A fundamental problem with p-values and how it can be avoided using a Bayesian approach.

(Chapter 5, Example 6 of "Risk Assessment and Decision Analysis with Bayesian Networks", by Norman Fenton and Martin Neil, CRC Press, 2012)
Note: The print version of the book actually has some errors in this example that were corrected in the errata page: http://bayesianrisk.com/Errata.pdf

Suppose that we have 999 fair coins and one coin that is known to be biased toward 'heads'. We assume that the probability of tossing a 'head' is 0.5 for a fair coin and 0.9 for a biased coin. We select a coin randomly. We now wish to test the null hypothesis that the coin is not biased. To do this we toss the coin 100 times and record the number of heads, *X*.

As good experimenters we set a significance level in advance at a value of 0.01. What this means is that if we observe a *p*-value less than or equal to 0.01 (see sidebar) then we will reject the null hypothesis if we see at least 63 heads (in fact 62 would be very close to rejection also).

So, assuming that *H* is the null hypothesis (fair coin) and *E* is the observed event "at least 63 heads in 100 throws" then since $P(E \mid H) < 0.01$ (in fact $P(E \mid H)=0.006$) the null hypothesis is rejected and statisticians conclude that the coin is biased. It is typical to claim in such situations something like 'there is only a 1% chance that the coin is fair, so there is a 99% chance that the coin is biased.' But such a conclusion wrongly assumes that $P(E \mid H) = P(H \mid E)$.

The coin is, in fact, still very likely to be a fair coin despite the evidence. The key point is that the correct prior probability for the null hypothesis, *H*, is 0.999, because only one of the 1000 coins is biased. Also, $P(E \mid \text{not } H)$ is 0.999999999999899. Hence we have:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E|H) \times P(H) + P(E|not\ H) \times P(not\ H)}$$

$$= \frac{0.006 \times 0.999}{0.006 \times 0.999 + 0.999999999999899 \times 0.001} = 0.86$$

So, given the evidence *E,* the probability that the coin is fair has come down from 0.999 to 0.86. In other words, there is still an 86% chance the coin is fair.

The conclusions drawn by people using the *p*-value are actually even more misleading than what we have already revealed. That is because the event they actually observe is not "at least 63 heads" but a *specific* number of heads. So suppose they actually observe the number 63. In this case (using the Binomial distribution) it turns out that $P(E \mid H)=0.0027$ and $P(E \mid \text{not } H)$ is 0.0000000000000448.

This means that, from Bayes:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E|H) \times P(H) + P(E|not\ H) \times P(not\ H)}$$

$$= \frac{0.0027 \times 0.999}{0.0027 \times 0.999 + 0.0000000000000448 \times 0.001} \approx 1$$

In other words, far from the evidence 'proving' that the coin is biased, the evidence actually proves the exact opposite: the coin is almost certainly unbiased. ***The statistical hypothesis test gets it completely wrong.***

We would need to see a *much* higher number of heads in 100 tosses (than suggested by the pre-defined *p*-value) to in any way justify rejecting the null hypothesis.

<sidebar>
The Binomial Theorem (explained in Chapter 4) with n = 100 and p = 0.5 provides a formula for the probability of tossing X number of heads from 100 tosses of a fair coin. But what we want to know is the 99th percentile of the distribution, i.e. the number X for which 99% of the probability distribution lies to the left of X. Or equivalently 1% of the distribution lies to the right of X. It turns out that the 99th percentile in this case is 62. This means the probability of tossing at least 62 heads in 100 tosses of a fair coin is 0.01.
</sidebar>

The key in this example is that we started with a very strong prior belief that the null hypothesis is true. The evidence alone is nowhere near sufficient to 'shift' the prior enough for us to believe it is false. The Bayesian approach to hypothesis testing is covered in Chapter 10.