Chapter 4

Models and Truth

All models are wrong, but some are useful – George E.P. Box

When considering the relationship between models and truth, it is useful to take a step back and first discuss different kinds of models. Modeling is a wide-ranging field with many distinctions made by modelers and mathematicians. Such distinctions are generally of little interest to us, as we believe focusing on them can often encourage a focus on jargon and formalism rather than the quality of a model. Furthermore, we will present our own classification that clarifies the core dichotomy at the heart of modeling. It helps, however, to briefly discuss the distinctions that are commonly made in order to obtain a deeper understanding of the choices underlying the development of a model.

Deterministic versus Stochastic Models

There are two polar opposite views of the world. One view says the fate of the universe is governed by strictly predictable laws of physics. In this view, the universe acts as if it were a giant machine, where if its current state is known (down to each individual atomic particle), its future states through the rest of time are predetermined. The opposite view is that the universe is ruled by chance and randomness. Random quantum mechanical fluctuations merge and amplify leading to an infinite range of diverging possibilities.

Which of these two views holds more of the truth? We certainly do not know and it is possible that this will be a question that physicists will never cease exploring. Albert Einstein had a viewpoint of special interest, however. He was a strong partisan of the more deterministic view, famously remarking, "God does not play dice with the world."

When creating a model of a system, we must choose how we treat chance. Do we build our model deterministically, such that each time we run it we obtain the same results? Or do we conversely incorporate elements of uncertainty so that each time the model is run we may see a different trajectory of outcomes?

Mechanistic versus Statistical Models

When beginning to build a model of a system, there are many questions you should ask, two of which are:

1. Do I know (or have a hypothesis of) the mechanisms that drive the system?
2. Do I have data that describe the observed behavior of the system?

If the first question is answered in the affirmative, you can build a mechanistic model that replicates your understanding (or hypothesis of) the true mechanisms in the system. If, on the other hand, the second question is answered in the affirmative, you can use statistical algorithms such as regressions to create a model of the system based purely on the data.

If neither question is answered affirmatively... well in that case there isn't much of anything you can build.

Aggregated versus Disaggregated

When building a model, the issue of scale becomes very important. Imagine we are concerned about the effects of Global Climate Change on water resources. We may wish to examine the question of whether there will be sufficient water supplies given a rise in future temperatures.

At what scale do we build this model? The range of possible scales is wide:

- At the most aggregate, we could estimate total worldwide water demands and supplies into the future.
- Maybe that is too coarse a scale, however, as clearly having excess water in Norway has little direct impact on the situation in Egypt. We could instead create a finer resolution model that separately looked at water demand and consumption in each country.
- Even that may still be too coarse, maybe we should make our model more granular to look at specific cities or population clusters around the globe.
- At the extreme disaggregated level, we might even want to model individual people – all 7 billion of them – and their needs and movements around the world.

There is no simple answer to this question of optimal scale. The best choice is highly context-sensitive and depends on the needs of the specific modeler and application.

Prediction, Inference and Narrative

The three distinctions just presented – deterministic vs. stochastic, mechanistic vs. statistical, aggregated vs. disaggregated – can be used to classify models.

We can even use them to classify the models we have discussed in this book. Most of our models would be classified as deterministic (random chance is generally not explicitly incorporated in these models), mechanistic (we generally assume mechanisms rather than estimating dependencies from data), and highly aggregated (the agent based models are an exception).

Outside of modelers, however, these distinctions are of little importance. Let's take off our modeler hats for a moment, and instead look at modeling from the perspective of a client (a loose term, it can include consulting arrangements but also work within an organization or in other contexts). As clients, we engage the modeler to build a model to achieve a specific purpose. Most of the choices the modeler makes are jus technical details. They are similar in importance to the choice of software or programming language used to build the model. It would make as little sense to say a model was fundamentally bad because it was written in a relatively ancient programming language – like Pascal – as it would be to say a model was fundamentally bad because it was, for instance, deterministic.

What is of true importance is the success of the model in fulfilling the client's goals whatever they may be. Technical details matter – they can affect maintainability and other factors – but they are of secondary interest. Let's look back at Box's quote at the beginning of this chapter. We know all models are wrong, what we should really care about is their utility in meeting a specific task.

So instead of using the aforementioned technical classifications to discuss models, we think it is more useful to base our discussions of models on the model's driving purpose. This allows us to leave behind relatively mundane technical and implementation details to focus on what we really care about. Among the many different reasons for building models, they all boil down basically to three broad purposes: prediction, inference and narrative.

Prediction : Models used for prediction are the most straightforward. They attempt to forecast some outcome given information about variables that may influence that outcome. A weather forecast is an example of a model used for prediction. Likewise, when you apply for a credit card at a bank, they run a predictive model to determine your risk of default. When you apply for life insurance, the company has an actuarial model to predict how much they should charge you for a given payout. All these models take in data (the current temperature for the weather forecast, the amount of money in your bank account for your risk of default, your age for the life insurance application) and apply various forms of analysis to generate a prediction of the outcome.

Inference : Models used for inference are most common in academic research. Often, academic research questions distills down to this simple template: "Does X affect Y?" These are inferential questions. As an example, a researcher may make a hypothesis statement such as, "The wealthier a high-school student's family is, then the higher the student's test scores will be." The researcher

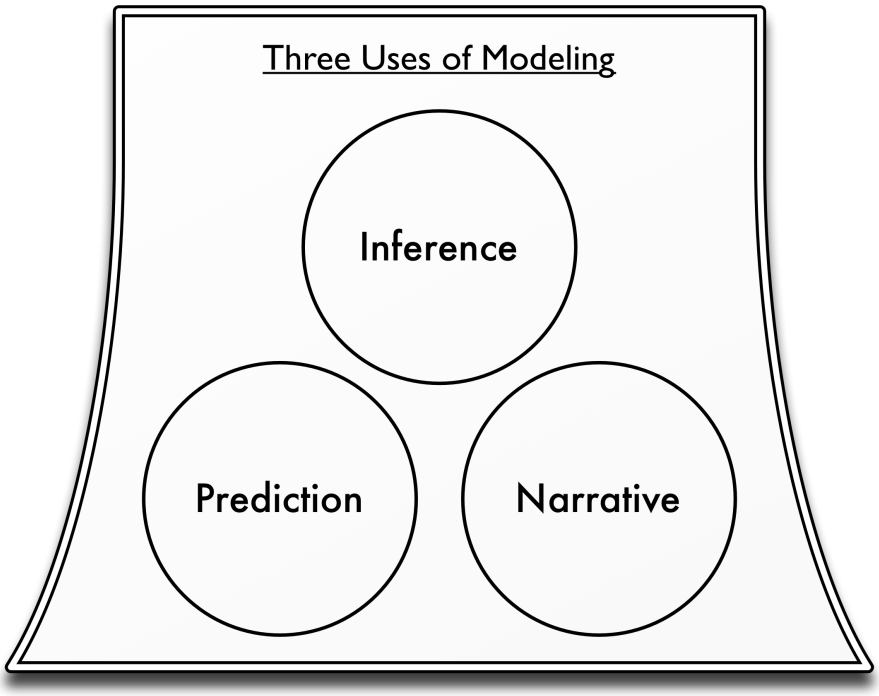

Figure 1. Three Usages of Models

may then build a model to test the validity of this hypothesis and the model's results will generally be phrased in terms of a p value indicating the statistical significance of the evidence in support of the hypothesis.

Narrative : Models are often used to tell a persuasive story. When the Obama administration wanted to persuade lawmakers and the public to support their economic stimulus, they famously published the graph shown in Figure 2. A great deal of complex modeling and mathematics surely went into constructing this figure. However its core purpose was to tell the nation a story: Things are going to be bad, but the recovery plan will make them less so. Such stories are at the heart of narrative models and we will return to this figure later on.

All models can be classified in terms of these three primary purposes and we will see how useful it is to discuss modeling projects in terms of them[ˆAnd we strongly recommend doing so. It is important to clearly define the purpose at the start of a project. The techniques used and data required depend significantly

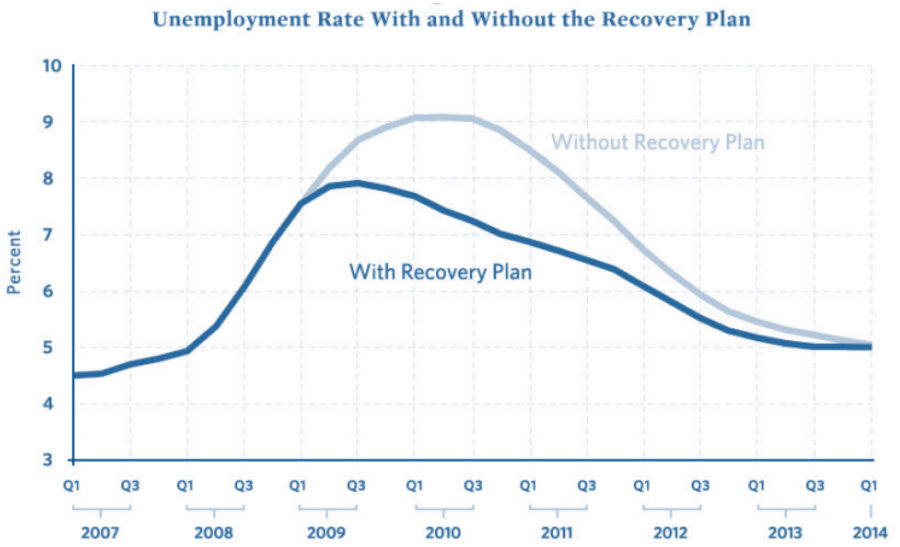

Figure 2. The Obama Administration's Predictions for the Effects of the Recovery Plan (Romer and Bernstein 2009)

on the models overall purpose. To be very clear, it is important to clarify at the outset whether your primary goal is to use a model for prediction or for narrative. Many clients and modelers expect to do both and end up with a model that does neither.].

The Strange Case of Inference

To help us get at this fundamental classification scheme, let's first talk for a moment about the process of inference. Take the earlier example of determining whether wealth results in increased high-school test scores. We phrased this hypothesis in a specific way: that increased wealth will always increase test scores. This illustrative statement, however, actually differs from what is often done in practice. In general, researchers simply asks the question "Does X affect Y?" rather than "Does X increase Y?" It's just a slight difference, but it is a more flexible question that allows for many forms of relationships. For our example, we would ask the question "Does wealth affect tests scores?"

The gold standard to answering questions like this is the controlled experiment. For our example, we could imagine an experiment where we took a sample of a thousand families from a school district. When these families' children enter high school we would randomly select them to be in a "poor" category and the other half to be in a "rich" category. Families in the rich category are given grants of $500,000 a year to spend how they wish while the parents in the poor category are fired from their jobs and have their savings frozen for the duration

of the experiment. Once the students graduate from high school, we would compare the scores for the students in the poor and rich categories.

These controlled randomized experiments are considered the ideal approach to answering inferential questions like these as they allow you to truly determine the effect of what your variables, in this case wealth. For many types of questions, such experiments can be implemented (for instance does treatment with a new drug help treat a disease). Unfortunately, in general complex social questions are simply impossible to answer with them. We can consider the testing procedure we just imagined to assess the effect of wealth on scores, but it would be impossible (and unethical) to undertake in a real community.

Traditional Model Based Inference

Given our general inability to undertake ideal controlled experiments, how do we answer inferential questions? The standard way is to collect data and then construct a model enabling us to measure the statistical significance of our hypothesis given the data. Due to history and simplicity, linear regression models are by far the most commonly used type of model today. A linear regression predicts an outcome (Y) based on the multiplication of variables (X's) by a set of coefficients determining the effect of the variables on the outcome (β's):

$$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 ...$$

For the education example we could collect data on a number of students, measuring their families' wealth (X_1 in the equation above) and the student's test scores (Y). We would then run the linear regression to determine the coefficient values (β_0 – the intercept – and β_1 – the effect of wealth on test scores). If we thought there were other factors that affected test scores, we could measure them and include them as addition X's in the regression.

In addition to obtaining the values of these coefficients, we also obtain as a result from the regression the statistical significances or "p values" of these coefficients. Although p values are commonly used in statistics, they are ubiquitously misunderstood(ˆThese misunderstandings are not only made by on-the-ground practitioners and analysts, they are frequently shared, and propagated, by university-level statistics instructors; see, for instance, Haller and Krauss 2002.) so it is useful to briefly review them.

In short a p value measures the probability of seeing the measured data (or more extreme data) assuming the null hypothesis is true. Generally the null hypothesis will be that there is no relationship between the variables and the outcomes.

When assessing the significance of coefficients, a p value means the probability of seeing that value of a coefficient (or one even further from 0), assuming that

the (unknown) truth is that the coefficient actually has a value of 0. In other words, it is the probability of seeing the observed non-zero value, assuming that the true value is in fact 0. Frequently, probabilities of 10%, 5% or 1% or smaller are taken as indicating statistical significance. These low values indicate that the coefficient value is so far from 0, and the probability of this occurring by chance so small, that we can reject the null hypothesis and accept the fact that the coefficient is not 0.

This is what a p value is. Now let's specifically mention what a p is not, as this is so often misunderstood. p values do not represent the following commonly used interpretations:

- The probability that the null hypothesis is true (that the coefficient is 0)
- One minus the probability that the alternative hypothesis is true (that the coefficient is not 0)
- Any sort of "proof" that the null or alternative hypotheses are correct or incorrect
- The probability that you are making the correct or incorrect decision if you accept or reject the null or alternative hypothesis

Using the p values enables inference by relying on the statistical significance of the coefficients. If the probability of β_1 (the coefficient for the effect of wealth) occurring due to chance (given it is 0 in reality) is less than, say 5%, we can claim with reasonable strength that wealth does in fact affect test scores. This is the standard approach researchers take to model-based inference and is used ubiquitously.

A Troubled Sea of Assumptions

Let's stop for a second and consider what we have done here. In carrying out these logical steps to apply model based inference to determine whether wealth affects test scores, we have had to make one very large assumption: that the relationship between test scores and wealth is linear.

Our linear regression equation assumes that for every increase in one unit of wealth (X_1), test scores (Y) will increase on average by the amount of the coefficient (β_1). What if this were not in fact the truth? For instance, we could easily imagine the case where wealth initially helped test scores by providing students more resources and opportunities to learn. After a certain point, however, wealth might negatively impact scores as very wealthy students might lack the pressure or motivation to study hard.

If we believed this were the case, then our linear regression model from earlier would be wrong as would the inferences we obtained from the model. We could correct our model and inferences by changing our regression formula to contain a squared term that could replicate this type of relationship:

$$Score = \beta_0 + \beta_1 \times \text{Wealth} + \beta_2 \times \text{Wealth}^2$$

Using this equation, at low values of wealth the $\beta_1 \times$ Wealth term will have the most effect on scores. Conversely, at high levels of wealth, the $\beta_2 \times \text{Wealth}^2$ term will have the most effect on scores. Thus by having a positive β_1 and a negative β_2 we can model wealth as havening an initially beneficial and then detrimental effect. If our assumptions about the quadratic relationship are correct, then this model will yield accurate inferences. If they are wrong, our inferences will be wrong again.

What are we really doing when we assume regression forms like this? Now it might not be immediately obvious, but what we are in fact doing is telling a story. Using our first equation, we are telling the story that as wealth increases test scores will almost always increase. Bill Gate's children will preform amazingly well here! Using the second equation we are telling a different story: As wealth increases test scores initially do as well but after a certain point increased wealth will hurt test scores. That picture isn't so rosy for the Bill Gates of the world!

And so we arrive at a key insight. By choosing our equations to tell a story, our inferences are in fact based on narrative modeling approaches. True, these inferences build upon numerous calculations and very advanced theoretical underpinnings, but ultimately what governs our conclusions and inferences are the stories or narratives we tell about our system. These are choices that we as narrators make and they not determined by an objective truth or reality.

Predictive Inference

Is there an alternative approach to inference that does not rely so heavily on narrative? Can we accomplish it without assuming the relationships between variables? The answer is yes. Although they are not often used, alternative prediction-based approaches to inference are available. In these approaches, rather than calculating statistical significances as a function of an assumed model, we calculate significances as a function of the simple question: "Does knowing X help us to predict Y?" This question is effectively identical to our earlier question – "Does X affect Y?" – but it is structured in an explicitly predictive manner. If the answer to the question is true, then we can say that there is a relationship between X and Y.

The techniques to accomplish prediction-based inference are much newer than classic techniques as linear regression as they rely upon extensive computing power and would not be possible without modern technology. One of these approaches is the *A3* method (XXX Citation) which uses resampling based algorithms to obtain estimates of predictive accuracy and statistical significance. *A3* focuses purely on predictive accuracy of a model to determine whether a variable is significant and often requires the automatic exploration of hundreds or thousands of competing models to find the one that best describes the data.

The results of these analyses are inferences that are founded in the data of model fits only, not on subjective assumptions.

Predictive versus Narrative Modeling

As we can see, inferential techniques can be split into two categories: those based on narrative modeling methods and those based on predictive modeling methods. So – and this is a key advance – from our original three categories of model purposes – prediction, inference, and narrative – we are left with just two fundamental types of modeling approaches: predictive modeling and narrative modeling.

This divide is not traditional used in the modeling field, but it is truly at the heart of modeling. Understanding the distinction between these two types of modeling proves below to be much more valuable than mastering fine technical details. The choice of whether to build a predictive or a narrative model is a fundamental one that shapes every aspect of a model and determines its ultimate utility for a specific purpose. The following sections will describe these two types of models in more detail.

Predictive Models

How do we define a predictive model? The naive answer is that a predictive model is one that makes predictions. If a model generates predictions for a future outcome or a given scenario, than it must be a predictive model. By this definition, a weather forecast is a predictive model as were the Obama administration's unemployment predictions we saw earlier.

Unfortunately, this straightforward definition is useless. Worse than being useless, it is actually quite dangerous.

Let us propose a model for next year's unemployment figures in the United States:

> Generate a random number from 0 to 1. If the number is less than 0.1, unemployment will be 20%. If the number is greater than or equal to 0.1, unemployment will be 0%.

There, we have just constructed a model of unemployment. Furthermore, our model creates predictions. With just a few calculations we can forecast unemployment for the coming year. Isn't that convenient?

Of course, this model is a joke. It is useless in predicting unemployment. However, using the naive definition of what it means to a predictive model, it would be classified as one.

What makes this simple model, such a poor model for prediction purposes?

There are several answers. We might start by saying it is too *simple*. If we are really trying to predict unemployment we should incorporate the current economic state and trends into our model. If the economy is improving, unemployment will probably drop and vice versa. This is a valid point. Let's address it by proposing an "improved" model:

> Generate a random number from 0 to 1. If the number is less than the percentage change in GDP over the past year, unemployment will be 20% plus the current unemployment rate. If the number is greater than or equal to 0.1, unemployment will be the net change in the consumer price index over the past 8 years.

Is this a better model? Clearly, it is more complex than the previous one and it incorporates some relevant economic data and indicators. Equally as clear, however, is that it is also a joke far from being a useful model.

These toy economic models show that just generating predictions is not a helpful criterion to define a predictive model. They also show that complexity and the use of relevant data is not a valid criterion. So how do we specify a predictive model? The answer is straightforward:

> A predictive model is a model not only creates predictions but also must contain an *accurate assessment of prediction error*.

Read that statement again. The key point is that the assessment of prediction error must be accurate, which is different from the accuracy of the predictions themselves. Of course, ideally the predictions will be accurate; however this is often not possible. Many systems are governed to a significant extent by chance and no model, no matter how good it is, will be able to create accurate predictions for the systems.

If you know the level of prediction error, you can instead contextualize poorly fitting models. You can determine how much to discount their predictions in your decision-making and analysis. Furthermore, and this is crucial, you can compare different predictive models. If your current model is insufficiently accurate, you can develop another one and objectively test it to determine whether it is better than the current model.

Without measures of predictive accuracy, discussing predictions or comparing models that create predictions is an almost nonsensical endeavor. Such discussions will be governed by political concerns and partisanship as there is no objective foundation on which to base them.

Our two proposed models to estimate unemployment are thus clearly not predictive as no estimate of predictive error has been established. We can

apply same this requirement to Obama's employment predictions we saw earlier. When we first presented the model, we called it a narrative model, which was probably slightly confusing as the model did generate predictions. However, using our above definition of a predictive model we can see also that it is in fact not a predictive model. The model contains no estimate of prediction error (and one is not available in the original report) so it simply cannot be considered to be predictive.

If accurate estimates of prediction error are available, you can directly compare the prediction errors between different models to select the one with the lowest error. We could estimate prediction errors for the two joke models we proposed here along with the Obama administration's model to find the one with the lowest error. We would hope that the one the Obama administration presented to Congress would be the most accurate. Before we test it however, we must not make the error of fallaciously accepting a model to be good based on who presented it to us or its complexity.

Why do we so rarely hear about the predictive accuracy of models? There are numerous reasons but they boil down to three basic ones:

1. Assessing prediction error accurately is quite difficult.
2. Sharing prediction error may perversely decrease an audience's belief in a model.
3. Most models people use for prediction are in reality narrative models and their predictive error is either abysmal or irrelevant.

Let's look at each point in detail. First consider the issue of the difficulty of assessing prediction error. In general, obtaining an accurate assessment of prediction error is much more difficult that developing the predictions themselves. Most commonly used approaches (for instance the standard R^2 from linear regression) have significant flaws. There are both theoretical and numerical methods that can be used to more accurate prediction errors in many cases (this will be discussed further in the section the Cost of Complexity; see also Fortmann-Roe (2012)). When dealing with time series data, however, like most of the models explored in this book, it is often almost impossible to accurately assess model prediction error. Recently, theoretical technique to approach these issues have just begun to be developed (e.g. He, Ionides, and King (2009) or A. A. King et al. (2008)) but they are still impractical to apply in many cases so far.

If the challenge of measuring prediction error is surmounted, there is an even more formidable barrier to its being published with the model. There is a perverse phenomena that the act of reporting prediction error can *decrease* the confidence an audience gives a model. An anecdote was relayed to us by a member of a team working on a model of disease spread. His team shared the predictions from the model with a group of policy-makers. Everything

was going fine until the audience saw the error bars around the predictions. Where his audience had been content with the raw predictions, they were quite unhappy with the predictions when accompanied by their accurately estimated uncertainties. Why was this? Was the team's model particularly bad or did these policy-makers have a better model at their disposal? No. In a world where policy-makers and clients are constantly shown models (like Obama's unemployment figures) with no measure of uncertainty (or even worse, poorly calculated, artificially low uncertainty), they come to have unrealistic expectations and often turn away good science in favor of magical thinking.

Finally, the most likely reason supposedly predictive models do not include prediction error is that they simply are not predictive. We have seen how models developed for a purportedly predictive purpose can actually be narrative models in disguise. Just why is this too often the case? You need only look at the reason for most modeling projects. It is very rare that models are commissioned solely for the purpose of generating an accurate prediction. Frequently, models are part of some political process within an organization or across them. Ultimately, the people funding the model expect it to prove a point to their benefit. In environments like these, it is to be expected that some predictive modeling efforts will be sidetracked by political concerns or compromised in the process.

We can see the results of such influences in the predictions generated for unemployment presented earlier. Figure 3 shows the projections for the unemployment rates with and without the stimulus plan just as in Figure 2. Overlaid on this are now the true values of unemployment the occurred after the predictions were made. As is readily evident, the original modeling and predictions were well off the mark. Not only was reality worse than the projections assuming the stimulus was enacted (which it was) it is much worse than the projections for the economy assuming the stimulus had never been enacted at all! This is just a small example – one that is sadly replicated over and over again in business and policy-making – of mistakenly treating a narrative model as a predictive one.

Narrative Models

In contrast to predictive models, a narrative model is one built to persuade by telling a story. When many people first hear the "narrative" terminology, they respond negatively. "It's just a story." We find this strange, as narratives are the fundamental human form of communication. We tell narratives to our friends and relatives. Politicians communicate their policies to us using narratives. Of course the vast majority of our entertainment is focused on narratives[^Even sports, a form of entertainment that innately contains no narrative, becomes wrapped in narrative as the announcers and commentators attempt to create stories to engage us.]. Business leaders and managers attempt to describe their strategies to us using storylines; and business books are in general dominated by anecdotes plotted along the way to making their points.

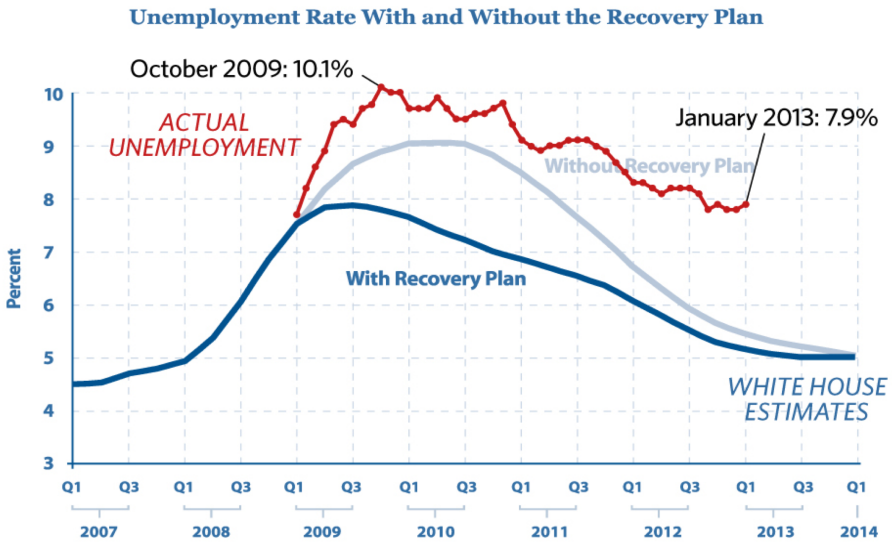

Figure 3. Unemployment predictions versus reality (The Heritage Foundation 2013)

We as a species do not view the world as a collection of numbers and probabilities; instead we see consequence and meaning. In short, narratives are how we see the world.

One critique of the term narrative is that it lacks numbers, quantified data, or mathematics. This could not be further off the mark. There are many ways to construct narratives. Words are one, pictures are another, and music is a third. Numbers and mathematics are just another way of telling a story.

In fact, most statistical and mathematical models are infused with narrative models. We looked earlier at the case of linear regression as a tool to predict test scores as a function of wealth. Again the mathematical equation for this simple model was:

$$Score = \beta_0 + \beta_1 \times Wealth$$

This equation defines a narrative. Translating this narrative into words, we would say:

> Test scores are only determined by the wealth of a student's family. A child whose family is broke will have a test score, on average, of β_0. For every dollar of wealth a child's family accumulates, the child will score, on average, better on tests by β_1.

You might or might not agree with this storyline (in our view it is a nonsensical and reductionist view of child achievement) but it shows the strict equivalence between this mathematical narrative and narrative prose. This process can be applied to all mathematical models. The mathematical definition of the model can be converted directly, with more or less lucidity, into a story describing how the system operates. The same can also be done in the reverse: we can take a descriptive narrative of a system and convert it into a mathematical description. As we have seen (will see? XXX) this is what tools like reference modes and pattern matching are designed to do efficiently: elicit a narrative from a subject in a way which can be reformulated quantitatively.

With predictive models, we can compare competing models based primarily on predictive accuracy[^Other criteria include ease of use, cost of filling data requirements, and computational requirements. But all those are generally secondary to prediction accuracy.]. But how do we evaluate and compare the quality of narrative models?

The key criterion in assessing a narrative model is its ability to be *persuasive*. Although persuasion is not an objective measure in the same sense prediction accuracy is, we can decompose persuasiveness into two components for our purposes: believability and clarity. A persuasive model is one that is both believable and effectively communicates its message.

When building a narrative it is very important to use tools that are well suited to meeting these components. Unfortunately, many statistical models, including regressions, are poorly suited to this two-fold task in many ways. Most statistical models depend on unrealistic and highly technical assumptions about the data. If these assumptions were enumerated in plain English, they would often conflict with people's understanding and in fact end up discrediting the model. The "alternative" has been to leave these assumptions hidden creating a black box model opaque to outside inspection.

This is a shame in our view. Such a stratagem can be successful if the authority presenting the model is prestigious enough. But the misdirection will quickly fail if any kind of rigorous scrutiny is applied to the model. Narrative models should never be given any real credence if the operation of the model is not transparent. Most statistical models are built on assumptions that are never made transparent to the audience. There is a reason for the adage, "there are lies, damned lies, and statistics."

The modeling techniques presented in this book, on the other hand, are well suited for narrative modeling. The techniques we present are "clear box" modeling where the workings of the model are transparently evident and accessible. Our models have their structure explicitly described using an accessible modeling diagram showing the interactions between the different components in the model. The equations governing the model's evolution are clear and readily available for each part of the model[^Admittedly, for complex models it may still require a significant investment on the part of

an audience to fully understand the logic and equations in the model. But the opportunity is available.]. Furthermore, these modeling techniques used here make it straightforward to generate animated illustrations and displays to clearly communicate model results.

Confidence Building Steps for Narrative Models

When used correctly, the transparency of these modeling techniques results in models that are powerful persuasive tools. As with any model, however, there are concerns and questions will invariably be raised which could cause users to doubt the result of the modeling work. There are a number of techniques that you can use to help preemptively address these concerns and increase an audience's confidence in your model.

The idea of building confidence in a model is closely tied to the standard concept of model verification and validation. We dislike this conceptual approach to assessing models as it seems to imply that a model can go through a process to get a big fat "VALID" or "VERIFIED" stamp on it. Returning to Pox's quote at the beginning of this chapter, in reality all models are wrong and none of them are valid, period. Models can however be useful, especially narrative models in which the audience has confidence.

We favor the conceptual approach put forth by Forrester and Senge (1978), that there is not any single test or suite of tests that will verify or validate a model and that validity should instead be thought of as a function of confidence. This is a view that differs from that held by some modelers and laypeople. As Forrester and Senge note, "the notion of validity as equivalent to confidence conflicts with the view many seem to hold which equates validity with absolute truth." We share their belief that model confidence is built up piece by piece from a variety of tests that, though they cannot prove anything, together comprise a persuasive case for the quality of a model.

There are three distinct areas where confidence needs to be developed:

1. That the model itself is well designed.
2. Given a design of the model, this design is implemented correctly.
3. The conclusions drawn from the model are accurate.

Model Design

Fundamentally the design of a narrative model is of utmost importance and needs to be justified to an audience[^This is different from predictive models where the results of the model are much more important than the design and the "proof is in the pudding" so to speak.]. There are two primary aspects to a model's design design: the structure of the model and the data used to parameterize the model.

Structure

The structure of the model should mirror the structure of the system being simulated. Depending on the system complexity, the model structure may need to carry out more or less aggregation and simplification of this reality. Nevertheless, all the primitives in the model should map on to reality in a way that is understandable and relatable to the audience. Furthermore, the model structure should include components that an audience thinks are important drivers of the system. Missing a factor that the audience considers to be a key driver can fatally discredit a model in an audience's mind irrespective of the performance or other qualities of a the model. This is true even if the factor has in fact a negligible effect. Generally speaking, it is much easier to include a factor an audience views as important than it is to convince the audience that the factor does not in actuality matter later on.

Data

The more a model uses real-world data, the more confidence an audience will have in the model. Ideally, you have empirical data to justify the value of every primitive in your model. In practice, such a goal may be a pipe dream. Indeed, for a complex model, obtaining data to parameterize every aspect of it is usually impossible[ˆLeading to the clichéd conclusion of many modeling studies: "We are unable to draw strong conclusions from this modeling work. Instead, our contribution has been to show where additional data needs to be collected."]. When faced with model primitives that do not have empirical data to parameterize them, an approach must be taken to ensure that it does not appear that their values were chosen without justification or to arrive at a predetermined modeling conclusion. Sensitivity testing, as discussed later on, is one way to achieve this. Another is to carry out a survey of experts in the field in order to solicit a set of recommended parameter values that can then be aggregated or used to justify the ultimate parameterization. Remember, you cannot be definitive but still timely when it comes to using models for policymaking; even if you could achieve a full model based on comprehensive point in time data, by the time you are done, the model is out of date.

Peer-Review

Going through a peer-review process can be extremely useful in establishing confidence in a model. Two general types of peer-review are available. In one, the model may be incorporated into an academic journal article and submitted for publication. The article will then peer-reviewed by generally two or three anonymous academics in the field who critique it and judge whether or not it is a worthy contribution to the literature, thus meriting publication. In the second type of peer-review, a peer-review committee may be assembled (hired) to review a specific model and provide conclusions and recommendations to clients.

Peer-review can be very useful in establishing the credibility of a model. A credible model is a model one can be more confident in, other things being equal. By engaging an independent group of exports to assess the model, their conclusions about its quality have the appearance of greater validity than those of the self-interested modelers[^When the peer review panel is hired by the client there is some conflict of interests, but the panel members should not be swayed by this.]. This can be especially useful when trying to meet some abstract standard such as that the model represents the "best available technology" or the "best available science".

A key risk of a peer-review is, of course, that the peer-review members will find a model deficient in important respects. Good criticism can be very useful and help improve a model. However, some criticism received in practice may be nitpicking details or detrimental advice that would make the model worse if followed.

Model Implementation

Although it is not as much a lightning rod as is model design, the implementation of a model specification is a point where significant error can occur. Programming mistakes or mistyped equations can introduce bugs into a model that can be hard to identify later on. This is a particular problem in black-box models but it is still an important point to consider for all types of models including those presented in this book. Fortunately, a number of steps can be taken to ensure the model is implemented correctly.

Primitive Constraints

For many of the primitives in the model, there will be natural constraints. For instance, a stock representing the volume of water in a lake can never fall below 0. Similarly, if a variable represents the probability of an event occurring, it must be between 0 and 1.

Often these constraints are implicit without being formally specified in the model. A modeler may think, water volume can never become negative so why would I need to specify it? However, the existence of these constraints provides an opportunity to implement a level of automatic model checking. By specifying that a primitive can never go above or below a value (using the *Max Value* and *Min Value* properties in Insight Maker), you can create in effect a canary in the coal mine that warns if something is wrong in the model. If these constraints are violated an error message can be given letting you know that you need to correct some aspect of your model.

Unit Specification

Since we introduced units in Chapter 3, we showed that they could be a useful tool in constructing models. Units can also be used to ensure that equations

are entered correctly. If you fully specify the units in a model, many types of equation errors will result in invalid units, which will create an immediate error. By employing units in your model you can automatically catch a whole class of errors and mistyped equations.

Regression Tests

Other tests than those specified above can be developed. For instance, the proper behavior of one part of the model may be determined and automated tests created to periodically confirm that the model continues to exhibit the correct behavior. Development of such tests are a common part of software engineering that we wish would see more use in model development. Insight Maker itself has a suite of over 1,000 individual regression tests that automatically test every aspect of its simulation engine.

In regards to regression testing, it is important to ensure these tests are automated. It is not enough to examine a portion of the model, determine it is currently working correctly, and leave it at that. The problem is that future changes may break the existing functionality. Especially for complex models, a change in one part of the model may have an unexpected effect in another part. By implementing a set of automatic checks, you can protect your model against unintended changes and regressions.

A Second Pair of Eyes

That is not to say, however, that spot and point-in-time checks are not worthwhile. It can be very useful to have a second modeler review your models and cross-check the equations. This helps not only to check simple mistakes but also to question and critique the fundamental structure and choices of the model.

The gold standard in verifying that a model is implemented correctly according to specification is to have a second modeler completely reimplement the model according to that specification. Such reimplementation should ideally occur without access to the original model's code base to ensure that the second modeler does not simply copy bugs from the original model into the reimplementation. If the results from the two implementations concur, that is strong evidence that the model has been implemented correctly. Although potentially an expensive exercise, it will also most likely identify numerous ambiguities in the specification, which could be valuable in and of itself.

Model Results

Given that the design of the model and its implementation are assumed to be correct, the burden still falls upon the modeler to transfer her confidence in the model's results to her audience. There are several different ways this can be done.

Expected Results

The first way is to demonstrate that the model generates expected results for normal inputs. For instance, if you had a model a reservoir, you would expect the volume of the reservoir to decline over time during the summer due to evaporation if no more water flowed into it. You can additionally test extreme scenarios and show that they generate the expected results. If, for example, your reservoir were empty, you would expect the amount of water to evaporate from it to be zero. By enumerating these standard cases and showing the model results match the expected results you can help build confidence in the model.

Counterintuitive Results

Another attempt to increase confidence in a model is to show unexpected results that are justifiable. Imagine a model that for a certain set of inputs would create what, at first glance, appeared to be the "wrong" behavior. Some lever in the model could lead to unexpected results. When first shown these results, they could decrease an audience's confidence in the model. If the audience was then walked through the model step by step to show how those results proved to be correct and mirrored reality, then that could well increase their confidence in the model results.

Forecasting

Possibly the most persuasive action to convince an audience of the effectiveness of a model is to forecast the future and then to show this forecast to be correct. This, of course, is difficult to do in practice for multiple reasons including the fact that the scale of a model is often such that it could take several years or decades to generate data to test the model. Additionally, it must be remembered that most narrative models are poor predictors and should not be used for predictive purposes solely.

Sensitivity Testing

Sensitivity testing is a broad field that has the potential to address many questions and doubts that may arise about a model. In general, the variables and numeric configuration values in a model will never be known with complete certainty. When the results from an election poll are published, the pollsters publish not only their predictions but also the uncertainty in the prediction (e.g., "the Democratic candidate will obtain $52\% \pm 3\%$ of the vote"). Similarly when a building is constructed, the materials used will have certain properties – such as strength – that again are only known up to some errors or tolerance. It is the engineer's and contractor's responsibilities to ensure that the materials are sufficient even given the uncertainty of their exact strengths.

The same occurs when modeling. Most primitive values in the model will have to be estimated by the modeler and there will be an error associated with these values. Of course the error will also be propagated through the model when it

is simulated and affect the results output by the model. This error is one factor that can create doubt about a model and reduce an audience's confidence.

As a modeler, one approach to address this doubt would be to try to measure all the model's variables with great accuracy. You could search the available literature, undertake a meta-analysis of current results, carry out new experiments, and survey experts to get as precise a set of parameter values as possible. If you were able to say with strong certainty that these values were so accurate and the errors so small that their effect on the results in negligible, then that would be one way of addressing the issue of uncertainty.

However, all of this is often impossible to do. When dealing with complex systems it is almost always the case that at least a couple variable values will never be known fully with certainty. In this case, no matter how much research or experiments you do, you will never be able to pin down the precise values of these variables. How do we handle these cases?

The answer is straightforward: Rather than trying to eliminate the uncertainty, we embrace it by explicitly including it in the model. If you can then show that the results of your model do not significantly change even given the uncertainty, you have a persuasive case for the validity of your results. Of course the results will always change when the uncertainty is introduced, but if the model conclusions persist even in the face of this uncertainty it will greatly increase your audience's confidence in the results.

Uncertainty can be explicitly integrated into a model by replacing constant primitive values with a construct that represents the uncertainty in that value. Imagine you had a simple population model of rabbits in a cage. You want to know how many rabbits you will have after two years. However, you don't know how many rabbits there initially are in the cage. You have been told that there are probably 12 rabbits, but the true number could range anywhere from 6 to 18.

If you model your population as a single stock, what should the initial value be? A naive model could be built where you the initial value of the rabbit stock was specified as 12. However, that does not incorporate the uncertainty and could be a source of criticism or doubt for the model. An alternative would be to specify that the initial value of the stock is a random number with a minimum value of 6 and a maximum value of 18. So each time you run the model you will get a different result. If you ran the model once, the initial value might be chosen to be 7 and you would obtain one result. If you ran the model again, the initial value might be 13 and you would get another result.

If you run this stochastic model many times, you obtain a range of results. These results can be automatically aggregated to show the range of outputs. For instance if you ran the model 100 times you could see what the maximum and minimum final populations were. This would give you a good feeling for how many rabbits you needed to prepare for after two years. In addition to the maximum and minimum you might be interested in the average of these

100 runs: the expected number of rabbits you would see. You could also plot the distribution of the final population sizes using a histogram to see how the results are distributed. This distribution would show how sensitive the outputs are to the uncertainty in the inputs: a form of sensitivity testing.

[XXX Embedded Demo]

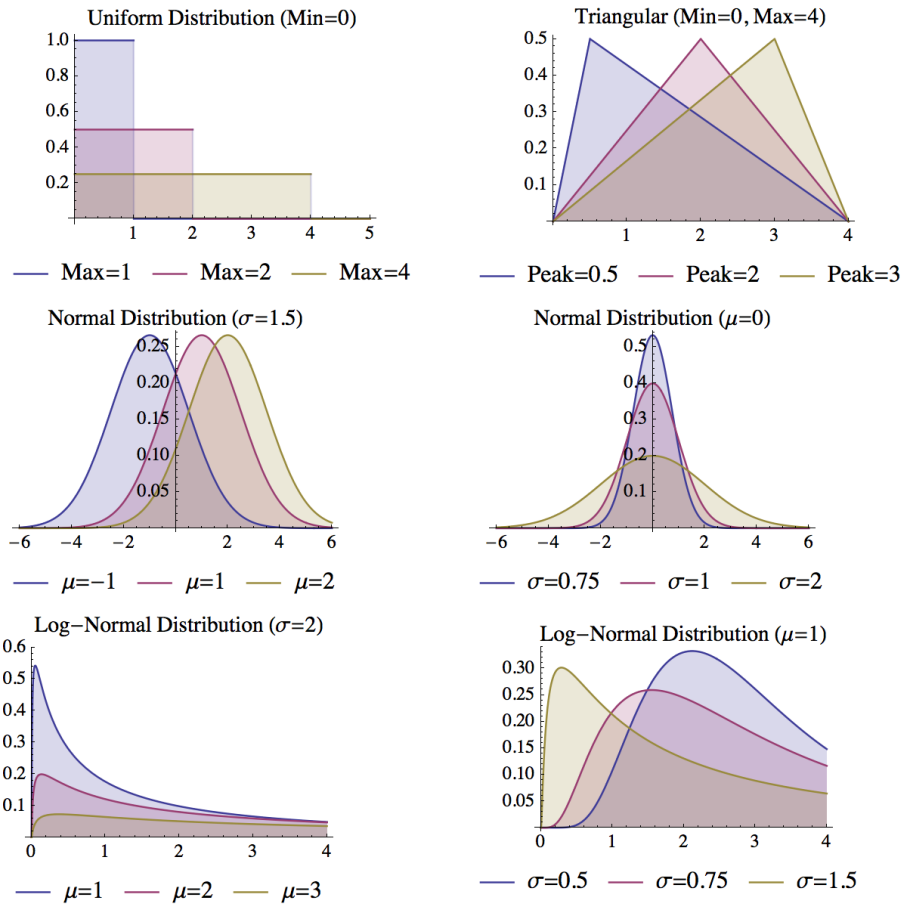

Figure 4. Common Distributions for Sensitivity Testing with Sample Parameter Values

There are four key distributions that are useful for specifying the uncertainty in a variable:

Uniform Distribution : The uniform distribution is defined by two parameters: a minimum and a maximum. Each number within these two boundaries has an equal probability of being sampled. The uniform distribution is useful when you know the boundaries on the values a variable can take on, but you do not

have any information on the likelihood of the different values within this region. The uniform distribution can be used in Insight Maker using the function *Rand(Minimum, Maximum)*, the two parameters are optional and will default to 0 and 1 if *Rand()* is called without them.

Triangular Distribution : The triangular distribution is defined by three parameters: the minimum, the maximum, and the peak. Like the uniform distribution, the triangular distribution will only generate numbers between the minimum and maximum. Unlike the uniform distribution, the triangular distribution will not sample all numbers between these boundaries with equal likelihood. The value specified by the peak will have the most likelihood of being sampled with the likelihood falling off as you move away from the peak towards either the minimum or maximum boundary. The triangular distribution is useful when you know the both the most likely value for a variable and you also know boundaries for the values a variable can take on. The triangular distribution can be used in Insight Maker using the function *RandTriangular(Minimum, Maximum, Peak)*.

Normal Distribution : The normal distribution is defined by two parameter: the mean of the distribution (generally denoted μ) and the standard deviation of the distribution (generally denoted σ). The most likely value to be sampled from the normal distribution is the mean. As you move away from the mean (in either a positive or negative direction), the likelihood of a number being sampled decreases. The standard deviation controls how fast this likelihood falls as you move away from the mean. Small standard deviations result in steep declines in the likelihood while large standard deviations result in more gradual declines. The normal distribution is useful when you do not have boundaries on the values for a variable but you do know what the most likely value for the variable should be (the mean). The normal distribution can be used in Insight Maker using the function *RandNormal(Mean, Standard Deviation)*.

Log-normal Distribution : The log-normal distribution is closely related to the normal distribution. In fact the logarithm of the values samples from a normal distribution will be log-normally distributed. Like the normal distribution, the log-normal distribution is defined by two parameters: the mean and standard deviation. Where the log-normal distribution differs from the normal distribution, is that negative values will never be generated by the log-normal distribution. Thus it is useful when you have a variable which you know cannot be negative but for which you do not have an upper bound. The log-normal distribution can be used in Insight Maker using the function *RandLogNormal(Mean, Standard Deviation)*. The log-normal distribution can also be used to represent other types of one-sided boundaries. For instance, the following equation could be used to represent a variable whose number was always less than 5: *5-RandLogNormal(2, 1)*

There are many other forms of probability distributions. Some notable ones are the Binomial Distribution (*RandBinomial(Count, Probability)*), the Negative Binomial Distribution (*RandNegativeBinomial(Successes, Probability)*), the

Poisson Distribution (*RandPoisson(Lambda)*), the Exponential Distribution (*RandExp(Lambda)*) and the Gamma Distribution (*RandGamma(Alpha, Beta)*). These distributions can be used to address very specific modeling usage cases and needs (for instance, the Poisson distribution can be used to model the number of arrivals over time), however, the four distributions described in detail above should generally be sufficient for most sensitivity testing needs.

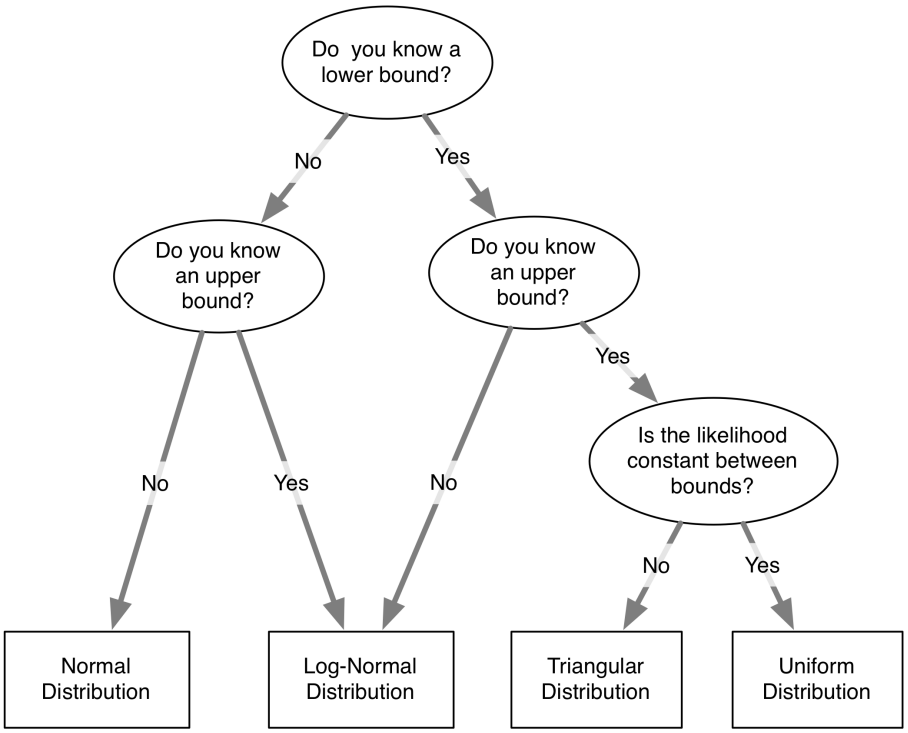

Figure 5. Choices in Selecting a Distribution for a Variable's Value

The astute reader will notice that our discussion up to this has failed to address an important point: how do we determine the uncertainty of a variable? It is very easy to say that we do not know the precise value of a variable, but it is much more difficult to define the uncertainty of it. One case where we can precisely define uncertainty is when you take a random sample of measurements. For instance, suppose our model included the height of the average American man as a variable. We could randomly select a hundred men and measure their heights. In this case our uncertainty would be normally distributed with a mean equal to the mean of our sample of one hundred men and a standard deviation equal to the standard error of our sample of one hundred men[^ Please note that this contradicts slightly what we said earlier. Clearly, a person cannot have a negative height while the normal distribution will sometimes generate negative values. So wouldn't a log-normal distribution be better than a normal

distribution? Mechanistically, it would, however statistically we can show that due to the Central Limit Theorem the normal distribution does asymptotically precisely model our uncertainty. Given a large enough sample size (100 is more than enough in this case), the standard deviations for uncertainty will be so small that the chances of seeing a negative number (or even one far from the mean) are effectively none.]. For any random sample of n values from a population, the same should hold true: you will be able to model your uncertainty using a normal distribution with:

$$\mu = \frac{Value_1 + Value_2 + Value_3 + ... + Value_n}{n}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Value_i - \mu)^2}$$

However, in most applied cases you will not be able to apply this normality assumption. Generally you will not have a nice random sample, or you might have no data at all and instead have some abstract variable you need to specify a value for. In these cases, it is up to you to make a judgment call on the uncertainty. Choose one of the four distributions detailed above and use whatever expert knowledge available to you to place an estimate on the parameterization of uncertainty. One rule of thumb, however, is that it is better to overestimate uncertainty than underestimate it. It is better to err on the side of overestimating your lack of knowledge than it is to obtain undue confidence in model results due to an underestimation of uncertainty.

Chapter 6

References

Forrester, Jay Wright, and Peter M. Senge. 1978. "Tests for building confidence in system dynamics models."

Fortmann-Roe, Scott. 2012. "Accurately Measuring Model Prediction Error" (apr). http://scott.fortmann-roe.com/docs/MeasuringError.html.

Haller, H., and S. Krauss. 2002. "Misinterpretations of Significance: A Problem Students Share with Their Teachers." *Methods of Psychological Research Online* 7 (1): 1–20.

He, D., E. L. Ionides, and A. A. King. 2009. "Plug-and-play inference for disease dynamics: measles in large and small populations as a case study." *Journal of The Royal Society Interface* 7 (43) (dec): 271–283.

King, Aaron A., Edward L. Ionides, Mercedes Pascual, and Menno J. Bouma. 2008. "Inapparent infections and cholera dynamics." *Nature* 454 (7206) (aug): 877–880.

Romer, Christina, and Jared Bernstein. 2009. "The job impact of the American recovery and reinvestment plan."

The Heritage Foundation. 2013. "Unemployment Rate January 2013" (feb). http://www.heritage.org/multimedia/infographic/2013/02/unemployment-rate-january-2013.