3 The Tyranny of Fisher

Suppose you have a number of dresses in your closet. You want to know which one fits you best. There are two basic strategies. You could try each dress on, one at a time, until you find one that seems to fit good enough for you. That is, you try to evaluate each dress relative to some standard of fit. Perhaps you want the hem line to reach no lower than 1 inch below the knee. The other approach would be to try each dress or suit on and select the one that fits best. This is a relative standard, finding the best dress by comparing them.

In statistical inference, these two approaches also exist, but the dresses are models and the fit is predictive accuracy, or some other criterion. Sometimes we compare models to an independent standard. Call this the *absolute evaluation*. Is the model good enough? Can it predict a crucial observation? Other times we compare models to one another. Call this the *relative evaluation*. Which model predicts best? Each approach has its strengths and weaknesses. The relative evaluation can more reliably find the best model in your closet, but even that model might fit badly. The absolute evaluation is good for telling us if we need to go buy a new dress or come up with another model. But it will be less good at finding the best fitting model, when several of them might be good enough. In that case, the order you evaluate the models in will determine which you choose. It might also be hard to determine the absolute standard to judge a model by. When is a model good enough? In contrast, it's a simpler matter to judge fits relative to one another, as you saw in the previous chapter, even though such comparisons tell us little about the absolute evaluations of models.

There is nothing stopping us from using both evaluation methods, of course. We could use relative evaluation to find the best dress in the closet, and then compare that dress to some absolute standard of fit. Indeed, this is the course that many statisticians practice and recommend, substituting dresses for models, of course. And often we don't want to choose any single model, but rather just summarize what the evidence

says about all of them. But much of applied statistics does little relative evaluation. Instead, many natural and social scientists have been trained to decide whether or not to accept or reject a statistical model based only upon a specific absolute evaluation, the *P*-value, of a single model that usually does not even correspond to a research hypothesis, but rather a *null hypothesis*.

A *P*-value is a kind of predictive check of a null model, a preferred dress from the closet. The standard of prediction for *P*-values is almost always that the probability of the observations or more extreme observations be greater than 0.05. If so, then it fits well enough, and usually no other model is checked. Indeed, often no other model is ever fit to the data. There may be models lurking in the closet that make much more accurate predictions than the null model does, but if one uses only this sort of absolute evaluation, one will never know it.

Moreover, as I argued in the first chapter, we should carefully distinguish *inference* and *decision*. The posterior density provides inferential power that can be used to make many different cost-benefit decisions. Much of the time in science, we are not trying to decide in the moment which model is correct. Rather we wish to infer what the evidence says about all of the models. The decision about which model to adopt is a community project. The *P*-value approach attempts to leap right past the inferential step to the decision step. Of course it must quantify how evidence reflects on a model, but because it uses only one model, it is prone to some odd behavior. And because it forces a decision, it encourages us to believe that we must make a ruling right now, instead of honestly reporting the posterior density to the community of scientists.

In this chapter, we consider in detail the Tyranny of Fisher, the dogmatic reliance upon null hypothesis significance testing for making decisions about models. By *null hypothesis significance testing*, I mean the procedure of computing a *P*-value for a point hypothesis of no effect or difference and then rejecting this null hypothesis only when $P < 0.05$. By *dogmatic reliance*, I mean the social context in which studies cannot get published and dissertations cannot get signed unless it is shown that $P < 0.05$.

In this chapter, I argue and show the reader how to prove that *P*-values are neither what most scientists believe them to be nor do they perform well at the jobs put to them. Absolute evaluations have a role to play, but not in this form and never alone. If you are like most scientists, the actual properties of a *P*-value will surprise you. Odds are that your elementary education in statistics actively misinformed you about them. Exploring and deconstructing the Tyranny of Fisher is partly a

problem in the sociology of science. Prominent statisticians have been trying for decades to correct misunderstandings and misuse of *P*-values, without success. Meanwhile, statistics departments continue to teach a procedure that they largely do not respect but that science departments require their majors to know.

I'm not going to have much new to say about the sociology of the problem, however. Instead, I will direct our attention to practical difficulties with *P*-values. The goal is to demonstrate, and teach the reader the tools needed to prove for themselves, that using *P*-values and typical null hypotheses provide unreliable help in finding good models. None of these arguments are new. But in my experience students and colleagues need to be confronted with them in an uncompromising way, before they will engage with more powerful approaches. At the end of the chapter, I recommend further reading to explore the history of these ideas.

The perspective I present here does reference Bayesian concepts, when they provide clarity. But it's important to realize that one does not have to accept Bayesian norms to accept most of these criticisms of *P*-values. Non-Bayesian statisticians have been just as fiercely critical of null hypothesis significance testing.

3.1. Defining the *P*-value

Before getting into misunderstandings and drawbacks of *P*-values, it is necessary to precisely define them and compare them to the measures in the previous chapter. Let's continue with the proportion of water on the globe example, just for sake of continuity and comparability.

In order to compute a *P*-value, we have to define a *null hypothesis*, which we have not done so far. In principle, the null hypothesis can be anything. But the reader is probably accustomed to the null hypothesis being equivalent to a model of no effect or no difference between groups in the data. The customary role of such a null model is to capture a hypothesis in which some casual effect is absent. So there is a natural attraction to such hypotheses, for many. But the logic of null hypothesis testing survives, no matter what value of a parameter we choose to be the null.

Which null model should we imagine for the proportion of water example? We could focus on the hypothesis that $p_W = 0.5$, but that seems bizarre. Why is half water and half land a hypothesis of no effect or no difference? So to make the example a little more interesting, suppose instead of sampling from our globe now, that we have launched a probe to an exoplanet with unknown proportion of surface water. When the

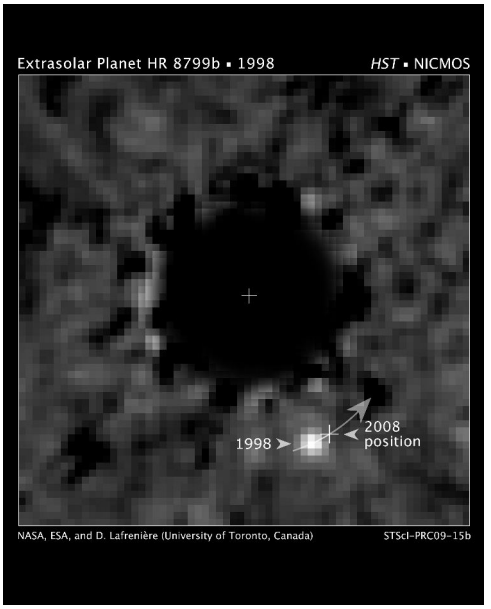

FIGURE 3.1. Our mysterious exoplanet. Image source: http://hubblesite.org/newscenter/archive/releases/2009/15/

probe arrives, it places 10 landing probes into essentially random decaying orbits. When each probe lands, it records only whether or not it landed in water and sends that message back to Earth. Since we can't see the exoplanet clearly at this distance, we will assume that the count n_W of the number of probes that said "water" is all the information we have to estimate the proportion of the exoplanet covered in water, p_W.

I don't happen to think that a null hypothesis makes much sense here, but that's because I don't think null hypotheses make much sense in hardly any circumstance. Many students I've taught find the proportion of water covering the Earth as a natural null, however, so let's adopt $p_W = 0.7$ as the null. Now that we have a null hypothesis, we define the P-value as:

> ***P*-value:** The likelihood of the observed data or any unobserved more extreme data, assuming that the null hypothesis is true.

How do we define "more extreme?" It is defined relative to the null hypothesis, H_0, and the scale of measurement we adopt. So if what we observe, D, is smaller than what is expected under the null hypothesis, then "more extreme" means all those likelihoods for values of D even smaller. This is a little confusing, so there will be plotted examples shortly.

In the case of our exoplanet problem, suppose half of the probes reported "water" and half reported "land." This means we observe $n_W = 5$, and the P-value in this case becomes:

$$P \equiv \Pr(n_W \leq 5 | p_W = 0.7).$$

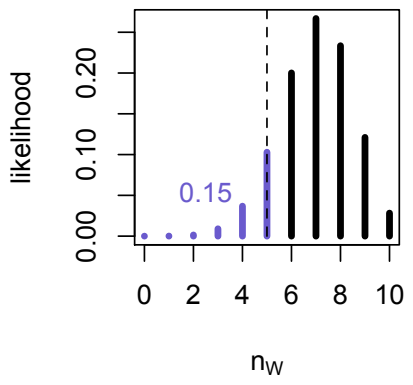

FIGURE 3.2. Computing the *P*-value when 5 of 10 probes report "water." Each bar is the likelihood of a possible observed count of "water." The vertical dashed line shows the location of the actual observation, $n_W = 5$. The blue bars sum to the *P*-value, $P = 0.15$.

This expression is actually a sum of likelihoods—probabilities of data given a model—and so expands to:

$$P = \sum_{n_W=0}^{5} \Pr(n_W|p_W = 0.7).$$

You can easily calculate this in R with code like:

```
p <- sum( dbinom( 0:5 , size=10 , prob=0.7 ) )
```

R code 3.1

This line of code just adds up all the likelihoods for those observations less-than-or-equal-to 5, holding the probability constant at 0.7. If you are not familiar with syntax like `0:5` for making lists, just enter that bit of code alone on the command prompt in R to see that it is just a shortcut for `seq(from=0,to=5,by=1)`.

Figure 3.2 illustrates this sum. On the horizontal axis are the possible observations, from zero reports of "water" to ten reports of "water." The vertical axis is likelihood, and the height of each bar in the figure is the likelihood of each possible observation, under the model that $p_W = 0.7$. So unlike all of the figures you saw in Chapter 2, here the model is constant and the data are what we are varying. The actual observation is shown by the location of the vertical dashed line at $n_W = 5$. The blue bars are those values of n_W equal to or more extreme than the actual data we observed. When we sum up the heights of these blue bars, we get the value shown in blue in the figure, 0.15. This is the *P*-value in this case, $P = 0.15$.

What do we do with this number? Following the widespread ritual of null hypothesis significance testing (NHST), we adopt $P \leq 0.05$ as the criterion for rejecting the null hypothesis. Since $P = 0.15 > 0.05$, we conclude that there is not sufficient evidence to reject the null hypothesis and probably even claim that there is no statistical evidence that the proportion of water on the exoplanet in question is different from 0.7. The result is *statistically insignificant*. If the result had instead been that $P \leq 0.05$, we would conclude—still following the ritual—that the result is *statistically significant* and that the proportion of water on the exoplanet is not 0.7.

Let's review the procedure. To conduct a null hypothesis significance test, one must:

(1) Define a null hypothesis, usually a statistical model equivalent to zero effect or no difference. Specify no other models.
(2) Compute the likelihood of observing the actual data or any observation more extreme than the actual data, assuming that the null hypothesis is true. Call this sum of likelihoods P.
(3) If $P \leq 0.05$, conclude that the null hypothesis is false. It has been rejected. The observations are *statistically significant*. If instead $P > 0.05$, conclude that we cannot reject the null hypothesis. The observations are *statistically insignificant*.

There are some elaborations of this ritual, such as power analysis and corrections for multiple comparisons. But this core above remains intact in almost all cases.

3.2. Hopeful Illusions

Judging from surveys of graduate students and professors and even professors who teach statistics, scientists commonly hold to a number of hopeful illusions about the meaning of P-values. Before we dig into these illusions and dispel them, I must stress again, as I did in Chapter 1, that the intention here is not to scold. A majority of scientists have been taught these illusions, either actively in poor instruction or passively in reading the scientific literature. Many textbooks contain incorrect definitions of P-values, so it becomes unfair to blame students and faculty who are not experts in statistical inference. And even when scientists know that these illusions are fallacies, they may still behave as if they were true. So if you find yourself guilty of endorsing any of them, you are in very good company. The important thing is to find the courage and mastery of the material to challenge those who would coerce you into participating in these illusions.

Figure 3.3 shows the results of two surveys, conducted in the years 1986 and 2000 (Oakes 1986, Haller and Krauss 2002). These surveys were of academic psychologists, which is particularly shocking, since psychologists perhaps rely upon *P*-values more than any other field. This figure shows the percent of different categories of people endorsing different absolutely false statements about the meaning of a significant result of $P = 0.01$ in a simple experimental context. The top group of bars shows the percent in each category endorsing at least one false statement. The other groups of bars correspond to individual false statements. These statements are abbreviated in the figure. They are in full, in order from top to bottom:

(1) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.
(2) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
(3) You can deduce the probability of the experimental hypothesis being true.
(4) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
(5) You have found the probability of the null hypothesis being true.
(6) You have absolutely disproved the null hypothesis.

What all of these statements have in common is that they are false and they are hopeful. All of them overestimate rather than underestimate what *P*-values can tell us (Gigerenzer 2004). In my experience, the general atmosphere of understanding is no better in biology nor anthropology nor economics and has not improved since.

Before exploring the computational details of *P*-values and the consequences of their use, it might be helpful for many readers to see explanations of why each of these statements is false. Haller and Krauss (2002) provide such explanations, as well as a number of valuable heuristics for teaching about and interpreting such tests. I will provide slightly different explanations of these fallacies here, but readers who can admit to themselves that they believed any one of these would do well to consult Haller and Krauss (2002) and Gigerenzer (2004) and follow citations therein.

(1) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain

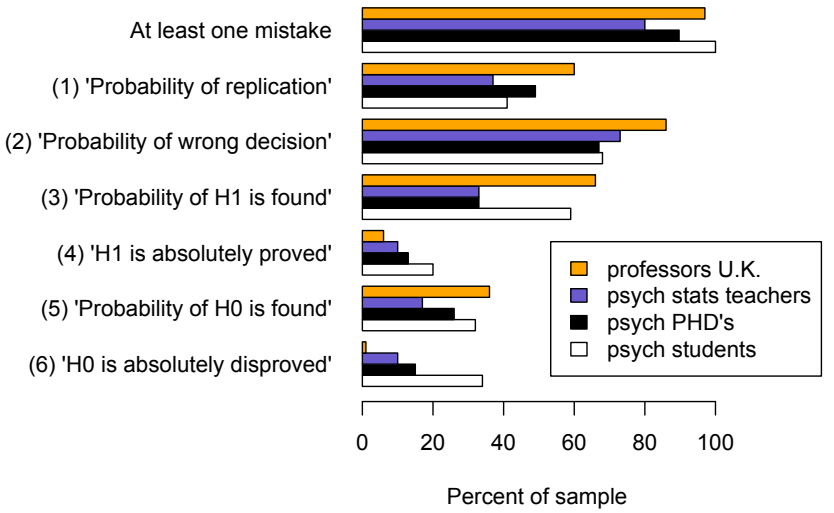

FIGURE 3.3. Two surveys of false beliefs about the meaning of *P*-values (Oakes 1986, Haller and Krauss 2002). See full text of these false statements in the main text.

a significant result on 99% of occasions. This statement is false foremost because any conclusions derived from a *P*-value are only valid under the assumption that the null hypothesis is true. If the null hypothesis is not true, then whatever the probability of replicating the results, it isn't provided by the *P*-value, which only applies to the null hypothesis. There is a deeper confusion usually lurking here, one of conflating Fisherian *P*-values with the kind of Type I error rates, α, advocated by Neyman and Pearson. These are not equivalent and usually not even compatible approaches to hypothesis testing.

(2) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision. Again, this is foremost false because any information we derive from the *P*-value is only valid when the null hypothesis is true. This statement implies that the *P*-value tells us the probability that the null hypothesis is true, which is instead what we must assume, in order to compute the *P*-value. As we learned in the previous chapter, $\Pr(D|M)$ is not the same as $\Pr(M|D)$. A *P*-value is more similar to a likelihood than it is to a posterior probability, and only posterior probability informs us about probabilities of models.

(3) You can deduce the probability of the experimental hypothesis being true. No alternative hypothesis factors in the calculation of a *P*-value, and so *P* contains no information about the truth value of any hypothesis other than the null. Even then, as I keep saying, *P* is not the probability that the null is true, because we must assume it is true to compute *P*.

(4) You have absolutely proved your experimental hypothesis (that there is a difference between the population means). There are a couple of key illusions here. The first is that probability can prove anything. Probabilities only corroborate hypotheses, and to pretend otherwise is to indulge in a kind of Tyranny of Popper (see Chapter 1). The second is again that we must assume the null hypothesis is true in order to compute *P*. As a result, *P* cannot tell us whether or not either it or any other hypothesis is true.

(5) You have found the probability of the null hypothesis being true. Yet again, since we must assume the null is true in order to compute *P*, it does not tell us if the null is true or not. $\Pr(D|M)$ is not the same as $\Pr(M|D)$.

(6) You have absolutely disproved the null hypothesis. Since probabilities cannot provide logical proof in the manner this illusion imagines, it is false. This is a common example of the Tyranny of Popper, manifested through the Tyranny of Fisher.

It is certainly true, however, that a *P*-value provides some information about the null hypothesis. What is it? Perhaps the easiest way to explain *P* is to compare it to the naive posterior we explored in the last chapter.

3.3. Comparing *P* to the Naive Posterior

A puzzling thing about the null hypothesis testing procedure is that no model other than the null model is ever made to predict the observations. It is a method of absolute evaluation. This is quite a change from the previous chapter, in which we used relative evaluation of models. In Chapter 2, we saw that the logic of probability gives us advice on which model (value of a parameter) the evidence supports, when we compute the posterior probability density. Under a uniform prior, the posterior is proportional to the likelihood. I called the resulting posterior a *naive* posterior, one informed only by the likelihood function. The likelihood function is computed (at least in the case of a single parameter family of models) by varying the model and recomputing the likelihood of the observed data. The data were constant, but the model was unknown, so we

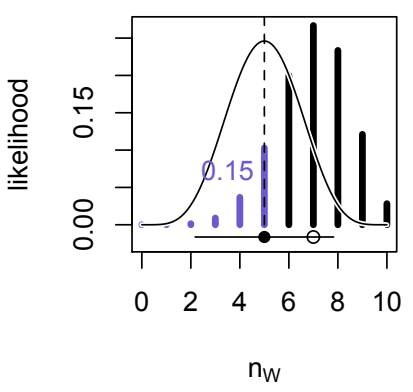

FIGURE 3.4. Computing both the *P*-value when 5 of 10 probes report "water" (blue bars) as well as the likelihood function for $n_W = 5$ (solid curve). The line underneath is the 95% confidence interval, with the maximum likelihood estimate (filled circle) and null model (open circle).

varied the model to see how models performed relative to one another in predicting the fixed data. Since the *P*-value is computed holding the model constant, at the null model, and instead varying the data, it's somewhat the opposite of using the naive posterior.

In FIGURE 3.4, I reproduce the plot from before, but now superimposing the likelihood function and the 95% confidence interval around the maximum likelihood estimate. The sum heights of the blue bars are again the *P*-value in this instance. The thin curve is the likelihood function, treating the horizontal axis as the expected samples from a family of models with different values of p_W. The horizontal line under the curve represents the 95% confidence interval. The solid circle on this line is the maximum likelihood estimate, $\hat{p}_W = 0.5$ in this case, and the open circle is the null model, $p_W = 0.7$. Notice that the fact that $P > 0.05$ does not also imply that the maximum likelihood estimate is $\hat{p}_W = 0.7$, the null model. Instead, we see that the null model, the open circle underneath, is included in the far right end of the 95% confidence interval, while the maximum likelihood estimate lies at the sample proportion.

When a test of significance fails, it is very common to act as if the best estimate of the parameter is the null model (usually zero, here 0.7). Are we really supposed to conclude here that there is not sufficient evidence to reject the null hypothesis? There are a couple of puzzles that arise immediate from this question. Let's consider each before turning to clearing up confusion about the meaning of *P*.

3.3.1. The maximum likelihood estimate always has higher likelihood. First, the null hypothesis has much lower likelihood than does

the maximum likelihood estimate, $p_W = 0.5$. You can compare the likelihoods of the two models quite easily:

R code 3.2

```
lik.mle <- dbinom( 5 , 10 , 0.5 )
lik.H0 <- dbinom( 5 , 10 , 0.7 )
lik.mle/lik.H0
```

```
[1] 2.391132
```

This number is the likelihood ratio of the maximum likelihood to the null model. The maximum likelihood estimate is more than twice as likely as the null, but our NHST procedure still has us failing to reject the null in favor of the maximum likelihood estimate. So obviously, whatever justification we might offer for favoring the null, it isn't simple likelihood alone. We might attempt to patch this up by comparing the likelihood of the null model to the total likelihood of all other models, or maybe just all the models less than the null. But this will just make things worse for the null, because most of the likelihood is bunched up around the maximum likelihood estimate. Let's try two other approaches, both of which will demonstrate concerns with NHST, but also teach some ways to compare approaches to statistical inference.

3.3.1.1. *A prior belief in the null model.* Another idea is to adopt a prior probability density that favors the null model. One can always find a prior that will lead to the null having higher posterior probability than the maximum likelihood estimate. The simplest way is just use a uniform prior with a spike in probability at the null. If that spike is tall enough, posterior probability will still favor the null model over the maximum likelihood model. In this way, NHST amounts to having an unstated prior preference for the null hypothesis.

Here's what such an approach would look like. I don't happen to think this is a credible way to arrive at the procedures of NHST, but going through this exercise now allows me to accomplish two goals. First, I get to introduce you to computing posterior probabilities using a simple technique called *grid approximation*. Second, this exercise will illustrate how strongly NHST tends to favor the null hypothesis.

First, we need to construct a prior probability distribution over possible values of p_W. So let's make a list of values of p_W in 0.001 increments. By dividing up a continuous variable like p_W in this way, we are doing something often called *grid approximation*, which approximates a continuous distribution with a finite number of discrete values. This is a handy technique for getting used to computing posterior probabilities,

especially for students who have never had a course in integral calculus. To make a list of models, we use our friend the `seq` command:

R code 3.3
```
models <- seq(from=0,to=1,by=0.001)
```

Now we want to start construction a prior probability distribution that is flat everyplace but at the null model. This will do it:

R code 3.4
```
prior <- rep( 1 , length(models) )
h0idx <- min( which( models >= 0.7 ) )
prior[h0idx] <- 10
prior <- prior/sum(prior)
```

The first line just makes a list of 1's of the same length as our list of models. The second and third lines then find the location (index) of the null model and place a 10 at that location in the prior. Finally, we normalize the prior so that it sums to one, because it is a probability density. This gives us a list of probabilities corresponding to each model in `models`. The probability of the null in this list is 10-times greater than that of any other model.

Now recall that Bayes' theorem is:

$$\text{Posterior} = \frac{\text{Likelihood}}{\text{Probability of data}} \times \text{Prior}.$$

In our current problem, this becomes:

$$\Pr(p_W|n_W, n) = \frac{\Pr(n_W, n|p_W)}{\Pr(n_W, n)} \Pr(p_W).$$

We've already constructed $\Pr(p_W)$, by assigning a probability to each unique value of p_W in `models`. Now we need the likelihoods and the probability of the data. After working through Chapter 2, computing the likelihoods is familiar to you:

R code 3.5
```
likelihood <- dbinom( 5 , size=10 , prob=models )
```

To compute the normalization constant, $\Pr(n_W, n)$, the unconditional probability of the observed data, it helps to realize that this probability is an average across models (values of p_W) of the likelihood of n_W, n. So if we multiply each likelihood by each prior probability, we get the average we are after. In code form:

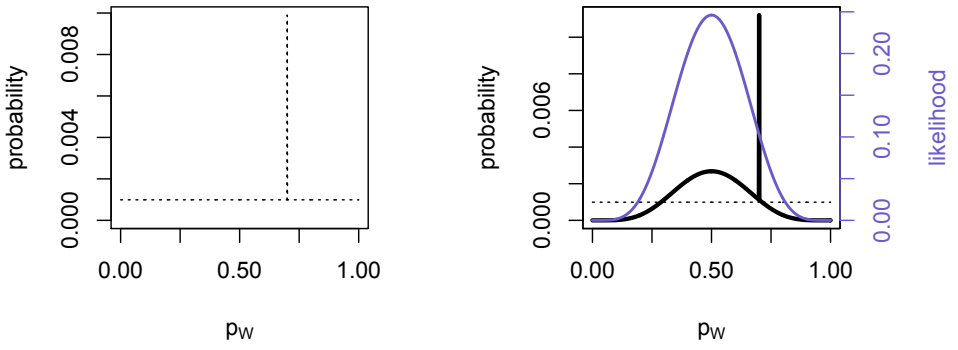

FIGURE 3.5. Prior probability and null hypotheses. On the left, an arbitrary prior probability density that strongly favors the model at $p_W = 0.7$, the null hypothesis. On the right, the posterior (black) still is maximized at $p_W = 0.7$, because the prior so strongly favored the null. With enough sampling, the prior can eventually be overwhelmed. In this way, prior probability of this kind behaves much like null hypothesis significance testing, in that it requires some threshold of evidence before a preexiting preference for the null hypothesis can be dislodged.

```
PrD <- sum( likelihood * prior )
```

R code 3.6

Finally, we combine all three pieces to compute the list of posterior probabilities of each value of p_W:

```
posterior <- likelihood * prior / PrD
```

R code 3.7

In FIGURE 3.5, I plot the prior we created and the posterior we computed from it. On the left, the plot with the dotted lines shows the prior probability of each value of p_W being very small and equal, but with a spike in probability at $p_W = 0.7$. On the right, the posterior probability in black is distinctly different from the likelihood curve in blue, because of the prior probability. The spike in posterior probability at $p_W = 0.7$ means that the null model still has larger posterior probability than any

other model, even the maximum likelihood estimate. If you play around with the code, you can show that, depending upon how strongly the prior favors the null, with enough sampling the maximum likelihood estimate will eventually surpass the null in posterior probability. But it might take a lot of sampling.

With enough data, any prior can be overwhelmed, as you saw in Chapter 2. So a prior probability like the one we created here will behave similarly to a null hypothesis test. However, this exercise is heuristic. Null hypothesis tests are not derived via any consideration of prior probabilities, except perhaps by rejecting their use. The point to take away here is that NHST acts as if we had a pre-existing preference for the null hypothesis.

3.3.1.2. *Is NHST accurate?* All that philosophizing is nice, but most of us are more interested in how a method performs. Does NHST work or doesn't it? To decide that, we need a comparison to some other method, as showing that NHST works better than guessing would hardly be surprising. But setting up a contest between NHST and some other approach is simple enough. Let's consider NHST as the choice between the null hypothesis, when $P > 0.05$, and the maximum likelihood estimate, when $P < 0.05$. In practice, that's what most people do, although as I've argued already in summarizing the posterior, it usually doesn't make much sense to discard the posterior in favor of a single model. But directly comparing inference under NHST and adopting maximum likelihood will make this lesson easier to understand.

Suppose we compute the accuracy of NHST and maximum likelihood estimation, across different true values of the parameter p_W, by simulating samples. Define accuracy as the distance of model (value of p_W) we adopt from the true value. Since we simulated the samples, we know the true value in each case, even though our NHST procedure acts as if it doesn't know. Here's the outline of how we might do this.

(1) Imagine the true value of p_W to be anything between zero and one.
(2) Under the true value of p_W from (1), generate a large number of simulated samples from 10 probes.
(3) For each sample from (2), calculate the P-value for the null hypothesis that $p_W = 0.7$ (or any other value).
(4) If $p < 0.05$, adopt the maximum likelihood estimate as the NHST estimate of p_W. Otherwise, adopt the null model as the NHST estimate of p_W.

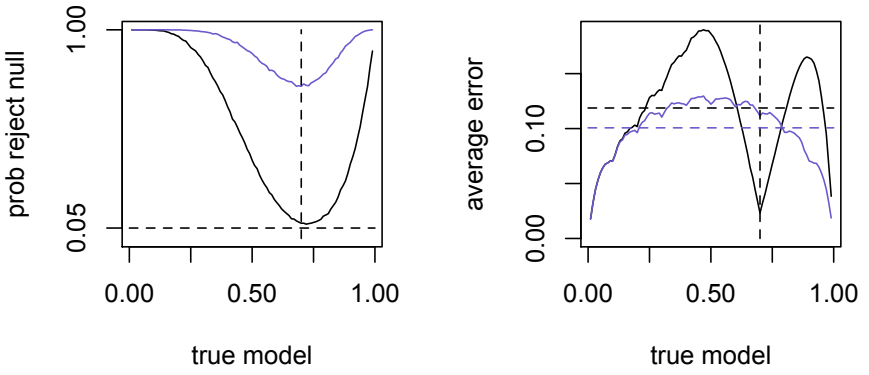

FIGURE 3.6. Behavior and accuracy of null hypothesis significance testing (NHST) compared to the maximum likelihood estimate. In both plots, the black curve represents NHST and the blue curve the maximum likelihood estimate. The vertical dashed line shows the location of the null model, at $p_W = 0.7$. For each possible value of p_W on the horizontal axes, 10-thousand simulated samples of 10 probes were used to compute decisions to accept/reject the null model that $p_W = 0.7$ (NHST) and to compute the maximum likelihood estimate. On the left, NHST adopts the null model much more often than maximum likelihood does. This behavior explains why in the right-hand plot NHST is less accurate, unless the true model is very close to the null model.

(5) Compute the distance between the true value of p_W from (1) and the value we adopted in (4).
(6) Go back to (1) until there are no more unique values of p_W to consider as true values.

In FIGURE 3.6, I show the results of doing this in R. The code is a little more involved than usual, so I include it at the very end of this section, rather than right here. The black curves in both plots represent decisions taken under NHST. The blue curves in both plots represent instead the decision of always adopting the maximum likelihood estimate. The horizontal axis in both plots are different true values of p_W used in simulations. The vertical dashed lines mark the location of

the null hypothesis, $p_W = 0.7$. At 0.001 increments, I simulated 10-thousand samples of $n = 10$ at each imagined true value of p_W. For each simulated sample, I calculated the P-value and the maximum likelihood estimate, $\hat{p}_W = n_W/n$.

In the lefthand plot, I show the proportion of simulations in which the null hypothesis was rejected. The horizontal dashed line is the significance threshold of 0.05. When the null is true, NHST is supposed to falsely reject the null model this often. Notice that NHST, in black, accepts the null much more often than adopting the maximum likelihood estimate. And NHST doesn't even reject the true null 5% of the time, as it advertises. It does get pretty close, though, when the true model is near $p_W = 0.7$. Adopting the maximum likelihood estimate, in blue, always "rejects" the null hypothesis, unless the null and the maximum likelihood estimate are the same. In a small sample like this, that is possible, but rare. It happens only when the null is very close to the true value. When the true value of p_W is very far from the null model, NHST usually rejects the null. But in a wide region around the null, NHST accepts the null over the maximum likelihood estimate.

What kind of accuracy does NHST achieve by acting this way? In the righthand plot, I show the average accuracy of each approach. The error on the vertical axis is the distance from the adopted model, whether null or maximum likelihood, to the true value on the horizontal axis. Notice that when the true value of p_W is very close to the null model, then NHST is more accurate than using the maximum likelihood estimate. A little further out, however, and NHST is less accurate than the maximum likelihood estimate. Very far out from the null model, both methods adopt the maximum likelihood estimate and are equally accurate. If we are willing to entertain the notion that all values of p_W on the horizontal are equally probable, in our state of ignorance of the exoplanet, then the average error for NHST in this example is 0.1188 (horizontal dashed line) and that of the maximum likelihood estimate 0.1007 (blue horizontal dashed line). If we make the sample size larger, this relationship does not change. NHST will still always perform worse, averaged across possible true models. And in case you think some other procedure like a likelihood ratio test will do better, likelihood ratio tests actually do worse. They do worse because they are based on normal approximations to binomial probability, in this case.

The comparison between MLE and NHST here is actually handicapping the Bayesian approach. NHST would look even worse, if we used the entire posterior, rather than just the MLE. The MLE averages worse accuracy than the average of the posterior, whenever they differ.

They will differ when the posterior is not symmetrical, as is common for binomial probabilities like these.

The only way to make NHST more accurate on average than naive maximum likelihood is if the null hypothesis, or values near it, are truly more probable in real samples. Then NHST's greater accuracy near the null model would make up for its terrible performance further away. So we see again that NHST is much like adopting some kind of prior probability that favors the null hypothesis.

How reasonable such a preference for the null appears to you will depend upon your view of how science should work. It will also probably depend upon whether or not the null is your pet hypothesis or not. What I think most of us can agree upon, however, is that one should either be explicit about the form and magnitude of this preference—make it into a real prior probability distribution—or rather leave it out of the analysis and include it instead in the discussion of how current results reflect on models. In it's usual form, the prior preference NHST endows upon the null model is unconscious and of unmeasured strength. It is leading many scientists to think that logical analysis has accepted the null hypothesis, when in fact some vague preference for the null has accepted it. Checking model predictions is fine, but why check the predictions of the null? Why not check instead the MLE?

Finally, here's the code to compute the values used in FIGURE 3.6. First, we make a convenient function that computes the *P*-values for the binomial sampling model. This code is functionally equivalent in this context to the built-in R command `binom.test`, but much simpler.

R code 3.8

```
calc.p <- function(n=10,nw=1,pw.h0=0.7) {
    if ( nw/n <= pw.h0 ) { p <- sum( dbinom(0:nw,n,pw.h0) ) }
    else { p <- sum( dbinom( n:nw , n , pw.h0 ) ) }
    p
}
```

Now we need a function that simulates sampling under some true value of p_W and simulates rejecting the null hypothesis. This function returns four different values: (1) the proportion of simulations in which the null was accepted, using NHST; (2) the proportion of simulations in which the maximum likelihood estimate (MLE) and the null coincide; (3) the average absolute distance between the accepted model and the true model, under NHST; (4) the average absolute distance between the maximum likelihood estimate and the true value.

R code 3.9

```
sim.false.accept <- function(n=10,pw.true=0.1,pw.h0=0.7,
    R=9999,alpha=0.05) {
    nw <- rbinom( R , size=n , prob=pw.true )
    p <- sapply( nw , function(z) calc.p(nw=z,n=n,pw.h0=pw.h0) )
    accept.null.nhst <- ifelse( p > alpha , 1 , 0 )
    accept.null.mle <- ifelse( nw/n==pw.h0 , 1 , 0 )
    avg.error.nhst <- mean( abs( accept.null.nhst*(pw.true-pw.h0)
        + (1-accept.null.nhst)*(pw.true-nw/n) ) )
    avg.error.mle <- mean( abs( pw.true - nw/n ) )
    c( mean(accept.null.nhst) , mean(accept.null.mle) ,
        avg.error.nhst , avg.error.mle )
}
```

We'll use this function to repeat such simulations across all possible true values of p_W. The valuable trick here is using `sapply` (Simplified Apply), which is a command that allows us to pass each value in a list to a function, storing the individual results in a new list. For example, to square each element of a list of numbers `x`, you could just enter `x^2` in R. But suppose you are working with a function that doesn't recognize lists. In that case, you want to use `sapply` or a similar command. You could do it like this:

R code 3.10

```
x <- 1:10
x.squares <- sapply( x , function(z) z^2 )
```

That `z` inside of `sapply` is just an arbitrary name. You could call it `donut` if you want to. It just takes on values from the list of numbers being fed to it, one at a time. Here's the code that applies our `models` to `sim.false.accept`:

R code 3.11

```
models <- seq(from=0.01,to=0.99,by=0.01)
sims <- sapply( models , function(z)
    sim.false.accept( pw.true=z , n=10 , pw.h0=0.7 ) )
```

The use of `sapply` above takes each value in `models`, hands it to `sim.false.accept`, and then stores the four-part result in the matrix `sims`. So we end up here with a matrix of results, in which each column is an individual call to `sim.false.accept` and each row is one of the elements returned by `sim.false.accept`. Now you can use standard plotting commands like `plot( 1-sims[1,] ~ models )` to replicate the figure. The logic of

`sapply` can be awkward at first. But there will be other examples. So be patient with yourself and this trick will make sense eventually.

3.3.2. Why predict what did not happen? A second puzzle is that it's not clear why the likelihoods of events that didn't happen should matter. Every blue bar to the left of $n_W = 5$ is the likelihood of an event that we have not observed. This isn't just about judging the fit of a dress by an absolute standard—the standard of fit here imagines ways that the dress could fit worse than it actually does and uses those imaginary flaws to decide how well the dress does fit. Statisticians have long worried about this issue, especially Bayesian statisticians. Here is what Harold Jeffreys had to say about *P*-values in 1939, in his book *Theory of Probability* (page 315):

> 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. On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law, not against it.

This would be like convicting an innocent man because we have not found his prints on the murder weapon. To return to the dress metaphor, this is like deciding a dress does not fit well enough, because it could fit worse.

In practice, however, the usual trouble with including these unobserved events in *P* is that they tend to overstate the evidence for the null hypothesis. So we might say more often that the use of *P* implies that a hypothesis that may be wrong may be accepted because it predicts observable results that have not occurred. This would be like acquitting a guilty man because you found his prints on the murder weapon. In this section, we'll examine a consequence of the fact that *P* is defined as a sum across imagined data rather than across models: The magnitude of *P* depends upon how sampling works, even when the parameter of interest doesn't affect sampling in any way. The major problem arises when sample size is not fixed but instead varies randomly or contingently upon what events occur.

Suppose our probe to the exoplanet carries 20 surface probes, each capable of landing in a different random location on the exoplanet. However, there is a chance that any surface probe burns up on entry into the exoplanet atmosphere, and as a result, only 10 of them ever report back to the mother probe. Suppose still that the observed data are

5 "water" in 10 reports, for a sample proportion of 0.5. If we could replicate this "experiment," the sample size would not always be $n = 10$, but would instead vary randomly according to the probability of a probe's being destroyed on entry. That is, sometimes all 20 would report back, sometimes only 3 would report back. This means we now have to consider variable sample size.

While comparing the maximum likelihood estimate and the null hypothesis based upon their respective likelihoods remains unaffected by the potential for variable sample sizes, NHST is dramatically altered by it. In practice, computer software always assumes constant sample size, but as I'll argue at the end of this section, only in the most disciplined experimental sciences is such an assumption reasonable.

Let d be the probability any surface probe burns up on entry. Then the likelihood of observing n_W reports of "water" from n surface probes that actually manage to land, out of 20 probes that were launched at the exoplanet, is:

$$\Pr(n_W, n|p_W, d) = \Pr(n|d, 20) \Pr(n_W|p_W, n). \tag{3.1}$$

This is just to say that the likelihood is the chance of both n probes landing successfully and receiving n_W reports of "water." $\Pr(n|d, 20)$ is the probability of achieving a sample size of n, conditional on d and 20 attempts. $\Pr(n_W|p_W, n)$ is just the same old binomial likelihood we've been working with since Chapter 2. So what exactly is the distribution $\Pr(n|d, 20)$? This too is just binomial, with 20 attempts and n successes, each success having probability $1-d$, the chance a surface probe is not destroyed during entry. We're assuming that the number of probes that land successfully does not influence where they land, and so the the probability on the left is just the product of the probabilities on the right. Note that the general conclusion we are about to arrive at does not depend upon this assumption of independence between landing location and landing successfully. The assumption of independence is just to make the example easier to work through.

Suppose the chance of a probe being destroyed is about 2%, $d = 0.02$. Now let's produce some code to make R compute P-values for our landing probes data. First, we need the likelihood function above, and just multiplying two calls to `dbinom` will do the trick. To compute the likelihood of observing 5 "water" when 10 of 20 surface probes survive, using the null hypothesis that $p_W = 0.7$:

R code 3.12

```
dbinom(10,size=20,prob=1-0.02) * dbinom(5,size=10,prob=0.7)
```

```
[1] 0.0840926
```

The first call to `dbinom` computes the probability of 10 out of 20 surface probes surviving. The second call computes the probability of getting 5 reports of "water" from 10 probes, with a 0.7 probability of "water" from each. We're going to want to use this code again and again, to sum up likelihoods of all the unobserved events that contribute to the *P*-value, so let's package it up into an R *function*. A function is just a bundle of R commands that we give a name, so we call use it again and again. Entering the following into R will define the function `dprobe`:

R code 3.13

```
dprobe <- function( nw , n , nmax , d , pw ) {
    prn <- dbinom( n , size=nmax , prob=1-d )
    prnw <- dbinom( nw , size=n , prob=pw )
    prn * prnw
}
```

This function computes the likelihood of observing `nw` reports of "water" from `n` probes, when `nmax` probes were launched, each with probability `d` of being destroyed on entry, and finally where each surviving probe has a chance `pw` of landing in water. Once you have defined the function in R, it works just like every other R command. For example, to compute the likelihood of the observed data under the null hypothesis:

R code 3.14

```
dprobe( nw=5 , n=10 , nmax=20 , d=0.02 , pw=0.7 )
```

```
[1] 1.590949e-13
```

Readers who didn't grow up coding like I did may reasonably wonder what that `e-13` means. That is how computers typically format scientific notation. The `e-13` just means $\times 10^{-13}$.

The *P*-value for the observation $n_W = 5, n = 10$ is the sum of all likelihoods for all possible events with sample proportion more extreme than the observed. So that means we want to add up the likelihoods of all possible combinations of n_W and n where $n_W/n \leq 0.5$. There are some slick compact ways to compute such a thing, but for the sake of clarity, I'll just use some loops. Here's a function to compute the *P*-value:

R code 3.15

```
pval.probe <- function( d ) {
    p <- 0
    for ( n in 1:20 ) {
        for ( nw in 0:n ) {
            if ( nw/n <= 0.5 )
```

```
                p <- p + dprobe(nw=nw,n=n,nmax=20,d=d,pw=0.7)
        } # nw
    } # n
    p
}
```

This function loops over all possible values of `n`, the number of surviving surface probes, from 1 to 20. We don't consider cases in which zero probes survive, because in that case, the sample proportion n_W/n is undefined—no information means no estimate. For each value of `n`, the function then loops over each value of `nw` from zero to `n`. These are all the possible counts of reports of "water." For each of these, it adds the likelihood under the null hypothesis, $p_W = 0.7$, to the growing P-value, stored in the symbol `p`.

Now we want to examine how changes in d affect the P-value. So we make a list of values of d and compute P at each:

R code 3.16

```
d.list <- seq(from=0,to=1,by=0.01)
p.list <- sapply( d.list , function(z) pval.probe(d=z) )
```

FIGURE 3.7 shows these P-values as a function of d. The solid curve is the P-value and the horizontal dashed line is just the conventional significance level of $P = 0.05$ for reference. Notice that p increases drastically as d increases, up to very high probabilities of destruction, at which point we begin to get no information from the probes. The blue line shows the likelihood ratio of the maximum likelihood estimate, $\hat{p}_W = n_W/n$, to the null model, $p_W = 0.7$. You can calculate this ratio as a function of d in a similar way:

R code 3.17

```
Lr.list <- sapply( d.list , function(z)
    dprobe(5,n=10,nmax=nmax,d=z,pw=5/10) /
    dprobe(5,n=10,nmax=nmax,d=z,pw=0.7) )
```

This ratio is insensitive to d. It has the same value, regardless of how probable it is for a probe to be destroyed during entry. Why is this?

We can use the analytical likelihood function for this model, Expression 3.1 (page 122), to compute the likelihood ratio between the maximum likelihood estimate and the null model:

$$\frac{\Pr(n_W, n|\hat{p}_W, d)}{\Pr(n_W, n|0.7, d)} = \frac{\Pr(n|d, 20)\Pr(n_W|\hat{p}_W, n)}{\Pr(n|d, 20)\Pr(n_W|0.7, n)} = \frac{\Pr(n_W|\hat{p}_W, n)}{\Pr(n_W|0.7, n)}.$$

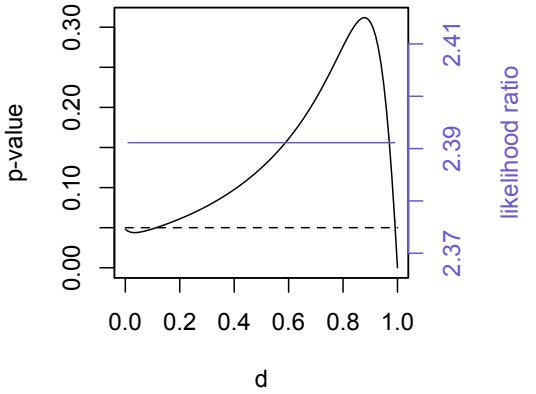

FIGURE 3.7. Comparing P and the likelihood ratio when sample size varies. The solid curve is the P-value of observing $n_W = 5, n = 10$, as a function of d. The dashed line is conventional significance, $P = 0.05$. The blue line is the likelihood ratio of the maximum likelihood estimate, $\hat{p}_W = n_W/n = 0.5$, to the null model, $p_W = 0.7$. While the P-value is dramatically influenced by how the data are sampled, the likelihood ratio is invariant to it.

The probability of destruction d has no influence on the likelihood ratio, because whatever the likelihood $\Pr(n|d, 20)$, it is the same on both the top and bottom, and so it just divides out.

What is going on here? The reason that the P-value is so dramatically altered by the way the sample size is determined is because P is a sum of events that have not been observed. As a result, the likelihood of these unobserved events can have big effects on the magnitude of P. The more likely it is for a probe to be destroyed during entry, the more ways there are to get a sample proportion equal-to-or-more-extreme than the observed proportion. Most of the time, the probabilities of these multiplying unobserved events contribute to P and make it harder and harder to reject the null model. But notice that initial increases in d actually reduce the P-value. Either way, the dependency of P on our assumptions about sampling is a logical consequence of how the P-value is defined. In contrast, the likelihood ratio is undisturbed by variation in sampling. The likelihood ratio is unaffected by the nature of sampling, because the

parameter p_W has no role in determining sample size. Likelihood ratios vary models and compare their ability to predict the sample we actually got. NHST, however, does not vary the model, p_W, but instead imagines new data. This partly bases the decision about which value of p_W to favor upon probabilities that p_W does not influence, $\Pr(n|d, 20)$.

One way to try and rescue P-values here is to note that it would make sense then to define the P-value as the probability of data-or-more-extreme data only when sample size is imagined to be constant across replications. In practice, this is what everyone (everyone's computer) implicitly assumes anyway. If we go this route, however, there are other problems. Sample size is rarely fixed in the real world, and usually we don't even know how it is determined. Many social psychologists, for example, just run experiments for a fixed duration and take however many participants they can get in that time frame. When we analyze data collected by someone else for another purpose, it's not clear what we should believe about sample size across replications. And while everyone knows it's forbidden, sometimes scientists collect data only until they achieve statistical significance. In all of these cases, if you want to actually interpret P-values according to their definition, as a probability of data-or-more-extreme-data, then imagining that sample size is fixed means that the probability you compute will be wrong. It will not have the long-run properties you want it to have. In the wicked case of collecting data until statistical significance, then $P = 1$, by definition, after correctly accounting for the sampling model.

In summary, anytime sample size can vary, then computing a P-value requires taking account of the distribution of sample sizes. Just comparing models (parameter values) using likelihoods (naive posterior probability) does not suffer from this problem, however. Almost no one ever worries about this in practice, but the issue is always there.

My own belief, however, is that this logical snag with P-values isn't the most important reason to reject their use. It certainly is hard to defend the use of events that did not happen when deciding upon the value of a parameter. But even if we could justify always computing them under constant sample size assumptions, it wouldn't make them perform better than maximum likelihood, unless the null hypothesis really were more probable or somehow less costly to adopt than another model. This turns our attention to other issues.

3.4. The need to develop the alternative model

My principle concern with the use of *P*-values, and indeed any other method of absolute evaluation used in isolation, is that they discourage scientists from developing research models. When all one has to do in order to support a research hypothesis is to reject the null model of "randomness," then there is little incentive to develop the research hypothesis into a mature predictive model. It is possible that null models do not make substantially different predictions than the implied alternatives. But this fact may never be discovered, unless we actually force both (or all) models to predict the observations. As a consequence, scientists sometimes accept the null hypothesis for no reason other than it was the only model they fit to the data. They conclude the alternative is worse, but they never fit the alternative to the data, so the claim is counterproductive and actively retards the progress of knowledge.

It's hard to force the proportion of water problem into an example of this problem. The possible models that we've learned so far are too simple to exhibit the difficulties. Later in the book, we'll examine in detail an infamous case of this problem, the use of "neutral" models in evolutionary biology. There's nothing wrong with neutral models, as long as they have company. Rather, the sort of data commonly used to test them is not very capable of differentiating them from non-neutral models. Many scientists never realized this, because they did not actually specify any selection model. They just fit the neutral model and it was good enough by an arbitrary standard, so they rejected selection as a candidate force. The plague of naive neutral modeling spread rapidly, to the point that a few papers actually conducted the same exercise of the distribution of baby names in North America, concluding that the choice of baby names is random, despite decades of work in sociology to the contrary. But since no non-neutral model of name choice was fit to the data, the possibility that the data could not detect non-neutrality, even if it is there, never came up.

As usual, I have a hard time holding the researchers responsible in these cases. They've been the victims of bad scientific norms. They were taught by prestigious PhD programs to test theories by constructing only models in which the forces of interest are not present. The norms are the villains here. In defiance of those norms, we must insist that research hypotheses be specified and explicitly compared to any null or neutral models. In doing so, any scientist will immediately realize when the data cannot distinguish the models. They will also realize that there are many ways to make a non-neutral model, and so rejecting the neutral model does little to tell us about nature.[42]

3.5. How to Interpret *P*-Values, If You Must

Sadly, journals are full of *P*-values and will be for some years to come. Norms are changing, such that it is harder to find papers that do not report at least estimates and standard errors, with *P*-values tacked on the end. In those cases, you can just ignore the *P*'s and use the estimates and standard errors to construct the approximate naive posterior. But it is still common enough to see publications in highly-ranked journals that report only estimates and *P*-values or the statistics *P*-values are derived from, like *t*-values. There is still a lot of valuable science in these papers, if we can cut through the vagueness of the significance tests. Here is some advice on extracting as much information as you can from these papers.

3.5.1. Confusing effect size and sample size. One difficulty in interpreting *P*-values is that they are an unclear mix of information about *effect size*—how far the maximum likelihood estimate is from the null model—and *precision*—how confident we can be in the maximum likelihood estimate. It's not possible to know from a *P*-value alone whether we have accepted/rejected the null model because the effect is small/large or rather because the precision is low/high. The distinction is crucial. Let's consider some examples.

Suppose for the sake of illustration that there are actually four different exoplanets and we have sent probes to each of them. The results of null hypothesis tests for each are displayed in the table below. In each case, the null model is $p_W = 0.7$.

Exoplanet	*P*-value
Fomalhaut b	0.15
HR 8799c	0.048*
83 LeonisBb	0.39
51 Pegasi b	0.044*

I display an asterisk next to each significant result, as is customary for the results of NHST. Now what can you conclude from this information? Which exoplanets are likely to have water coverage similar to the Earth? It is true that HR 8799c and 51 Pegasi b are significantly different from the Earth, but without knowledge of how many probes reported back in each case, it's impossible to know whether this is due to estimates very different from $n_W/n = 0.7$ or rather few surface probes surviving entry and managing to report back a reading. For example, 51 Pegasi b could be statistically significant either because the sample proportion of water reported back by the surface probes was very small, relative to the null model, or rather because the sample proportion was close to

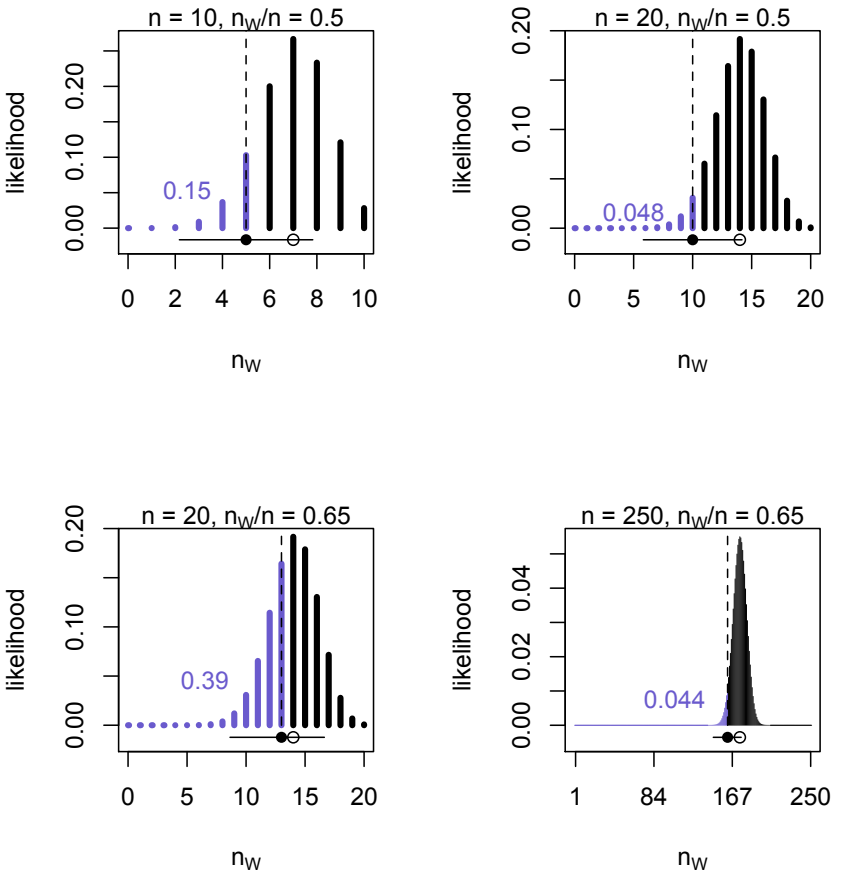

FIGURE 3.8. The ambiguity of *P*-values with respect to effect size and precision. See explanation in main text.

0.7 but many probes reported back, producing a precise estimate. The *P*-value alone cannot tell us which of these two things happened. This point is widely appreciated. What is less well appreciated is that, once we have the information necessarily to resolve ambiguity in the *P*-value, the *P*-value provides no additional information of value.

It is easier to appreciate how little the table above actually tells us if we fill in the examples with sample size and precision information. In Figure 3.8, I show all four null hypothesis tests, resulting from our

probe reconnaissance of each exoplanet. Each plot shows possible observations along the horizontal axis and likelihoods under the null hypothesis on the vertical axis. The blue bars are included in summing the *P*-value, which is shown in blue. The observation in each case is shown by the location of the vertical dashed line. The filled circle under each set of likelihoods is the maximum likelihood estimate, and the open circle is the null model. The horizontal line under the likelihoods is the 95% confidence interval. Now we can distinguish large effects from precise small effects. In the upper-left, 10 surface probes managed to report back, and half of them said "water." This result is statistically insignificant, even though the sample proportion is identical to the plot in the upper-right, in which 20 surface probes survived. The results in the upper-right are statistically significant, with exactly the same effect size, because they are more precise. In the lower-left, again 20 surface probes survive to report back, but this time 13 of them, a proportion 0.65, reported "water." This is a very small effect size, due to how close 0.65 is to the null model. As a result, this result is statistically insignificant. But in the lower-right, the exact same effect size from a stunning 250 successful probes (perhaps this mission was launched later than the others and so had a larger payload of surface probes to deploy) leads to a statistically significant result, even though this exoplanet we must conclude has a coverage of water very similar to the Earth.

Again, the point of these examples is that small *P*-values alone can result from either estimates that differ greatly from the null model or from the good precision of small estimates. To resolve this ambiguity, we need to know the estimate and the sample size, with which we could describe the precision by a confidence interval (as in Figure 3.8) or just by plotting a likelihood function. Let's augment our previous table with maximum likelihood estimates and 95% confidence intervals, to see how much more information this gives us.

Exoplanet	*P*-value	$\hat{p}_W$	95% confidence interval
Fomalhaut b	0.15	0.50	0.22–0.78
HR 8799c	0.048*	0.50	0.29–0.71
83 LeonisBb	0.39	0.65	0.43–0.83
51 Pegasi b	0.044*	0.65	0.59–0.71

Now we can more easily see that the evidence supports stronger inferences about the proportions of water on HR 8799c and 51 Pegasi b than for the other two exoplanets. Notice also that in every case, the 95% confidence interval includes the null model, $p_W = 0.7$. This should drive home the point that inferences from the naive posterior are not

the same as inferences from the *P*-value. In special cases, statistical significance is attained exactly when the null model falls outside the 95% confidence interval. But in general, this is not true. And since we already saw that one can move the *P*-value around greatly just by changing the sampling model, there may be very little correspondence at all, once the *P*-value is correctly computed. At the end of Chapter 2, I already argued that conducting dogmatic tests of statistical significance with confidence intervals abuses the nature of those intervals. Here I emphasize also that tests using confidence intervals give different answers than tests using *P*-values. We have now not only an arbitrary threshold at 5%, but now also more than one arbitrary way to compute that 5%, and these two methods disagree.

So let's ask again: What do we learn from the *P*-value? First, if we know sample size, we can with some difficulty infer effect size from *P*. The *P*-value is strongly correlated with sample size, because for any given effect size, *P* declines as sample size increases. So it is possible to compute, for a given *P*-value and sample size, the value of the maximum likelihood estimate. But it would be much easier to just report effect size and precision. And once you know the effect size and precision, all that knowing *P* adds is information about how strong your prior preference for the null hypothesis is.

Second, if we know effect size, then with some difficulty we can estimate precision from *P*. For a given estimate, it is possible to compute the sample size at which the estimate will become statistically significant. Unless the estimate is exactly the same as the null model, with enough data, the effect will always become significant. For very small effect sizes, it will take a huge amount of data. For very large effects, it will take only a few samples. Again, it would be much easier on everyone—scientist and readers alike—to just report the estimate and its precision to begin with.

While I still encounter papers that report nothing more than *P*-values and some vague measure of effect size like an *F* value, most people are aware that *P*-values alone are not sufficient for interpreting statistical models. But many people still base decisions about research hypotheses upon the value of *P*. They are using *P* to choose parameter values, to choose among models. And since computing *P* doesn't actually compare models, the procedure is an absolute evaluation where we need a relative one. Certainly one could invent other criteria that lead us to prefer the null model, but those criteria will not be how well the model predicts the data, compared to the other models.

So again we see that appeals to *P*-values are like appeals to prior belief in the null model. When $P > 0.05$, the evidence at hand is not strong enough to overcome the prior. When $P < 0.05$, it is. In neither case is the prior belief specified with any precision, so it's difficult to follow the chain of inference that leads to a concern with *P*.

I'm pretty down on classical *P*-values, as the reader can see. But I did say at the beginning of this chapter that I reserve an important role for checking individual models against data in an absolute evaluation. It's just that *P*-values, because they actively reject or ignore relative evaluation, and because they are sums of events that we have not observed, are not the form of absolute evaluation that we seek. Or at least, there is no reason to think that the same threshold of predictive quality, and the same measure of it, is the right standard to adopt for all models, for all types of data, in all fields, whatever the purpose of our statistics. Indulge me for a couple more difficulties that arise from reliance on *P*-values, to support that point.

3.5.2. Statistical significance is not itself significant. Here's a common scenario. A researcher has two variables that each might help predict an outcome of interest. A regression (coming in detail in the next chapter) is fit that includes both of the variables. The researcher's software produces a *P*-value to correspond to each variable, and only one of them is significant. At this point, it is very common to conclude that only the variable with a significant outcome helps predict the outcome variable.

The logical fallacy in this kind of reasoning is the assumption that the difference in significance is itself significant.[43] It might be, but it might also not be. This is not a criticism that relies upon rejecting NHST, as your author does. Instead, it is an inconsistency in how scientists interpret tests of significance. If we adopt a uniform criterion of significance and compute *P*-values as is standard, then it does not necessarily follow that a difference in statistical significance implies that the difference between two estimates would itself be significant.

This is confusing, without examples. Unfortunately, since we haven't introduced a way to include more than one parameter (model dimension) in our models yet, we're not ready to analyze examples of this issue. So let's use an abstract example that goes straight to estimates and the standard errors of those estimates (Figure 3.9). The curves in this figure are naive posteriors, so they show the maximum likelihood estimates and uncertainty for three different parameters. To take a common example from the scientific literature, suppose these estimates refer to rates of recovery from different drugs, labeled "1" and "2." The effect of

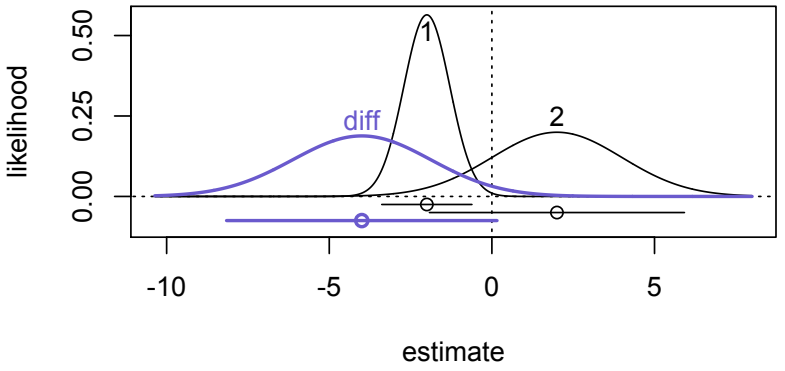

FIGURE 3.9. A difference in statistical significance is not always statistically significant. In this figure, the naive posteriors for three different estimates are shown, as well as corresponding 95% confidence intervals. The estimates labeled "1" and "2" differ in their statistical significance: "1" is significantly different from zero, while "2" is not. The naive posterior of their difference, labeled "diff" and shown in blue, is however not significantly different from zero, meaning that by the same standard of statistical significance, one cannot conclude that "1" and "2" are not the same.

the first drug is negative and significantly different from zero, if we use the confidence interval to conduct a significance test—you know I am not endorsing this procedure, but it is very common. The effect of the second drug is positive, but more variable, and so this drug's effect is not significantly different from zero. In such a situation, it is routine to see scholars concluding that the drug with the significant result is less effective than the one with the insignificant result. This does not follow. You can see this clearly, if you now inspect the blue curve, which is the naive posterior of the difference between the two black curves. The maximum likelihood difference is indeed negative, but the variance of this density is very large, and so the difference is not significantly different from zero, by the NHST based upon the confidence interval. The same logic applies however you compute the *P*-value, but using confidence intervals in this case makes the lesson more transparent.

The key reason that differences in significance are not always significant is that such null hypothesis tests do not compute *P*-values for the model that the two drugs—or two planets, or whatever—are the same. To do that, we'd need to structure the statistical model differently, so that it includes a parameter (a model dimension) corresponding to the *difference* between the two drugs. In the next chapter, we'll see how to do that, but even then, it won't be very useful to test the null hypothesis of zero difference, for all the reasons we've already encountered in this chapter. But it will help you understand why concluding that two estimates are significantly different from one another, just because one or both is significantly different from the null model, is illogical.

If you doubt that scientists commit this fallacy, start looking for it. Nieuwenhuis et al (2011) surveyed the neuroscience literature and found that about half of the papers published contained this fallacy. Whenever you see a claim that some variable has an effect and some other does not, because the first is significant but the other is not, then you are seeing an example of this fallacy. Now, it might be true that the difference between the effect sizes of the two variables is indeed significant, but it does not follow from testing if each effect is different from zero (or any other null model). In a later chapter, you will see how the use of *interactions* can address this problem. But even then, it won't make sense to focus on significance. Instead, the interest is in estimating the difference between two categories in the data. Estimating the response in each category is different than estimating the difference.

A closely related fallacy is to use the overlap of confidence intervals of different estimates to conclude whether they are significantly different from one another. Figure 3.10 illustrates this scenario. Again we see the estimated effects of two drugs, labeled "1" and "2," and their difference, in blue. I have drawn this example such that neither drug is significantly different from zero, and their confidence intervals overlap considerably. Nevertheless, the naive posterior of their difference is indeed significantly different from zero, by the same standard. Never use overlap or lack of overlap in confidence intervals to make inferences about the posterior density of a difference. Just estimate the difference, to be sure.

I encourage the reader to experiment with examples of this kind. You can use the `show.naive.posterior` function from the code LIBRARY that accompanies this book to make these plots yourself. The code to produce Figure 3.10 is:

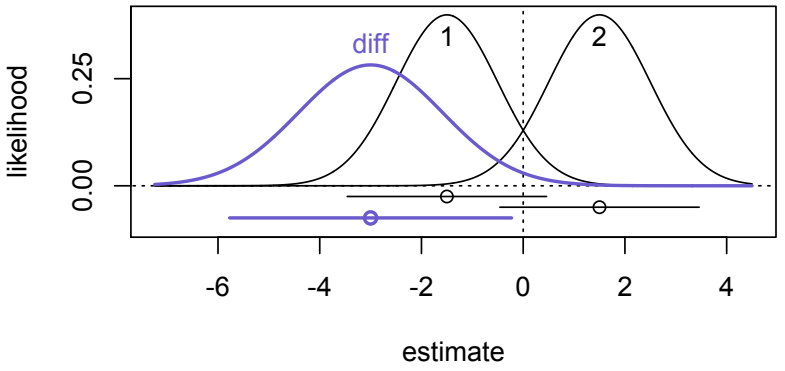

FIGURE 3.10. A difference between estimates may be significantly different, even when the confidence intervals of the estimates overlap.

R code 3.18

```
estimates <- c( -1.5 , 1.5 )
variances <- c( 1 , 1 )
ests <- c( estimates , estimates[1]-estimates[2] )
ses <- c( sqrt(variances) , sqrt( sum(variances) ) )
show.naive.posterior( ests , ses , label=c(1,2,"diff") ,
    lwidths=c(1,1,2) , cols=c("black","black","slateblue") )
```

Just change the values in the `estimates` and `variances` symbols to change the location and precision of "1" and "2." The difference will be computed by the third and fourth lines. The last line plots the estimates.

3.6. Defenses of *P*-values

Published defenses and rebuttals thereof.

That weird BBS article. That paper in PRSB. The econ defense. Is there an example in psychology, other than the BBS article?

What about Mayo? Not really a defense of NHST, but seems like it much of the time. Best paper is probably the regression assumptions paper: Mayo and Spanos 2004. Agree that we need more than posterior probability to evaluate model assumptions and absolute performance. This segues to next section.

3.7. Checking assumptions

Brief comments on other ways to provide absolute evaluations, aimed at critiquing model structure. Basic message is that no universal method exists, because models are meant for different purposes.

Jaynes on testing assumptions with significance tests (Jaynes 1985 page 351):

> It would be very nice to have a formal apparatus that gives us some optimal way of recognizing unusual phenomena and inventing new classes of hypotheses that are most likely to contain the true one; but this remains an art for the creative human mind. In trying to practice this art, the Bayesian has the advantage because his formal apparatus already developed gives him a clearer picture of what to expect; and therefore a sharper perception for recognizing the unexpected.

Segue from there to model building and estimation in next chapter.