

Statistical Rethinking²

A BAYESIAN COURSE
WITH EXAMPLES
IN R AND STAN

Second Edition

Richard McElreath

This version compiled February 22, 2020


1 The Golem of Prague

In the sixteenth 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. The Sun shone always on some portion of it. Its ruler was also Holy Roman Emperor, and his seat of power was Prague. The Emperor in the late sixteenth 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 from 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 sixteenth 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.  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.

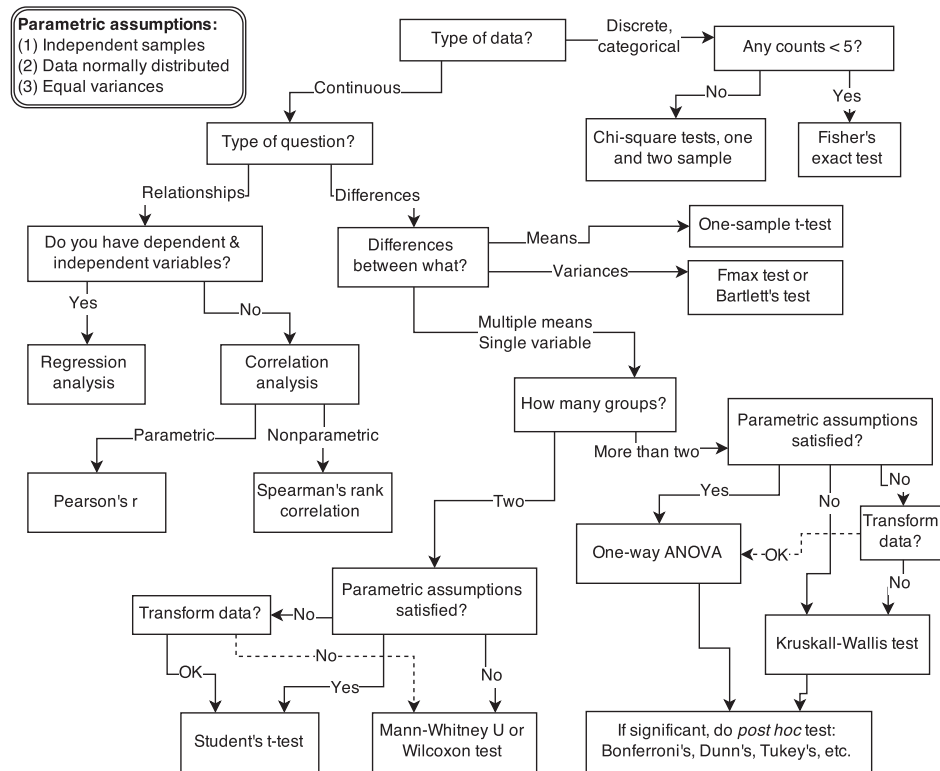


FIGURE 1.1. Example decision tree, or flowchart, for selecting an appropriate statistical procedure. Beginning at the top, the user answers a series of questions about measurement and intent, arriving eventually at the name of a procedure. Many such decision trees are possible.

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 (nearly ∞) 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 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 [7](#)).

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.

Furthermore, no statistical tool does anything on its own to address the basic problem of inferring causes from evidence. Statistical golems do not understand cause and effect.

They only understand association. Without our guidance and skepticism, pre-manufactured golems may do nothing useful at all. Worse, they might wreck Prague.

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 goal 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, making 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 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.¹ 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, and Popper recognized that.² 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 often work, then learning that it

doesn't work via falsification shouldn't change your mind. But it might help you do better science. It might open your eyes to 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.⁸

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,⁹ 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, evolutionary biologists became interested in the proposal that the majority of evolutionary 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.¹⁰ 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 ecology¹¹ and anthropology¹² 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 hundreds of possible detailed processes that can be described as "neutral," depending upon choices about 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.¹³

An important feature of these process models is that they express causal structure. Different process models formalize different cause and effect relationships. Whether analyzed mathematically or through simulation, the direction of time in a model means that some things cause other things, but not the reverse. You can use such models to perform experiments and probe their causal implications. Sometimes these probes reveal, before we even turn to statistical inference, that the model cannot explain a phenomenon of interest.

In order to challenge process models with data, they have to be made into statistical models. Unfortunately, statistical models do not embody specific causal relationships. A

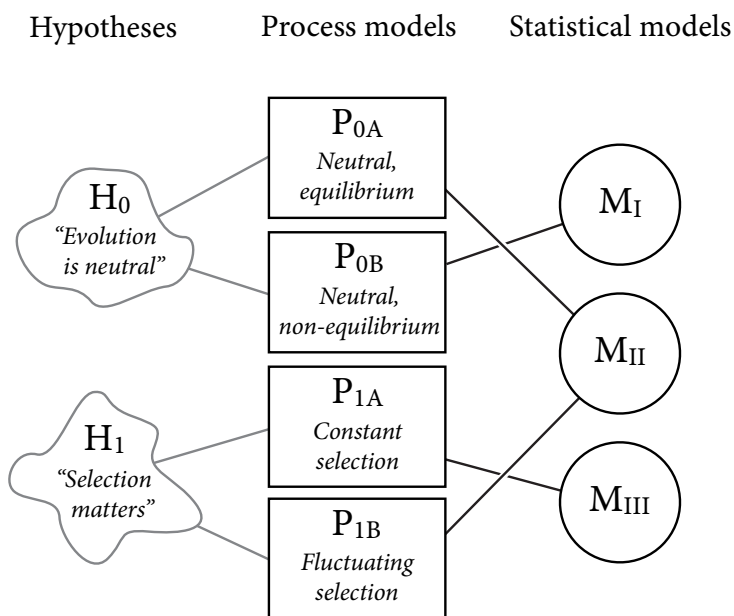


FIGURE 1.2. Relations among hypotheses (left), detailed process models (middle), and statistical models (right), illustrated by the example of “neutral” models of evolution. Hypotheses (H) are typically vague, and so correspond to more than one process model (P). Statistical evaluations of hypotheses 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.

statistical model expresses associations among variables. As a result, many different process models may be consistent with any single statistical model.

How do we get a statistical model from a causal model? One way is to derive the expected frequency distribution of some quantity—a “statistic”—from the causal model. For example, a common statistic in this context is the frequency distribution (histogram) of the frequency of different genetic variants (alleles). 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 clear what 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. [□](#) 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.

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. Explicitly compare predictions of more than one model, and you can save yourself from some ordinary kinds of folly.

Statistical models can be confused in other ways as well, such as the confusion caused by unobserved variables and sampling bias. Process models allow us to design statistical models with these problems in mind. The statistical model alone is not enough.

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 10](#). 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.

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 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 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 satisfying evidence has yet (by 2020) 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.[□] 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.^[3] 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 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.^[4] My reading of the history of science is that these sorts of measurement problems are the norm, not the exception.^[5]

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.^[6]

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.^[7]

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. 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 models^[20]—but they can also be used to design, forecast, and argue. Doing research benefits from the ability to produce and manipulate 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. You may not even know which statistical model to use, unless you have a generative model in addition.

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 engineering training, you're always at someone's mercy.

We want to use our models for several distinct purposes: designing inquiry, extracting information from data, and making predictions. In this book I've chosen to focus on tools to help with each purpose. These tools are:

- (1) Bayesian data analysis
- (2) Model comparison
- (3) Multilevel models
- (4) Graphical causal models

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 remainder of this chapter, I provide introductions to these tools.

1.3.1. Bayesian data analysis. Supposing you have some data, how should you use it to learn about the world? There is no uniquely correct answer to this question. Lots of approaches, both formal and heuristic, can be effective. But one of the most effective and general answers is to use Bayesian data analysis. Bayesian data analysis takes a question in the form of a model and uses logic to produce an answer in the form of probability distributions.

In modest terms, Bayesian data analysis is no more than counting the numbers of ways the data could happen, according to our assumptions. Things that can happen more ways are more plausible. Probability theory is relevant because probability is just a calculus for counting. This allows us to use probability theory as a general way to represent plausibility, whether in reference to countable events in the world or rather theoretical constructs like

parameters. The rest follows logically. Once we have defined the statistical model, Bayesian data analysis forces a purely logical way of processing the data to produce inference.

Chapter 2 explains this in depth. For now, it will help to have another approach to compare. Bayesian probability is a very general approach to probability, and it includes as a special case another important approach, the **FREQUENTIST** approach. The frequentist approach requires that all probabilities be defined by connection to the frequencies of events in very large samples.^[21] This leads to frequentist uncertainty being premised on imaginary resampling of data—if we were to repeat the measurement many many times, we would end up collecting a list of values that will have some pattern to it. It means also that parameters and models cannot have probability distributions, only measurements can. The distribution of these measurements is called a **SAMPLING DISTRIBUTION**. This resampling is never done, and in general it doesn't even make sense—it is absurd to consider repeat sampling of the diversification of song birds in the Andes. As Sir Ronald Fisher, one of the most important frequentist statisticians of the twentieth century, put it:^[22]

[...] the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination [...]

But in many contexts, like controlled greenhouse experiments, it's a useful device for describing uncertainty. Whatever the context, it's just part of the model, an assumption about what the data would look like under resampling. It's just as fantastical as the Bayesian gambit of using probability to describe all types of uncertainty, whether empirical or epistemological.^[23]

But these different attitudes towards probability do enforce different trade-offs. Consider this simple example where the difference between Bayesian and frequentist probability matters. In the year 1610, Galileo turned a primitive telescope to the night sky and became the first human to see Saturn's rings. Well, he probably saw a blob, with some smaller blobs attached to it (FIGURE L.3). Since the telescope was primitive, it couldn't really focus the image very well. Saturn always appeared blurred. This is a statistical problem, of a sort. There's uncertainty about the planet's shape, but notice that none of the uncertainty is a result of variation in repeat measurements. We could look through the telescope a thousand times, and it will always give the same blurred image (for any given position of the Earth and Saturn). So the sampling distribution of any measurement is constant, because the measurement is deterministic—there's nothing “random” about it. Frequentist statistical inference has a lot of trouble getting started here. In contrast, Bayesian inference proceeds as usual, because the deterministic “noise” can still be modeled using probability, as long as we don't identify probability with frequency. As a result, the field of image reconstruction and processing is dominated by Bayesian algorithms.^[24]

In more routine statistical procedures, like linear regression, this difference in probability concepts has less of an effect. However, it is important to realize that even when a Bayesian procedure and frequentist procedure give exactly the same answer, our Bayesian golems aren't justifying their inferences with imagined repeat sampling. More generally, Bayesian golems treat “randomness” as a property of information, not of the world. Nothing in the real world—excepting controversial interpretations of quantum physics—is actually random. Presumably, if we had more information, we could exactly predict everything. We just use randomness to describe our uncertainty in the face of incomplete knowledge. From the perspective of our golem, the coin toss is “random,” but it's really the golem that is random, not the coin.



FIGURE 1.3. Saturn, much like Galileo must have seen it. The true shape is uncertain, but not because of any sampling variation. Probability theory can still help.

Note that the preceding description doesn't invoke anyone's "beliefs" or subjective opinions. Bayesian data analysis is just a logical procedure for processing information. There is a tradition of using this procedure as a normative description of rational belief, a tradition called **BAYESIANISM**.²⁵ But this book neither describes nor advocates it. In fact, I'll argue that no statistical approach, Bayesian or otherwise, is by itself sufficient.

Before moving on to describe the next two tools, it's worth emphasizing an advantage of Bayesian data analysis, at least when scholars are learning statistical modeling. This entire book could be rewritten to remove any mention of "Bayesian." In places, it would become easier. In others, it would become much harder. But having taught applied statistics both ways, I have found that the Bayesian framework presents a distinct pedagogical advantage: many people find it more intuitive. Perhaps the best evidence for this is that very many scientists interpret non-Bayesian results in Bayesian terms, for example interpreting ordinary p -values as Bayesian posterior probabilities and non-Bayesian confidence intervals as Bayesian ones (you'll learn posterior probability and confidence intervals in Chapters 2 and 3). Even statistics instructors make these mistakes.²⁶ Statisticians appear doomed to republish the same warnings about misinterpretation of p -values forever. In this sense then, Bayesian models lead to more intuitive interpretations, the ones scientists tend to project onto statistical results. The opposite pattern of mistake—interpreting a posterior probability as a p -value—seems to happen only rarely.

None of this ensures that Bayesian analyses will be more correct than non-Bayesian analyses. It just means that the scientist's intuitions will less commonly be at odds with the actual logic of the framework. This simplifies some of the aspects of teaching statistical modeling.

Rethinking: Probability is not unitary. It will make some readers uncomfortable to suggest that there is more than one way to define "probability." Aren't mathematical concepts uniquely correct? They are not. Once you adopt some set of premises, or axioms, everything does follow logically in mathematical systems. But the axioms are open to debate and interpretation. So not only is there "Bayesian" and "frequentist" probability, but there are different versions of Bayesian probability even,

relying upon different arguments to justify the approach. In more advanced Bayesian texts, you'll come across names like Bruno de Finetti, Richard T. Cox, and Leonard "Jimmie" Savage. Each of these figures is associated with a somewhat different conception of Bayesian probability. There are others. This book mainly follows the "logical" Cox (or Laplace-Jeffreys-Cox-Jaynes) interpretation. This interpretation is presented beginning in the next chapter, but unfolds fully only in Chapter 10.

How can different interpretations of probability theory thrive? By themselves, mathematical entities don't necessarily "mean" anything, in the sense of real world implication. What does it mean to take the square root of a negative number? What does it mean to take a limit as something approaches infinity? These are essential and routine concepts, but their meanings depend upon context and analyst, upon beliefs about how well abstraction represents reality. Mathematics doesn't access the real world directly. So answering such questions remains a contentious and entertaining project, in all branches of applied mathematics. So while everyone subscribes to the same axioms of probability, not everyone agrees in all contexts about how to interpret probability.

Rethinking: A little history. Bayesian statistical inference is much older than the typical tools of introductory statistics, most of which were developed in the early twentieth century. Versions of the Bayesian approach were applied to scientific work in the late 1700s and repeatedly in the nineteenth century. But after World War I, anti-Bayesian statisticians, like Sir Ronald Fisher, succeeded in marginalizing the approach. All Fisher said about Bayesian analysis (then called *inverse probability*) in his influential 1925 handbook was:²⁷

[...] the theory of inverse probability is founded upon an error, and must be wholly rejected.

Bayesian data analysis became increasingly accepted within statistics during the second half of the twentieth century, because it proved not to be founded upon an error. All philosophy aside, it worked. Beginning in the 1990s, new computational approaches led to a rapid rise in application of Bayesian methods.²⁸ Bayesian methods remain computationally expensive, however. And so as data sets have increased in scale—millions of rows is common in genomic analysis, for example—alternatives to or approximations to Bayesian inference remain important, and probably always will.

1.3.2. Model comparison and prediction. Bayesian data analysis provides a way for models to learn from data. But when there is more than one plausible model—and in most mature fields there should be—how should we choose among them? One answer is to prefer models that make good predictions. This answer creates a lot of new questions, since knowing which model will make the best predictions seems to require knowing the future. We'll look at two related tools, neither of which knows the future: **CROSS-VALIDATION** and **INFORMATION CRITERIA**. These tools aim to compare models based upon expected predictive accuracy.

Comparing models by predictive accuracy can be useful in itself. And it will be even more useful because it leads to the discovery of an amazing fact: Complex models often make worse predictions than simpler models. The primary paradox of prediction is **OVERFITTING**.²⁹ Future data will not be exactly like past data, and so any model that is unaware of this fact tends to make worse predictions than it could. And more complex models tend towards more overfitting than simple ones—the smarter the golem, the dumber its predictions. So if we wish to make good predictions, we cannot judge our models simply on how well they fit our data. *Fitting is easy; prediction is hard.*

Cross-validation and information criteria help us in three ways. First, they provide useful expectations of predictive accuracy, rather than merely fit to sample. So they compare models where it matters. Second, they give us an estimate of the tendency of a model to

overfit. This will help us to understand how models and data interact, which in turn helps us to design better models. We'll take this point up again in the next section. Third, cross-validation and information criteria help us to spot highly influential observations.

Bayesian data analysis has been worked on for centuries. Information criteria are comparatively very young and the field is evolving quickly. Many statisticians have never used information criteria in an applied problem, and there is no consensus about which metrics are best and how best to use them. Still, information criteria are already in frequent use in the sciences, appearing in prominent publications and featuring in prominent debates.^[80] Their power is often exaggerated, and we will be careful to note what they cannot do as well as what they can.

Rethinking: The Neanderthal in you. Even simple models need alternatives. In 2010, a draft genome of a Neanderthal demonstrated more DNA sequences in common with non-African contemporary humans than with African ones. This finding is consistent with interbreeding between Neanderthals and modern humans, as the latter dispersed from Africa. However, just finding DNA in common between modern Europeans and Neanderthals is not enough to demonstrate interbreeding. It is also consistent with ancient structure in the African continent.^[81] In short, if ancient northeast Africans had unique DNA sequences, then both Neanderthals and modern Europeans could possess these sequences from a common ancestor, rather than from direct interbreeding. So even in the seemingly simple case of estimating whether Neanderthals and modern humans share unique DNA, there is more than one process-based explanation. Model comparison is necessary.

1.3.3. Multilevel models. In an apocryphal telling of Hindu cosmology, it is said that the Earth rests on the back of a great elephant, who in turn stands on the back of a massive turtle. When asked upon what the turtle stands, a guru is said to reply, “it’s turtles all the way down.”

Statistical models don’t contain turtles, but they do contain parameters. And parameters support inference. Upon what do parameters themselves stand? Sometimes, in some of the most powerful models, it’s parameters all the way down. What this means is that any particular parameter can be usefully regarded as a placeholder for a missing model. Given some model of how the parameter gets its value, it is simple enough to embed the new model inside the old one. This results in a model with multiple levels of uncertainty, each feeding into the next—a **MULTILEVEL MODEL**.

Multilevel models—also known as hierarchical, random effects, varying effects, or mixed effects models—are becoming *de rigueur* in the biological and social sciences. Fields as diverse as educational testing and bacterial phylogenetics now depend upon routine multilevel models to process data. Like Bayesian data analysis, multilevel modeling is not particularly new. But it has only been available on desktop computers for a few decades. And since such models have a natural Bayesian representation, they have grown hand-in-hand with Bayesian data analysis.

One reason to be interested in multilevel models is because they help us deal with overfitting. Cross-validation and information criteria measure overfitting risk and help us to recognize it. Multilevel models actually do something about it. What they do is exploit an amazing trick known as **PARTIAL POOLING** that pools information across units in the data in order to produce better estimates for all units. The details will wait until Chapter **13**.

Partial pooling is the key technology, and the contexts in which it is appropriate are diverse. Here are four commonplace examples.

- (1) *To adjust estimates for repeat sampling.* When more than one observation arises from the same individual, location, or time, then traditional, single-level models may mislead us.
- (2) *To adjust estimates for imbalance in sampling.* When some individuals, locations, or times are sampled more than others, we may also be misled by single-level models.
- (3) *To study variation.* If our research questions include variation among individuals or other groups within the data, then multilevel models are a big help, because they model variation explicitly.
- (4) *To avoid averaging.* Pre-averaging data to construct variables can be dangerous. Averaging removes variation, manufacturing false confidence. Multilevel models preserve the uncertainty in the original, pre-averaged values, while still using the average to make predictions.

All four apply to contexts in which the researcher recognizes clusters or groups of measurements that may differ from one another. These clusters or groups may be individuals such as different students, locations such as different cities, or times such as different years. Since each cluster may well have a different average tendency or respond differently to any treatment, clustered data often benefit from being modeled by a golem that expects such variation.

But the scope of multilevel modeling is much greater than these examples. Diverse model types turn out to be multilevel: models for missing data (imputation), measurement error, factor analysis, some time series models, types of spatial and network regression, and phylogenetic regressions all are special applications of the multilevel strategy. And some commonplace procedures, like the paired *t*-test, are really multilevel models in disguise. Grasping the concept of multilevel modeling may lead to a perspective shift. Suddenly single-level models end up looking like mere components of multilevel models. The multilevel strategy provides an engineering principle to help us to introduce these components into a particular analysis, exactly where we think we need them.

I want to convince the reader of something that appears unreasonable: *multilevel regression deserves to be the default form of regression.* Papers that do not use multilevel models should have to justify not using a multilevel approach. Certainly some data and contexts do not need the multilevel treatment. But most contemporary studies in the social and natural sciences, whether experimental or not, would benefit from it. Perhaps the most important reason is that even well-controlled treatments interact with unmeasured aspects of the individuals, groups, or populations studied. This leads to variation in treatment effects, in which individuals or groups vary in how they respond to the same circumstance. Multilevel models attempt to quantify the extent of this variation, as well as identify which units in the data responded in which ways.

These benefits don't come for free, however. Fitting and interpreting multilevel models can be considerably harder than fitting and interpreting a traditional regression model. In practice, many researchers simply trust their black-box software and interpret multilevel regression exactly like single-level regression. In time, this will change. There was a time in applied statistics when even ordinary multiple regression was considered cutting edge, something for only experts to fiddle with. Instead, scientists used many simple procedures, like *t*-tests. Now, almost everyone uses multivariate tools. The same will eventually be true of multilevel models. Scholarly culture and curriculum still have some catching up to do.

Rethinking: Multilevel election forecasting. One of the older applications of multilevel modeling is to forecast the outcomes of elections. In the 1960s, John Tukey (1915–2000) began working for the National Broadcasting Company (NBC) in the United States, developing real-time election prediction models that could exploit diverse types of data: polls, past elections, partial results, and complete results from related districts. The models used a multilevel framework similar to the models presented in Chapters 13 and 14. Tukey developed and used such models for NBC through 1978.³² Contemporary election prediction and poll aggregation remains an active topic for multilevel modeling.³³

1.3.4. Graphical causal models. When the wind blows, branches sway. If you are human, you immediately interpret this statement as causal: The wind makes the branches move. But all we see is a statistical association. From the data alone, it could also be that the branches swaying makes the wind. That conclusion seems foolish, because you know trees do not sway their own branches. A statistical model is an amazing association engine. It makes it possible to detect associations between causes and their effects. But a statistical model is never sufficient for inferring cause, because the statistical model makes no distinction between the wind causing the branches to sway and the branches causing the wind to blow. Facts outside the data are needed to decide which explanation is correct.

Cross-validation and information criteria try to guess predictive accuracy. When I introduced them above, I described overfitting as the primary paradox in prediction. Now we turn to a secondary paradox in prediction: *Models that are causally incorrect can make better predictions than those that are causally correct.* As a result, focusing on prediction can systematically mislead us. And while you may have heard that randomized controlled experiments allow causal inference, randomized experiments entail the same risks. No one is safe.

I will call this the **IDENTIFICATION** problem and carefully distinguish it from the problem of raw prediction. Consider two different meanings of “prediction.” The simplest applies when we are external observers simply trying to guess what will happen next. In that case, tools like cross-validation are very useful. But these tools will happily recommend models that contain confounding variables and suggest incorrect causal relationships. Why? Confounded relationships are real associations, and they can improve prediction. After all, if you look outside and see branches swaying, it really does predict wind. Successful prediction does not require correct causal identification. In fact, as you’ll see later in the book, predictions may actually improve when we use a model that is causally misleading.

But what happens when we intervene in the world? Then we must consider a second meaning of “prediction.” Suppose we recruit many people to climb into the trees and sway the branches. Will it make wind? Not much. Often the point of statistical modeling is to produce understanding that leads to generalization and application. In that case, we need more than just good predictions, in the absence of intervention. We also need an accurate causal understanding. But comparing models on the basis of predictive accuracy—or p -values or anything else—will not necessarily produce it.

So what can be done? What is needed is a causal model that can be used to design one or more statistical models for the purpose of causal identification. As I mentioned in the neutral molecular evolution example earlier in this chapter, a complete scientific model contains more information than a statistical model derived from it. And this additional information contains causal implications. These implications make it possible to test alternative causal models. The implications and tests depend upon the details. Newton’s laws of motion for

example precisely predict the consequences of specific interventions. And these precise predictions tell us that the laws are only approximately right.

Unfortunately, much scientific work lacks such precise models. Instead we must work with vaguer hypotheses and try to estimate vague causal effects. Economics for example has no good quantitative model for predicting the effect of changing the minimum wage. But the very good news is that even when you don't have a precise causal model, but only a heuristic one indicating which variables causally influence others, you can still do useful causal inference. Economics might, for example, be able to estimate the causal effect of changing the minimum wage, even without a good scientific model of the economy.

Formal methods for distinguishing causal inference from association date from the first half of the twentieth century, but they have more recently been extended to the study of measurement, experimental design, and the ability to generalize (or *transport*) results across samples.^[24] We'll meet these methods through the use of a **GRAPHICAL CAUSAL MODEL**. The simplest graphical causal model is a **DIRECTED ACYCLIC GRAPH**, usually called a **DAG**. DAGs are heuristic—they are not detailed statistical models. But they allow us to deduce which statistical models can provide valid causal inferences, assuming the DAG is true.

But where does a DAG itself come from? The terrible truth about statistical inference is that its validity relies upon information outside the data. We require a causal model with which to design both the collection of data and the structure of our statistical models. But the construction of causal models is not a purely statistical endeavor, and statistical analysis can never verify all of our assumptions. There will never be a golem that accepts naked data and returns a reliable model of the causal relations among the variables. We're just going to have to keep doing science.

Rethinking: Causal salad. Causal inference requires a causal model that is separate from the statistical model. The data are not enough. Every philosophy agrees upon that much. Responses, however, are diverse. The most conservative response is to declare “causation” to be unprovable mental candy, like debating the nature of the afterlife.^[25] Slightly less conservative is to insist that cause can only be inferred under strict conditions of randomization and experimental control. This would be very limiting. Many scientific questions can never be studied experimentally—human evolution, for example. Many others could in principle be studied experimentally, but it would be unethical to do so. And many experiments are really just attempts at control—patients do not always take their medication.

But the approach which dominates in many parts of biology and the social sciences is instead **CAUSAL SALAD**.^[26] Causal salad means tossing various “control” variables into a statistical model, observing changes in estimates, and then telling a story about causation. Causal salad seems founded on the notion that only omitted variables can mislead us about causation. But *included* variables can just as easily confound us. When tossing a causal salad, a model that makes good predictions may still mislead about causation. If we use the model to plan an intervention, it will get everything wrong. There will be examples in later chapters.

1.4. Summary

This first chapter has argued for a rethinking of popular statistical and scientific philosophy. Instead of choosing among various black-box tools for testing null hypotheses, we should learn to build and analyze multiple non-null models of natural phenomena. To support this goal, the chapter introduced Bayesian inference, model comparison, multilevel models, and graphical causal models. The remainder of the book is organized into four parts.

- (1) Chapters 2 and 3 are foundational. They introduce Bayesian inference and the basic tools for performing Bayesian calculations. They move quite slowly and emphasize a purely logical interpretation of probability theory.
- (2) The next five chapters, 4 through 8, build multiple linear regression as a Bayesian tool. This tool supports causal inference, but only when we analyze separate causal models that help us determine which variables to include. For this reason, you'll learn basic causal reasoning supported by causal graphs. These chapters emphasize plotting results instead of attempting to interpret estimates of individual parameters. Problems of model complexity—overfitting—also feature prominently. So you'll also get an introduction to information theory and predictive model comparison in Chapter 7.
- (3) The third part of the book, Chapters 9 through 12, presents generalized linear models of several types. Chapter 9 introduces Markov chain Monte Carlo, used to fit the models in later chapters. Chapter 10 introduces maximum entropy as an explicit procedure to help us design and interpret these models. Then Chapters 11 and 12 detail the models themselves.
- (4) The last part, Chapters 13 through 16, gets around to multilevel models, as well as specialized models that address measurement error, missing data, and spatial covariation. This material is fairly advanced, but it proceeds in the same mechanistic way as earlier material. Chapter 16 departs from the rest of the book in deploying models which are not of the generalized linear type but are rather scientific models expressed directly as statistical models.

The final chapter, Chapter 17, returns to some of the issues raised in this first one.

At the end of each chapter, there are practice problems ranging from easy to hard. These problems help you test your comprehension. The harder ones expand on the material, introducing new examples and obstacles. Some of the hard problems are quite hard. Don't worry, if you get stuck from time to time. Working in groups is a good way to get unstuck, just like in real research.