach itself is what scientists want or need), because strict rules that can in principle be automated constrains potential effects of subjective researchers’ inferential and interpretational whims. For instance, close to all psychologists have probably encountered the term “marginally significant”, denoting a $p$-value between .5 and .10. Often such results are taken to mean “not quite there, but almost, and that counts for something and grants some support against the null”. This leans toward the (problematic) Fisherian interpretation, but often this is mixed up with the Neyman-Pearson approach, because decision error control is also often assumed to be on the table. If the latter approach is taken, then there is no such thing as marginally significant in the traditional sense.\textsuperscript{2} In the end, though, even the Neyman-Pearson approach is subjective in the sense that it inherently incorporates statistics that depend on what did not happen.

A classic recommendation in the methodological and statistical literature within psychology is that researchers consider replication, effect size and confidence intervals, rather than simply rely on $p$ (Cohen, 1994). This is in the general spirit of recommending that one take a variety of indices and strategies into account when evaluating research. However, it is important to be clear about the benefits and limits of what each piece of information can be expected to provide. For current purposes, it is of main interest to evaluate how these complementary strategies relate to the concept of evidence, a concept for which $p$ is unable to serve coherently.

\textit{Replication, effect size, and confidence intervals}

Replication is an ambiguous term, but even when specified further (e.g., an effect size $> 0$ in the predicted direction) it is clear that replication does not easily translate into any obvious evidence metric. Successful replications just result in more data and the question becomes “how do we evaluate data in terms of evidence?” The concept of replication has little to offer in this regard,
except from the elusively obvious notion that “the more replicable the better”, but better for what? The fact that a phenomenon is incredibly replicable has, per se, no direct impact on evidence. To rely on replication is simply to, explicitly or implicitly, be faced with the question of how to evaluate the data that comes out of those replications.

Effect size is also insufficient as a index of evidence. In many circumstances effect size will be related to the concept of evidence, but ultimately it depends on the questions asked. For example, suppose hypothesis 1 (H1) predicts a really small difference between an experimental and a control group, H2 predicts a medium difference, and H3 predicts a huge difference. The results show a really small difference and the data would seem to provide more evidence for H1 over H3 than for H1 over H2, because the results diverge more from H3 than from H2. This is not captured by the effect size, which stays the same regardless of the hypotheses considered. Effect size is just a descriptive tool and a property of the data, independent of the hypotheses considered.

Turning to confidence intervals, an x % confidence interval means that when using a specified sampling procedure the true value of an estimated parameter will fall within the boundaries of the resulting intervals x % of the time. It does not mean that the true value will fall within this particular one interval x % of the time and it does not mean that one can be x % certain that the true value is within any particular interval. As can be seen from the definition of confidence intervals, they are intimately related to the logic of the Neyman-Pearson framework. Within that framework they provide an index of precision and sensitivity in terms of sampling and long-term objective probability. Translating that into an index of evidence for or against a particular hypothesis is neither obvious nor feasible. Confidence intervals simply do not speak directly to the question of what these and only these data imply for any hypothesis, because confidence intervals depend on considering a whole class of (imagined or real) replications.

Replication, effect size, and confidence intervals can surely constitute useful tools in the course of scientific investigation. The claim here is simply that they do not constitute straightforward
candidates to take on the role of evidence indices, except in the general and informal sense in which, for example, a successful replication may be said to constitute additional evidence for a hypothesis. However, in such statements, the evaluation of evidence is presupposed and implicitly buried within the statement itself. A replication is only evidence for x if the original data were evidence for x, and the question is “in what sense were they evidence?”.

The efforts of trying to squeeze the concept of evidence into the Neyman-Pearson framework are both futile and surprising, because 1) the framework was not built to handle it and 2) there are alternative frameworks that are able to handle the concept of evidence in a straightforward and consistent way and that respect our basic intuitions concerning the concept of evidence. The following section discusses the most direct way of representing the strength of statistical evidence, namely the likelihood ratio.

**The Likelihood Ratio**

Likelihood was an important concept for both Fisher and Neyman-Pearson, but for neither of their legacies did likelihood assume the leading role designed to directly embody the concept of evidence. For Fisher, mainly the \( p \)-value emerged as a measure of evidence, while Neyman-Pearson avoided the entire issue of evidence.

A \( p \)-value denotes the probability of these or more extreme unobserved data under a fixed hypothesis. In contrast, the likelihood of a hypothesis (for continuous data) denotes the probability density (multiplied by any arbitrary constant) of exactly these observed data under the hypothesis. A likelihood function is obtained by calculating the likelihood over a range of specific hypotheses, where the different hypotheses specify different values of a parameter. The *likelihood ratio* requires two specific hypotheses (H1 and H2) and is given by the likelihood of H1 divided by the likelihood of H2 (Dienes, 2008; Edwards, 1972; Glover & Dixon, 2004; Royall, 1997). The likelihood ratio forms a relative measure of statistical evidence: the more likely the data are under H1 compared to
H2, the more the data are evidence for H1 over H2. The data is evidence for one hypothesis over another if the former more strongly predicts the data relative to the latter.

The likelihood ratio does not in itself say what one should do or how one should act in light of the data. Such considerations depend on factors outside the data space (e.g., cost, risk aversion, prior commitment). Instead, it forms a measure of statistical evidence, with respect to the observed data and two hypotheses, that is free from the problems with $p$-values as measures of evidence. The likelihood ratio can indicate support for the null hypothesis, it is conditioned on two hypotheses and forms a relative concept, it does not confuse the strength of evidence with the probability of obtaining evidence, and it does not depend on unobserved data. In sum, it avoids the problems with evidential $p$-values and provides researchers with a direct and intuitive concept of statistical evidence.

Royall (1997) gives a clear general overview of likelihood ratios and on how to eliminate nuisance parameters (parameters not of immediate interest, such as the variance of a distribution when the mean is of interest). Dienes (2008) provides an introduction to various schools of statistical inference, the likelihood being one of them, along with examples and computational tools. Glover and Dixon (2004) provide an approach to calculate likelihood ratios from SPSS output along with possible ways of correcting for the complexity of hypotheses. Blume (2002) provides a likelihood tutorial with, among other things, examples and discussion of nuisance parameters. Royall (2000) discusses the probability of observing misleading evidence when using likelihood ratios.

Likelihood ratios have not yet received the detailed computational and implementational treatment in psychology that they deserve. Every textbook of statistical inference aimed at psychologists dives into $p$-values and (often incorrectly) explains their basis, but very few books include likelihood ratios. There is a cure for the commandment “Hail the impossible” and that cure is simply to find alternative candidates to embody the concept of evidence. The likelihood ratio is
one such candidate. Given that one is interested in the evidence for a hypothesis over another in light of the data, \( p \)-values are much less than ideal.

*Example: Punk Rock Revisited*

What is the evidence conveyed by the data with respect to the null hypothesis \( H_0 (\mu = 0.00) \) in the punk rock memory study? This question is generically meaningless unless we specify an alternative hypothesis \( H_1 \). Suppose that we expected a mean score of 3.00 among the beer-deprived punk rockers, corresponding to a medium effect size \( (d = (3.00-0.00)/6.00 = 0.50) \). If we treat this expectation as a relevant comparison against the null, what does the data say about the null \( (\mu = 0.00) \) and the alternative \( (\mu = 3.00) \)? Recall that according to the significance testing procedure we have evidence against the null \( (Fisher, p = .049) \) or should reject the null \( (Neyman-Pearson, p < .05) \).

Assuming a normal population and a known population standard deviation of 6.00, the likelihood ratio comparing \( H_0 \) and \( H_1 \) is given by the probability of the observed data \( (M = 1.40 \) of our 50 scores) under \( H_0 \) divided by the probability of the observed data under \( H_1 \). Consulting the relevant normal sampling distributions under \( H_0 \) and \( H_1 \) respectively, we find that the probability (density) of the observed data is about 0.12 under \( H_0 \) and about 0.08 under \( H_1 \). The likelihood ratio given by \( p(\text{observed data}|H_0) / p(\text{observed data}|H_1) \) indicates support for \( H_0 \) over \( H_1 \) if larger than 1, and evidence for \( H_1 \) over \( H_0 \) if smaller than 1. Ratios larger than 8 (or smaller than 1/8) indicate “pretty strong evidence” (Royall, 1997). Our likelihood ratio is 0.12/0.08 = 1.50, indicating very weak evidence for the null over the alternative hypothesis.

Figure 2 shows the entire likelihood function for the punk rock memory data. The likelihood ratio for any two hypotheses (population means) is obtained by the ratio of the curve heights of the population means being compared. The observed mean is the maximum likelihood estimate of the population mean, that is, the most likely estimate of the population mean given the data. The
horizontal line in Figure 2 shows the 1/8 likelihood interval: the range of alternative population means which are at most 8 times less likely than the maximum likelihood estimate. Population means outside the 1/8 likelihood interval are at least 8 times less likely than the maximum likelihood estimate.

The strength of evidence against or for the null depends on the alternative hypothesis, just as it should be. Evidence is relative and it may or may not be informative to know that the punk rock data is (weak) evidence for the null over a specific alternative. Figure 3 gives a more complete picture by depicting the likelihood ratio of the null against various alternatives. Here, the likelihood ratios are transformed by the natural logarithm, resulting in an additive scale. Negative ratios indicate evidence for the alternative over the null, positive ratios indicate evidence for the null over the alternative, and 0 indicate no evidence either way (Goodman & Royall, 1988, provide interpretational guidelines).

The Fisherian interpretation would have us believe that we have obtained evidence against the null generically (not relative evidence, but evidence in general), because $p = .049$. The likelihood ratio says otherwise. For example, for an alternative hypothesis specifying a medium effect size, we have very weak evidence for the null. The Neyman-Pearson interpretation implies that we should reject the null, if we had set $\alpha = .05$ (or any value $> p$, before the data comes in). For a medium effect size ($d = 0.50$) statistical power is $1-\beta = .97$. Thus, the error rates for decisions with respect to the two hypotheses ($\mu = 0.00$ and $\mu = 3.00$) are small. Still, the likelihood ratio indicates very weak evidence for the null compared to an alternative hypothesis specifying a medium effect size. There is no contradiction here, because the likelihood ratio does not say that we should not reject the null. It does not tell us what to do at all. It simply tells us what the evidence, as conveyed by our data, really is, something that different usages of $p$-values (Fisherian or Neyman-Pearsonian) cannot properly do. Judging from research articles and the way researchers talk and write, it seems that what researchers really want actually is a measure of statistical evidence.
It is ultimately unclear what it means to reject a hypothesis on the basis of significance testing anyway. If a hypothesis is rejected in a research paper, does this mean that we from now on act as if the hypothesis is not true. What if a paper contains two experiments, one that rejects the hypothesis and one that does not. How should we act? Does it depend on the order of the experiments? What are the consequences for science, not for the industrial manufacturing of ball bearings, but for science and the growth and communication of knowledge? What do all these artificial rejections really amount to?

The example shows that the notion of evidential $p$-values and the notion of likelihood ratios can convey different strength of evidence (cf. Dixon, 2003). The likelihood ratio rests on coherent and intuitive principles as measures of evidence (Royall, 1997) while $p$-values do not.

**Nuisance parameters**

In the punk rock memory example, the variance of the population was treated as known and we were interested in the mean. Typically, in real cases with similar data, the variance is not known, yet we are still interested primarily in the mean. The variance is then referred to as a “nuisance” parameter. Given a normal population, the likelihood depends on both the mean and the variance, so how can we get the likelihood if the variance is unknown? More generally, how can we obtain the likelihood in the presence of nuisance parameters?

The problem of nuisance parameters does not have a general single-strike solution, but there is a variety of different approaches. The Bayesian solution is to specify a prior probability distribution for the nuisance parameters and integrate them out, but this introduces subjectivity in specifying a prior probability distribution and goes beyond summarizing statistical evidence per se as indicated by observed data. For a variety of alternative ways of eliminating nuisance parameters, along with various pros and cons of the alternatives, see Royall (1997, 2000) and Blume (2002).

Suppose the population standard deviation for the punk rock memory data ($\sigma = 6.00$ in the
example above) is not known, but just an estimated population standard deviation, \( \sigma_{\text{estimate}} = \sqrt{(SS/(n-1))} = 6.00 \). For the punk rock memory data, one solution if the variance is unknown is a so-called profile likelihood. This is not a real likelihood, but a likelihood obtained in our case by maximizing the joint likelihood of the parameters for each value of the parameters of interest. The resulting likelihood for the mean with unknown variance is then proportional to a \( t \)-distribution (Blume, 2002, Appendix B).

Figure 4 shows the profile likelihood for the punk rock memory data when the variance is unknown (using the \( n-1 \) modification suggested by Royall, 1997, pp. 133-134). Figure 5 shows the log likelihood ratio of the null against various alternative hypotheses (population means) when the variance is unknown. Figure 4 and 5 (variance unknown) convey essentially the same information as Figure 2 and 3, except that the former are flatter than the latter, reflecting the fact that the variance is unknown. Profile likelihoods are often useful in many of the situations common to psychologists, although they do have limitations, such as when the number of parameters is large relative to sample size (Blume, 2002).

A one-sample \( t \)-test for the punk rock memory data against a population mean of 0 gives \( t(49) = 1.65, p = .053 \). The situation is of course pretty much the same as for the \( z \)-test, except that we now cross the boundary of .05. For the Fisherian interpretation, nothing has changed really. Supposedly, we are licensed to think that we have obtained some degree of evidence against the null, and if the evidence was of some degree when \( p = .049 \) then it is reasonable to suppose that \( p = .053 \) is also evidence against the null. Yet, the log likelihood ratio in Figure 5 says otherwise, whether we have evidence against the null or not depends on the alternative hypothesis entering the comparison. For the Neyman-Pearson interpretation, the situation is different. If \( \alpha = .05 \) we should not reject the null. Yet, the evidence is practically the same as before. The Neyman-Pearson framework controls long-term decision errors and likelihood ratios quantify statistical evidence. The question is: what do researchers really want? Let us put the question differently: How often do researchers really adhere
to the Neyman-Pearson methodology and act as if a hypothesis is either true or false based on properly controlled probabilistic error rates? What does that even mean, to act as if a hypothesis is either true or false? Psychological researchers are generally not inclined to follow Neyman-Pearson. One could say that researchers calculate $p$-values in terms of Neyman-Pearson, but interpret them in terms of Fisher. For example, take a random research article in psychology and you are quite likely to find that researchers say things like “the results constitute evidence that…” or “we found evidence for an effect of:…”. Yet, what is often missing is a proper measure of the strength of that evidence in statistical terms (although the $p$-value is often given that role). Likelihood ratios provide such a measure. Before concluding this article, I consider some common objections to likelihood ratios and argue that the objections are all misleading.

**Objections to Likelihood**

1) “With likelihood ratios one will almost always find evidence against the null.”

The reasoning behind this objection is that in most cases the likelihood of some specific hypothesis other than the null will be higher than the likelihood of the null. For example, for our punk rock study, an observed mean $M = .00$ is the only observed mean that will generate a higher likelihood for the null than for any other specific hypothesis. All other observed means will enable at least some likelihood ratio that constitutes evidence (regardless of how weak) for at least some alternative hypothesis over the null.

Objection 1 above is a misleading objection. The objective of likelihood ratios is to quantify statistical evidence, so that given a probability model and observed data they quantify the relative likelihood of one hypothesis in relation to another. This is a measure of the strength of evidence. This question should not be confused with the probability of observing misleading evidence (Blume, 2002; Royall, 2000). The probability of observing misleading evidence is relevant before conducting an experiment. Once the experiment is conducted that probability becomes immaterial.
and has nothing to do with the strength of observed evidence. Once the data are in the evidence is either misleading or not, and there is no way to know which is the case. The probability of observing misleading evidence can be calculated and should be calculated before an experiment is conducted, in order to minimize the risk of observing it.

Even though it is often possible to find evidence against the null, it is certainly not the case that one is likely to find strong misleading evidence against the null. The universal bound \(1/k\) provides a maximum boundary for the probability of observing misleading evidence, where \(k\) is the likelihood ratio for any two specific hypotheses (Royall, 2000). In practice, the probability is often considerably smaller. Misleading evidence for \(H_1\) in this context means observing a likelihood ratio favouring \(H_1\) when \(H_2\) is true. Furthermore, the more data you collect (regardless of your sampling plan), the less likely you are to observe misleading evidence. Contrast this with \(p\)-values that are guaranteed to be smaller than .05 if you just keep collecting data and stop when it suits you. With likelihood ratios, you are generally unlikely to observe strong misleading evidence for a hypothesis over the null even if you set out to selectively find it (Royall, 2000). Likelihood ratios encourage researchers to collect more data, while \(p\)-values punish researchers for collecting data (unless it was part of the subjective sampling plan). Using likelihood, the more data you collect the more likely it is that the true hypothesis is the one receiving the greatest likelihood. What more could you ask for?

2) “Likelihood ratios may encourage researchers to accept or believe strongly in a hypothesis that was implausible to begin with.”

This objection is also misleading. The reasoning behind the objection is that an experiment may generate evidence for an implausible alternative hypothesis (e.g., a particular parameter value indicating strong extrasensory perception) relative to a null (e.g., no extrasensory perception). Because likelihood ratios do not require specifying the implausible alternative hypothesis (or any hypothesis) beforehand and because they do not take prior beliefs into account, one may end up rejecting the null or believing strongly in the alternative on the basis of a likelihood ratio that
indicates evidence for the alternative relative to the null.

The objection is misleading because it presupposes that a measure of statistical evidence (the likelihood ratio) taken by itself should tell you what to do and what to believe. This is not the case. There may very well be factors (such as prior belief or cost) that result in rejecting a hypothesis that is supported by the data relative to another hypothesis or believing in a hypothesis that is less supported by the data relative to another hypothesis. In fact, the Neyman-Pearson lemma implies that the former situation may very well happen and Bayes theorem implies that the latter situation may very well happen. There is no contradiction here because what the data say in terms of evidence for or against an hypothesis relative to another is a different issue than decision making or specifying what to believe.

3) “Likelihood ratios require that researchers specify specific relevant hypotheses, such as $\mu = 1.45$, but researchers are often unable to specify such hypotheses. Significance testing and Bayes factors can deal with composite hypotheses, such as $\mu \neq 0$, while likelihood ratios cannot”.

As with the other objections, this one is misleading as well, or simply wrong. The interpretation of likelihood ratios is guided by the Law of Likelihood which says that the magnitude of the likelihood ratio represents the strength of evidence for any two specific hypotheses under a probability model and observed data (Hacking, 1965; Royall, 1997). Composite hypotheses, such as $\mu \neq 0$, do not have a likelihood so there is no way of directly representing relative evidence for or against a composite hypothesis (Royall, 1997). $p$-values do not represent evidence at all. Bayes factors constitute the Bayesian version of likelihood ratios and can deal with composite hypotheses (see Dienes, 2008, for an introduction), but they do so by 1) incorporating prior probabilities for a hypothesis (and thus do not depend solely on the data) and 2) by changing the problem from evidence for or against a composite hypothesis to average weighted evidence for or against a range of specific hypotheses. As emphasized by Royall (2000) it is simply not the case that significance testing or Bayes factors achieve representing relative evidence for or against a composite hypothesis.
where likelihood ratios cannot achieve this. None of the approaches achieve it, because it cannot be done, but significance testing and Bayes factors may be said to represent different ways to deal with situations that involve composite hypotheses.

Researchers are not required to specify specific hypotheses beforehand in order to use likelihood ratios. The likelihood function depicts everything you need to know about the data with regard to their status in terms of relative evidence. However, in order to calculate the probability of misleading (or of weak) evidence, you have to consider at least two specific hypotheses (just as you have to consider two specific hypotheses when calculating the frequentist corresponding Neyman-Pearson notion of power).

Psychological theories are often not constrained to the extent that specific hypotheses can be generated from a theory. In effect, one may get the impression that evidence cannot be quantified in terms of likelihood ratios in psychology. However, the data will speak for or against a specific hypothesis relative another one in terms of statistical evidence, regardless of whether the researcher is able to map those hypotheses onto a theory. The likelihood function depicts all of the information in a single blow and will provide the relevant statistical evidence of data regardless of whether the researcher is interested in it or not. This is useful for science because it represents the entire statistical evidence, given only a probability model and data, in a way that is consistent with the law of likelihood.

Conclusions

Many of the traditional objections against significance testing and p-values (e.g. .05 is an arbitrary convention) miss the target, because they do not go to the core of the problem. The core of the problem is this: p-values are unsuited as measures of statistical evidence, as illustrated by the problems discussed in this paper (Table 1) and using p-values as a tool for long-term decision error control, while feasible in theory, may not be what psychological researchers typically want.
What use is there for $p$-values for any given researcher who wants to communicate the evidence of observed data for or against a hypothesis relative to another hypothesis? The proper answer might be: *none*, because there are better alternatives. Incorporating effect size, confidence intervals, and a million replications does little to offer, not only a useful, but a precise and coherent measure of statistical evidence. In contrast, likelihood ratios can serve the purpose of quantifying statistical evidence. If the question of assigning belief to a hypothesis is on the table (and not just the evidence given by data), the entire Bayesian route is available (Kruschke, 2009; Rouder et al., 2009; Wagenmakers, 2007).
References


Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*, 289-337.


Table 1. *Four fundamental problems with p-values as a measure of evidence and the resulting consequences.*

<table>
<thead>
<tr>
<th>Problem</th>
<th>Consequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$ is uniformly distributed under the null</td>
<td>$p$ cannot indicate evidence for the null</td>
</tr>
<tr>
<td>$p$ is conditioned solely on the null</td>
<td>$p$ does not quantify evidence, which is relative</td>
</tr>
<tr>
<td>$p$ is about probability of obtaining evidence</td>
<td>$p$ does not quantify strength of evidence</td>
</tr>
<tr>
<td>$p$ depends on unobserved data and intentions</td>
<td>$p$ does not quantify evidence from observed data</td>
</tr>
</tbody>
</table>
Figure 1. Sampling distribution of the mean based on 50 scores for each sample mean from a normal population with a mean of 0 and a standard deviation of 6. The grey area extending beyond the obtained sample mean ($M = 1.4$) constitutes about 4.9% of the total area under the curve, $p = .049$ (one-tailed).
Figure 2. Likelihood function for the punk rock memory data. The parameter of interest is the population mean (normal population, variance known). The horizontal line shows the 1/8 likelihood interval (see text).
Figure 3. Log likelihood ratio of the null ($\mu = 0$) against various alternatives for the punk rock memory data. The points below the horizontal line indicate evidence for the alternative over the null, while points above the horizontal line indicate evidence for the null over the alternative.
Figure 4. Profile likelihood function for the punk rock memory data. The parameter of interest is the population mean (normal population, variance unknown). The horizontal line shows the 1/8 likelihood interval (see text).
Figure 5. Log (profile) likelihood ratio of the null ($\mu = 0$) against various alternatives for the punk rock memory data. The points below the horizontal line indicate evidence for the alternative over the null, while points above the horizontal line indicate evidence for the null over the alternative.
Footnotes

1 For an alternative, arguably more intuitive, but not unproblematic, interpretation of significance testing in terms of generating samples from a process with certain propensities rather than sampling from and generalizing to a population of individuals, see Frick (1998).

2 That these things really do confuse psychologists is testified by an email listserv discussion for social and personality psychologists in 2004, where responses in the hundreds to the question of whether there marginal significance exists sparked a controversy that, in the light of most responses, did little to illuminate the underlying issues. The question that sparked the controversy and some of the responses can be found at http://www.mail-archive.com/tips@acsun.frostburg.edu/msg11881.html

3 The probability density (and likelihood) here is
\[
\frac{1}{\sigma_M \sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma_M^2)},
\]
where the sample mean is \(x = 1.4\), the standard error of the mean is \(\sigma_M = 0.85\), and the population mean \(\mu\) varies as given by the hypothesis.

4 The profile likelihood here is given by
\[
\left[ 1 + \frac{t^2}{n-1} \right]^{(n-1)/2},
\]
where \(t = (x-\mu)/SE\), the sample mean is \(x = 1.4\), the estimated standard error of the mean is \(SE = 0.85\), and the population mean \(\mu\) varies as given by the hypothesis.