

Statistical Rethinking

A BAYESIAN COURSE
WITH EXAMPLES
IN R AND STAN

Richard McElreath

This version compiled November 9, 2015

1 The Golem of Prague

In the 16th century, the House of Habsburg controlled much of Central Europe, the Netherlands, and Spain, as well as Spain's colonies in the Americas. The House was maybe the first true world power, with the Sun always shining on some portion of it. Its ruler was also Holy Roman Emperor, and his seat of power was Prague. The Emperor in the late 16th century, Rudolph II, loved intellectual life. He invested in the arts, the sciences (including astrology and alchemy), and mathematics, making Prague into a world center of learning and scholarship. It is appropriate then that in this learned atmosphere arose an early robot, the Golem of Prague.

A golem (GOH-lem) is a clay robot known in Jewish folklore, constructed from dust and fire and water. It is brought to life by inscribing *emet*, Hebrew for “truth,” on its brow. Animated by truth, but lacking free will, a golem always does exactly what it is told. This is lucky, because the golem is incredibly powerful, able to withstand and accomplish more than its creators could. However, its obedience also brings danger, as careless instructions or unexpected events can turn a golem against its makers. Its abundance of power is matched by its lack of wisdom.

In some versions of the golem legend, Rabbi Judah Loew ben Bezalel sought a way to defend the Jews of Prague. As in many parts of 16th century Central Europe, the Jews of Prague were persecuted. Using secret techniques from the *Kabbalah*, Rabbi Judah was able to build a golem, animate it with “truth,” and order it to defend the Jewish people of Prague. Not everyone agreed with Judah's action, fearing unintended consequences of toying with the power of life. Ultimately Judah was forced to destroy the golem, as its combination of extraordinary power with clumsiness eventually led to innocent deaths. Wiping away one letter from the inscription *emet* to spell instead *met*, “death,” Rabbi Judah decommissioned the robot.

1.1. Statistical golems

Scientists also make golems.¹ Our golems rarely have physical form, but they too are often made of clay, living in silicon as computer code. These golems are scientific models. But these golems have real effects on the world, through the predictions they make and the intuitions they challenge or inspire. A concern with “truth” enlivens these models, but just like a golem or a modern robot, scientific models are neither true nor false, neither prophets nor charlatans. Rather they are constructs engineered for some purpose. These constructs are incredibly powerful, dutifully conducting their programmed calculations.

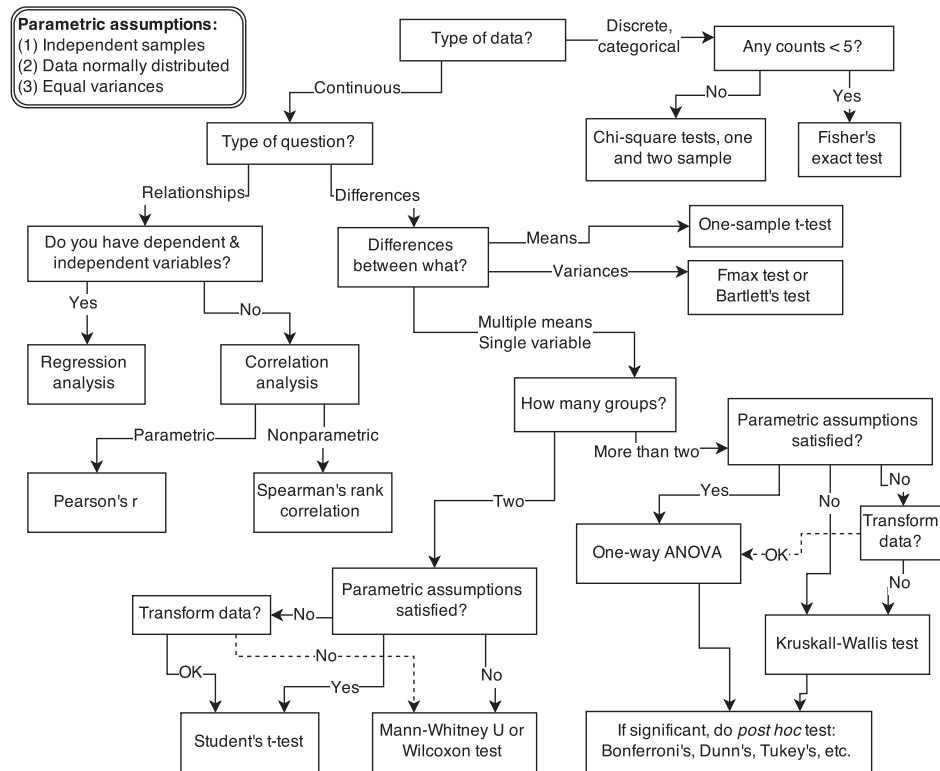


FIGURE 1.1. Example decision tree, or flowchart, for selecting an appropriate statistical procedure. Beginning at the top, the user answers a series of questions about measurement and intent, arriving eventually at the name of a procedure. Many such decision trees are possible.

Sometimes their unyielding logic reveals implications previously hidden to their designers. These implications can be priceless discoveries. Or they may produce silly and dangerous behavior. Rather than idealized angels of reason, scientific models are powerful clay robots without intent of their own, bumbling along according to the myopic instructions they embody. Like with Rabbi Judah's golem, the golems of science are wisely regarded with both awe and apprehension. We absolutely have to use them, but doing so always entails some risk.

There are many kinds of statistical models. Whenever someone deploys even a simple statistical procedure, like a classical t -test, she is deploying a small golem that will obediently carry out an exact calculation, performing it the same way (nearly²) every time, without complaint. Nearly every branch of science relies upon the senses of statistical golems. In many cases, it is no longer possible to even measure phenomena of interest, without making use of a model. To measure the strength of natural selection or the speed of a neutrino or the number of species in the Amazon, we must use models. The golem is a prosthesis, doing the measuring for us, performing impressive calculations, finding patterns where none are obvious.

However, there is no wisdom in the golem. It doesn't discern when the context is inappropriate for its answers. It just knows its own procedure, nothing else. It just does as it's told.

And so it remains a triumph of statistical science that there are now so many diverse golems, each useful in a particular context. Viewed this way, statistics is neither mathematics nor a science, but rather a branch of engineering. And like engineering, a common set of design principles and constraints produces a great diversity of specialized applications.

This diversity of applications helps to explain why introductory statistics courses are so often confusing to the initiates. Instead of a single method for building, refining, and critiquing statistical models, students are offered a zoo of pre-constructed golems known as “tests.” Each test has a particular purpose. Decision trees, like the one in [FIGURE 1.1](#), are common. By answering a series of sequential questions, users choose the “correct” procedure for their research circumstances.

Unfortunately, while experienced statisticians grasp the unity of these procedures, students and researchers rarely do. Advanced courses in statistics do emphasize engineering principles, but most scientists never get that far. Teaching statistics this way is somewhat like teaching engineering backwards, starting with bridge building and ending with basic physics. So students and many scientists tend to use charts like [FIGURE 1.1](#) without much thought to their underlying structure, without much awareness of the models that each procedure embodies, and without any framework to help them make the inevitable compromises required by real research. It’s not their fault.

For some, the toolbox of pre-manufactured golems is all they will ever need. Provided they stay within well-tested contexts, using only a few different procedures in appropriate tasks, a lot of good science can be completed. This is similar to how plumbers can do a lot of useful work without knowing much about fluid dynamics. Serious trouble begins when scholars move on to conducting innovative research, pushing the boundaries of their specialties. It’s as if we got our hydraulic engineers by promoting plumbers.

Why aren’t the tests enough for innovative research? The classical procedures of introductory statistics tend to be inflexible and fragile. By inflexible, I mean that they have very limited ways to adapt to unique research contexts. By fragile, I mean that they fail in unpredictable ways when applied to new contexts. This matters, because at the boundaries of most sciences, it is hardly ever clear which procedure is appropriate. None of the traditional golems has been evaluated in novel research settings, and so it can be hard to choose one and then to understand how it behaves. A good example is *Fisher’s exact test*, which applies (exactly) to an extremely narrow empirical context, but is regularly used whenever cell counts are small. I have personally read hundreds of uses of Fisher’s exact test in scientific journals, but aside from Fisher’s original use of it, I have never seen it used appropriately. Even a procedure like ordinary linear regression, which is quite flexible in many ways, being able to encode a large diversity of interesting hypotheses, is sometimes fragile. For example, if there is substantial measurement error on prediction variables, then the procedure can fail in spectacular ways. But more importantly, it is nearly always possible to do better than ordinary linear regression, largely because of a phenomenon known as **OVERFITTING** (Chapter 6).

The point isn’t that statistical tools are specialized. Of course they are. The point is that classical tools are not diverse enough to handle many common research questions. Every active area of science contends with unique difficulties of measurement and interpretation, converses with idiosyncratic theories in a dialect barely understood by other scientists from other tribes. Statistical experts outside the discipline can help, but they are limited by lack of fluency in the empirical and theoretical concerns of the discipline. In such settings, pre-manufactured golems may do nothing useful at all. Worse, they might wreck Prague. And if we keep adding new types of tools, soon there will be far too many to keep track of.

Instead, what researchers need is some unified theory of golem engineering, a set of principles for designing, building, and refining special-purpose statistical procedures. Every major branch of statistical philosophy possesses such a unified theory. But the theory is never taught in introductory—and often not even in advanced—courses. So there are benefits in rethinking statistical inference as a set of strategies, instead of a set of pre-made tools.

1.2. Statistical rethinking

A lot can go wrong with statistical inference, and this is one reason that beginners are so anxious about it. When the framework is to choose a pre-made test from a flowchart, then the anxiety can mount as one worries about choosing the “correct” test. Statisticians, for their part, can derive pleasure from scolding scientists, which just makes the psychological battle worse.

But anxiety can be cultivated into wisdom. That is the reason that this book insists on working with the computational nuts and bolts of each golem. If you don’t understand how the golem processes information, then you can’t interpret the golem’s output. This requires knowing the statistical model in greater detail than is customary, and it requires doing the computations the hard way, at least until you are wise enough to use the push-button solutions.

There are conceptual obstacles as well, obstacles with how scholars define statistical objectives and interpret statistical results. Understanding any individual golem is not enough, in these cases. Instead, we need some statistical epistemology, an appreciation of how statistical models relate to hypotheses and the natural mechanisms of interest. What are we supposed to be doing with these little computational machines, anyway?

The greatest obstacle that I encounter among students and colleagues is the tacit belief that the proper objective of statistical inference is to test null hypotheses.³ This is the proper objective, the thinking goes, because Karl Popper argued that science advances by falsifying hypotheses. Karl Popper (1902–1994) is possibly the most influential philosopher of science, at least among scientists. He did persuasively argue that science works better by developing hypotheses that are, in principle, falsifiable. Seeking out evidence that might embarrass our ideas is a normative standard, and one that most scholars—whether they describe themselves as scientists or not—subscribe to. So maybe statistical procedures should falsify hypotheses, if we wish to be good statistical scientists.

But the above is a kind of folk Popperism, an informal philosophy of science common among scientists but not among philosophers of science. Science is not described by the falsification standard, as Popper recognized and argued.⁴ In fact, deductive falsification is impossible in nearly every scientific context. In this section, I review two reasons for this impossibility.

- (1) Hypotheses are not models. The relations among hypotheses and different kinds of models are complex. Many models correspond to the same hypothesis, and many hypotheses correspond to a single model. This makes strict falsification impossible.
- (2) Measurement matters. Even when we think the data falsify a model, another observer will debate our methods and measures. They don’t trust the data. Sometimes they are right.

For both of these reasons, deductive falsification never works. The scientific method cannot be reduced to a statistical procedure, and so our statistical methods should not pretend. Statistical evidence is part of the hot mess that is science, with all of its combat and egotism and

mutual coercion. If you believe, as I do, that science does very often work, then learning that it doesn't work via falsification shouldn't change your mind. But it might help you do better science, because it will open your eyes to the many legitimately useful functions of statistical golems.

Rethinking: Is NHST falsificationist? Null hypothesis significance testing, NHST, is often identified with the falsificationist, or Popperian, philosophy of science. However, usually NHST is used to falsify a null hypothesis, not the actual research hypothesis. So the falsification is being done to something other than the explanatory model. This seems the reverse from Karl Popper's philosophy.⁵

1.2.1. Hypotheses are not models. When we attempt to falsify a hypothesis, we must work with a model of some kind. Even when the attempt is not explicitly statistical, there is always a tacit model of measurement, of evidence, that operationalizes the hypothesis. All models are false,⁶ so what does it mean to falsify a model? One consequence of the requirement to work with models is that it's no longer possible to deduce that a hypothesis is false, just because we reject a model derived from it.

Let's explore this consequence in the context of an example from population biology (FIGURE 1.2). Beginning in the 1960s, many evolutionary biologists became interested in the proposal that the vast majority of evolution—changes in gene frequency—are caused not by natural selection, but rather by mutation and drift. No one really doubted that natural selection is responsible for functional design. This was a debate about genetic sequences. So began several productive decades of scholarly combat over “neutral” models of molecular evolution.⁷ This combat is most strongly associated with Motoo Kimura (1924–1994), who was perhaps the strongest advocate of neutral models. But many other population geneticists participated. As time has passed, related disciplines such as community ecology⁸ and anthropology⁹ have experienced (or are currently experiencing) their own versions of the neutrality debate.

Let's use the schematic in FIGURE 1.2 to explore connections between motivating hypotheses and different models, in the context of the neutral evolution debate. On the left, there are two stereotyped, informal hypotheses: Either evolution is “neutral” (H_0) or natural selection matters somehow (H_1). These hypotheses have vague boundaries, because they begin as verbal conjectures, not precise models. There are thousands of possible detailed processes that can be described as “neutral,” depending upon choices about, for example, population structure, number of sites, number of alleles at each site, mutation rates, and recombination.

Once we have made these choices, we have the middle column in FIGURE 1.2, detailed process models of evolution. P_{0A} and P_{0B} differ in that one assumes the population size and structure have been constant long enough for the distribution of alleles to reach a steady state. The other imagines instead that population size fluctuates through time, which can be true even when there is no selective difference among alleles. The “selection matters” hypothesis H_1 likewise corresponds to many different process models. I've shown two big players: a model in which selection always favors certain alleles and another in which selection fluctuates through time, favoring different alleles.¹⁰

In order to challenge these process models with evidence, they have to be made into statistical models. This usually means deriving the expected frequency distribution of some quantity—a “statistic”—in the model. For example, a very common statistic in this context is the frequency distribution (histogram) of the frequency of different genetic variants (alleles).

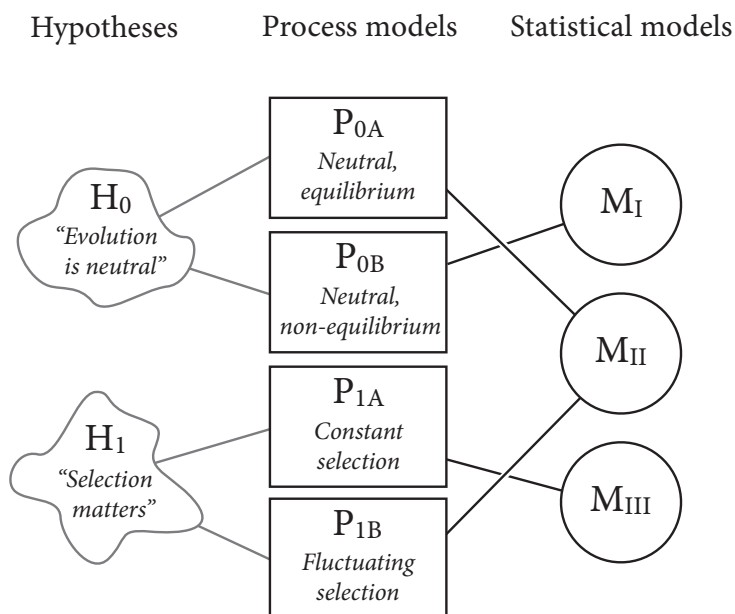


FIGURE 1.2. Relations among hypotheses (left), detailed process models (middle), and statistical models (right), illustrated by the example of “neutral” models of evolution. Hypotheses (H) are typically vague, and so correspond to more than one process model (P). Statistical evaluations of hypotheses rarely address process models directly. Instead, they rely upon statistical models (M), all of which reflect only some aspects of the process models. As a result, relations are multiple in both directions: Hypotheses do not imply unique models, and models do not imply unique hypotheses. This fact greatly complicates statistical inference.

Some alleles are rare, appearing in only a few individuals. Others are very common, appearing in very many individuals in the population. A famous result in population genetics is that a model like P_{0A} produces a *power law* distribution of allele frequencies. And so this fact yields a statistical model, M_{II} , that predicts a power law in the data. In contrast the constant selection process model P_{1A} predicts something quite different, M_{III} .

Unfortunately, other selection models (P_{1B}) imply the same statistical model, M_{II} , as the neutral model. They also produce power laws. So we’ve reached the uncomfortable lesson:

- (1) Any given statistical model (M) may correspond to more than one process model (P).
- (2) Any given hypothesis (H) may correspond to more than one process model (P).
- (3) Any given statistical model (M) may correspond to more than one hypothesis (H).

Now look what happens when we compare the statistical models to data. The classical approach is to take the “neutral” model as a null hypothesis. If the data are not sufficiently similar to the expectation under the null, then we say that we “reject” the null hypothesis. Suppose we follow the history of this subject and take P_{0A} as our null hypothesis. This implies data corresponding to M_{II} . But since the same statistical model corresponds to a selection model P_{1B} , it’s not at all clear what we are to make of either rejecting or accepting the

null. The null model is not unique to any process model nor hypothesis. If we reject the null, we can't really conclude that selection matters, because there are other neutral models that predict different distributions of alleles. And if we fail to reject the null, we can't really conclude that evolution is neutral, because some selection models expect the same frequency distribution.

This is a huge bother. Once we have the diagram in [FIGURE 1.2](#), it's easy to see the problem. But few of us are so lucky. While population genetics has recognized this issue, scholars in other disciplines continue to test frequency distributions against power law expectations, arguing even that there is only one neutral model.¹¹ Even if there were only one neutral model, there are so many non-neutral models that mimic the predictions of neutrality, that neither rejecting nor failing to reject the null model carries much inferential power.

And while you might think that more routine statistical models, like linear regressions (Chapter 4), don't carry such risk, think again. A typical "null" in these contexts is just that there is zero *average* difference between groups. But there are usually many different ways for this average to be close to or consistent with zero, just as there are many different ways to get a power law. This recognition lies behind many common practices in statistical inference, such as consideration of unobserved variables and sampling bias.

So what can be done? Well, if you have multiple process models, a lot can be done. If it turns out that all of the process models of interest make very similar predictions, then you know to search for a different description of the evidence, a description under which the processes look different. For example, while P_{0A} and P_{1B} make very similar power law predictions for the frequency distribution of alleles, they make very dissimilar predictions for the distribution of changes in allele frequency over time. In other words, explicitly compare predictions of more than one model, and you can save yourself from some ordinary kinds of folly.

Rethinking: Entropy and model identification. One reason that statistical models routinely correspond to many different detailed process models is because they rely upon distributions like the normal, binomial, Poisson, and others. These distributions are members of a family, the [EXPONENTIAL FAMILY](#). Nature loves the members of this family. Nature loves them because nature loves entropy, and all of the exponential family distributions are [MAXIMUM ENTROPY](#) distributions. Taking the natural personification out of that explanation will wait until Chapter 9. The practical implication is that one can no more infer evolutionary process from a power law than one can infer developmental process from the fact that height is normally distributed. This fact should make us humble about what typical regression models—the meat of this book—can teach us about mechanistic process. On the other hand, the maximum entropy nature of these distributions means we can use them to do useful statistical work, even when we can't identify the underlying process. Not only can we not identify it, but we don't have to.

1.2.2. Measurement matters. The logic of falsification is very simple. We have a hypothesis H , and we show that it entails some observation D . Then we look for D . If we don't find it, we must conclude that H is false. Logicians call this kind of reasoning *modus tollens*, which is Latin shorthand for "the method of destruction." In contrast, finding D tells us nothing certain about H , because other hypotheses might also predict D .

A compelling scientific fable that employs *modus tollens* concerns the color of swans. Before discovering Australia, all swans that any European had ever seen had white feathers. This led to the belief that all swans are white. Let's call this a formal hypothesis:

H_0 : All swans are white.

When Europeans reached Australia, however, they encountered swans with black feathers. This evidence seemed to instantly prove H_0 to be false. Indeed, not all swans are white. Some are certainly black, according to all observers. The key insight here is that, before voyaging to Australia, no number of observations of white swans could prove H_0 to be true. However it required only one observation of a black swan to prove it false.

This is a seductive story. If we can believe that important scientific hypotheses can be stated in this form, then we have a powerful method for improving the accuracy of our theories: look for evidence that disconfirms our hypotheses. Whenever we find a black swan, H_0 must be false. Progress!

Seeking disconfirming evidence is important, but it cannot be as powerful as the swan story makes it appear. In addition to the correspondence problems among hypotheses and models, discussed in the previous section, most of the problems scientists confront are not so logically discrete. Instead, we most often face two simultaneous problems that make the swan fable misrepresentative. First, observations are prone to error, especially at the boundaries of scientific knowledge. Second, most hypotheses are quantitative, concerning degrees of existence, rather than discrete, concerning total presence or absence. Let's briefly consider each of these problems.

1.2.2.1. *Observation error.* All observers will agree under most conditions that a swan is either black or white. There are few intermediate shades, and most observers' eyes work similarly enough that there will be little, if any, disagreement about which swans are white and which are black. But this kind of example is hardly commonplace in science, at least in mature fields. Instead, we routinely confront contexts in which we are not sure if we have detected a disconfirming result. At the edges of scientific knowledge, the ability to measure a hypothetical phenomenon is often in question as much as the phenomenon itself.

Here are two examples.

In 2005, a team of ornithologists from Cornell claimed to have evidence of an individual Ivory-billed Woodpecker (*Campephilus principalis*), a species thought extinct. The hypothesis implied here is:

H_0 : The Ivory-billed Woodpecker is extinct.

It would only take one observation to falsify this hypothesis. However, many doubted the evidence. Despite extensive search efforts and a \$50,000 cash reward for information leading to a live specimen, no evidence satisfying all parties has yet (by 2015) emerged. Even if good physical evidence does eventually arise, this episode should serve as a counterpoint to the swan story. Finding disconfirming cases is complicated by the difficulties of observation. Black swans are not always really black swans, and sometimes white swans are really black swans. There are mistaken confirmations (false positives) and mistaken disconfirmations (false negatives). Against this background of measurement difficulties, scientists who already believe that the Ivory-billed Woodpecker is extinct will always be suspicious of a claimed falsification. Those who believe it is still alive will tend to count the vaguest evidence as falsification.

Another example, this one from physics, focuses on the detection of faster-than-light (FTL) neutrinos.¹² In September 2011, a large and respected team of physicists announced detection of neutrinos—small, neutral sub-atomic particles able to pass easily and harmlessly through most matter—that arrived from Switzerland to Italy in slightly faster-than-light-speed time. According to Einstein, neutrinos cannot travel faster than the speed of light. So this seems to be a falsification of special relativity. If so, it would turn physics on its head.

The dominant reaction from the physics community was not “Einstein was wrong!” but instead “How did the team mess up the measurement?” The team that made the measurement had the same reaction, and asked others to check their calculations and attempt to replicate the result.

What could go wrong in the measurement? You might think measuring speed is a simple matter of dividing distance by time. It is, at the scale and energy you live at. But with a fundamental particle like a neutrino, if you measure when it starts its journey, you stop the journey. The particle is consumed by the measurement. So more subtle approaches are needed. The detected difference from light-speed, furthermore, is quite small, and so even the latency of the time it takes a signal to travel from a detector to a control room can be orders of magnitude larger. And since the “measurement” in this case is really an estimate from a statistical model, all of the assumptions of the model are now suspect. By 2013, the physics community was unanimous that the FTL neutrino result was measurement error. They found the technical error, which involved a poorly attached cable, among other things.¹³ Furthermore, neutrinos clocked from supernova events are consistent with Einstein, and those distances are much larger and so would reveal differences in speed much better.

In both the woodpecker and neutrino dramas, the key dilemma is whether the falsification is real or spurious. Measurement is complicated in both cases, but in quite different ways, rendering both true-detection and false-detection plausible. Popper himself was aware of this limitation inherent in measurement, and it may be one reason that Popper himself saw science as being broader than falsification. But the probabilistic nature of evidence rarely appears when practicing scientists discuss the philosophy and practice of falsification.¹⁴ My reading of the history of science is that these sorts of measurement problems are the norm, not the exception.¹⁵

1.2.2.2. *Continuous hypotheses.* Another problem for the swan story is that most interesting scientific hypotheses are not of the kind “all swans are white” but rather of the kind:

H_0 : 80% of swans are white.

Or maybe:

H_0 : Black swans are rare.

Now what are we to conclude, after observing a black swan? The null hypothesis doesn't say black swans do not exist, but rather that they have some frequency. The task here is not to disprove or prove a hypothesis of this kind, but rather to estimate and explain the distribution of swan coloration as accurately as we can. Even when there is no measurement error of any kind, this problem will prevent us from applying the *modus tollens* swan story to our science.¹⁶

You might object that the hypothesis above is just not a good scientific hypothesis, because it isn't easy to disprove. But if that's the case, then most of the important questions about the world are not good scientific hypotheses. In that case, we should conclude that the definition of a “good hypothesis” isn't doing us much good. Now, nearly everyone agrees that it is a good practice to design experiments and observations that can differentiate competing hypotheses. But in many cases, the comparison must be probabilistic, a matter of degree, not kind.¹⁷

1.2.3. **Falsification is consensual.** The scientific community does come to regard some hypotheses as false. The caloric theory of heat and the geocentric model of the universe are no

longer taught in science courses, unless it's to teach how they were falsified. And evidence often—but not always—has something to do with such falsification.

But falsification is always *consensual*, not *logical*. In light of the real problems of measurement error and the continuous nature of natural phenomena, scientific communities argue towards consensus about the meaning of evidence. These arguments can be messy. After the fact, some textbooks misrepresent the history so it appears like logical falsification.¹⁸ Such historical revisionism may hurt everyone. It may hurt scientists, by rendering it impossible for their own work to live up to the legends that precede them. It may make science an easy target, by promoting an easily attacked model of scientific epistemology. And it may hurt the public, by exaggerating the definitiveness of scientific knowledge.¹⁹

1.3. Three tools for golem engineering

So if attempting to mimic falsification is not a generally useful approach to statistical methods, what are we to do? We are to model. Models can be made into testing procedures—all statistical tests are also models²⁰—but they can also be used to measure, forecast, and argue. Doing research benefits from the ability to produce and manipulate statistical models, both because scientific problems are more general than “testing” and because the pre-made golems you maybe met in introductory statistics courses are ill-fit to many research contexts. If you want to reduce your chances of wrecking Prague, then some golem engineering know-how is needed. Make no mistake: You will wreck Prague eventually. But if you are a good golem engineer, at least you'll notice the destruction. And since you'll know a lot about how your golem works, you stand a good chance to figure out what went wrong. Then your next golem won't be as bad. Without the engineering training, you're always at someone else's mercy.

It can be hard to get a good education in statistical model building and criticism, though. Applied statistical modeling in the early 21st century is marked by the heavy use of several engineering tools that are almost always absent from introductory, and even many advanced, statistics courses. These tools aren't really new, but they are newly popular. And many of the recent advances in statistical inference depend upon computational innovations that feel more like computer science than classical statistics, so it's not clear who is responsible for teaching them, if anyone.

There are many tools worth learning. In this book I've chosen to focus on three broad ones that are in demand in both the social and biological sciences. These tools are:

- (1) Bayesian data analysis
- (2) Multilevel models
- (3) Model comparison using information criteria

These tools are deeply related to one another, so it makes sense to teach them together. Understanding of these tools comes, as always, only with implementation—you can't comprehend golem engineering until you do it. And so this book focuses mostly on code, how to do things. But in the rest of this section, I provide brief introductions to the three tools.

1.3.1. Bayesian data analysis. For the classical Greeks and Romans, wisdom and chance were enemies. Minerva (Athena), symbolized by the owl, was the personification of wisdom. Fortuna (Tyche), symbolized by the wheel of fortune, was the personification of luck, both good and bad. Minerva was mindful and measuring, while Fortuna was fickle and unreliable. Only a fool would rely on Fortuna, while all wise folk appealed to Minerva.²¹