Probabilistic Modeling and Likelihoods Notes

Srikar Katta and Edoardo Airoldi

Contents

Probabilistic Modeling

Data science is an interdisciplinary field that is incredibly collaborative. Given the great heterogeneity between disciplines, many times "modeling" in one field is different than "modeling" in another, so we must translate this into the language that data scientists, statisticians, and machine learners use: probabilistic models. Using, this information, we can construct the data science pipeline.

Definition 0.1: Scientific and Probabilistic Models

A **scientific model** is a physical, conceptual, or mathematical representation of a real world system, process, or event.

As compared to scientific models, **probabilistic models** specifically deal with the study of how data was generated, taking advantage of randomness and variability through the assignment of probability distributions to different quantities.

Most real world problems typically deal with estimating some quantity given a set of other quantities. After gathering the data and positing a process by which the real data was produced, we calculate the likelihood function; we then maximize the likelihood function to find unknown quantities. Using our guesses for the unknown quantities, we then estimate our objective.

In general, variables are quantities that have a theoretical variation; that is, in our probabilistic model, these values come from a probability distribution. On the other hand, constants are quantities that do not have theoretical variation: they are fixed. We classify our quantities into four groups:

1. Known constants, values that are observed and have no theoretical variation
2. Unknown constants, values that are unobserved and have no theoretical variation
3. Latent/omitted variables, values that are unobserved and have no theoretical variation
4. Observations/data/random variables, values that are observed and have theoretical variation

We often refer to the theoretical results as the model statement, which describes the probabilistic process by which our data was generated and informs us on which quantities are variables and constants. The empirical information is given through the problem/data statement and describes what information is actually observed. Combining the information from both sources, we will estimate our quantities.

It is important to note one factor of modeling: "all models are wrong but some are useful." As a data scientist or researcher, our responsibility is to evaluate the trade-offs between accurately representing the real world

and being able to infer information from our models. For instance, suppose we assume that the only quantity that can impact income is age. Someone may criticize that model because it excludes other relevant features, such as education. However, by including education in our model, we would increase the complexity of the problem; if we do not have the time or tools to consider education in the model, then we are at a disadvantage because we would not be able to infer anything. Throughout this book, our goal is to employ the readers with the skills necessary to actually model real world processes and evaluate the trade-offs between accuracy and complexity. Early on, we do not want to overwhelm ourselves, which is why we have incredibly simple and potentially inaccurate representations of the real world early on in this book.

Empirical and Theoretical Distributions

In statistical modeling, there are two ideas: what we *believe* will happen – the theoretical – and what will actually happen – the empirical. In the assumption/model statement phase of any situation, we must assume whether a quantity has variability or not. While that sounds simple enough, there may be issues that arise. Take the following example: we assume that a fair, six-sided die has a uniform distribution, so the theoretical variation is greater than 0. However, suppose we roll only 1s, a completely possible outcome. Should we classify this as a known constant or a known variable? The model statement will guide the researcher, not the data itself. So because the theoretical variation is greater than zero, this is known variable – the empirical result has no impact on our modeling.

Consider the following two scenarios:

Let X_i be a random variable that represents the outcome of rolling a 6-sided die with 6 on **all sides**. That means that $x_i = 6$ with probability 1 (recall that the big letter means a random variable while the little letter means the realization of the random variable). Because $\mathbb{P}(X_i = 6) = 1$, X_i has no theoretical variation and is a constant. When we look at the data (i.e., the empirical distribution) and our theoretical distribution (i.e., $X_i = 6$ with probability 1), we should see that the two align.

In another scenario, suppose we have a random variable Y_i representing the outcome of rolling a 6-sided die with $1, 2, ..., 6$ on its sides, so each outcome has probability $\frac{1}{6}$. Then, $\mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i = 2) = \ldots \mathbb{P}(Y_i = 6) = \frac{1}{6}$. When we actually roll the die, it is possible that we roll only a sequence of 1s (i.e. $y_1 = 1, y_2 = 1, \ldots$); or we could roll another sample like $y_1 = 1, y_2 = 2, y_3 = 1, y_4 = 6, \ldots$, in which there is empirical variation. The empirical and theoretical distributions do not necessarily have to be equivalent to one another.

The first example is a probabilistic representation of a constant: x_i will always have the same outcome. On the other hand, in the second situation, y_i is not guaranteed to be the same outcome each time *empirically*. The distinction between the empirical and theoretical variability comes from the sample outcome versus the expected outcome.

So, we want to estimate something, given that only a subset of quantities and given some assumptions about how those quantities relate. The problem and model statements will help us categorize our variables into the following table:

	Observed	**Unobserved**
Variable		
Constant		

Example 1. Suppose we want to estimate some quantity, τ, which is a function of $y_1...y_n, x_1...x_n, \theta$, where each i is a user and $y_1...y_n$ is a series of expenditures that each i has made, x_i is a bunch of covariates for each user i (i.e., age, gender, race, etc), and θ is a set of scalar quantities for the model. Our goal is to now estimate θ given $y_1...y_n, x_1...x_n$.

First, we know θ is unknown. We must ask another question: "is θ an unknown constant or an unknown variable?" This information comes from the model statement.

Suppose our probabilistic model reads as follows:

$$\text{for } i = 1, \ldots, n : y_i = \theta x_i$$

So, if given an age bracket (x_i), then y_i–the amount that user i will spend–is just some scalar multiple of their age bracket. In this model, there is no variability because we did not make any assumptions about variables having a distribution. Because we never explicitly stated which distribution θ comes from, we classify θ as a constant. The 2x2 table for this problem reads as follows:

	Observed	**Unobserved**
Variable	NA	NA
Constant	$x_1...x_n, y_1...y_n$	θ

Now, let us try a new example with the same problem setting but that will lead to a different two-by-two table.

Example 2. Suppose we want to estimate some quantity τ that is a function of $y_1...y_n, x_1...x_n, \theta, \sigma$, where each i is a user, y_i is a series of expenditures that each i has made, x_i is a bunch of covariates for each user i (i.e., age, gender, race, etc), and θ is a set of scalar quantities for the model. Similar to the earlier example, our goal is to estimate θ given $y_1...y_n, x_1...x_n$.

Now, suppose we posit the following model assumptions:

$$x_i \sim \mathcal{N}(0, \sigma^2)$$
$$y_i = \theta x_i$$

We know that x_i is a random quantity that has a normal distribution with population variance σ^2, which is now different from the earlier problem. Let us classify our variables as before. Because θ does not explicitly come from a probability distribution, θ is a constant.

Before, we had assumed that our x_i was not distributed. But our x_i are still observed. So, we can classify them as observed variables. Now, notice that we have a new quantity: σ^2 an observed value. Since we don't make any assumptions about its distribution, it is an observed constant.

Classifying our y_i is a little bit more complicated than before, but we will learn later on that because x_i has some variability, y_i must also have some variability, even if θ is a constant. So, y_i is an *observed variable* also.

	Observed	**Unobserved**
Variable	$x_1...x_n, y_1...y_n$	NA
Constant	σ^2	θ

We will now outline how the same *problem* statement but with different *model* statements can lead to different 2x2 tables.

Example 3.

Consider the following problem/model statement:

$$y_i = \text{number of days } i \text{ purchases something}$$
$$z_i = \begin{cases} 1 & age(i) \geq 30 \\ 0 & age(i) < 30 \end{cases}$$

In this problem, we have four settings, each of which will lead to a different 2x2 table.

- **Example 4**: z_i is observed and $z_i \sim iid\ Bernoulli(p)$
- z_i is observed and no assumptions about distribution
- z_i is not given and $z_i \sim iid\ Bernoulli(p)$
- z_i is not given and no assumptions about distribution

Let's think about each of these settings individually.

Example 4. Consider the first problem/model statement from Example 3. We are given $y_1...y_n, z_1...z_n$, and our model assumptions are as follows:

$$y_i|z_i, \theta \sim Bernoulli(\theta + \alpha z_i) = \begin{cases} y_i|z_i = 1, \theta \sim Bernoulli(\theta + \alpha) \\ y_i|z_i = 0, \theta \sim Bernoulli(\theta) \end{cases}$$

Let's create the 2x2 table for this problem. We know $y_1...y_n$ is definitely observed because it is given. Now is it variable or constant? Well, it follows a distribution. And even though we don't know explicitly what $\mathbb{V}(y_i)$ is, we can calculate it as $\mathbb{V}(y_i) = \mathbb{E}(\mathbb{V}(y_i|z_i)) + \mathbb{V}(\mathbb{E}(y_i|z_i))$. Now, we know that z_i are given, *but* there is nothing about their distribution, so z_i is not variable.

	Observed	**Unobserved**
Variable	$y_1...y_n$	NA
Constant	$z_1...z_n$	θ, α

Now, let's discuss the theoretical versus empirical distributions for z_i a little more deeply:

- $\mathbb{V}(z_i) = 0$ since we made no assumptions about the distribution of v_i theoretically
- Now, if we computed the *empirical* variance:

$$\frac{1}{n}\sum_{i=1}^{n}\left(z_i - \frac{\sum_{i=1}^{n} z_i}{n}\right)^2 > 0 \text{ (most likely).}$$

  However, the empirical variance has *no* bearing in our classification of z_i as constant or variable. What we believe z_i to be comes only from our problem and model statements. In the absence of something that explicitly states z_i has a distribution, we consider z_i a constant.

Example 5. Consider the second problem/model statement from Example 3. We are given $y_1...y_n, z_1...z_n$. But now our model is as follows:

$$z_i \sim Bernoulli(p), p = 0.4$$
$$y_i|z_i \sim \mathcal{N}(\theta + \alpha z_i, \sigma^2) = \begin{cases} y_i|z_i = 0 \sim \mathcal{N}(0, \sigma^2) \\ y_i|z_i = 0 \sim \mathcal{N}(0 + \alpha, \sigma^2) \end{cases}$$

Let's first classify all of our variables. Just as before in Example 4, y_i has a distribution and is given to us, so it is a known variable. Also, θ–a parameter describing the distribution for y_i–is not given and has no theoretical probability distribution, so it is an unknown constant. Likewise, α is not given and has no theoretical variability, so it too is an unknown constant. Now, notice that z_i has a distribution and is given, so we now classify it as a known variable. The 2x2 table for this problem would look as follows:

	Observed	**Unobserved**
Variable	$y_1...y_n, z_1...z_n$	NA
Constant	NA	θ, α

Example 6. Consider the third problem/model statement from Example 3. We are given $y_1...y_n$, and our model assumptions are as follows:

$$z_i \sim Bernoulli(p), p = 0.4$$
$$y_i|z_i \sim \mathcal{N}(\theta + \alpha z_i, \sigma^2) = \begin{cases} y_i|z_i = 0 \sim \mathcal{N}(0, \sigma^2) \\ y_i|z_i = 0 \sim \mathcal{N}(0 + \alpha, \sigma^2) \end{cases}$$

Notice, the model statement is the exact same as in Example 5, but our model statement is now different: we do not know what $z_1...z_n$ are, so it is unknown. So, since we can calculate a *theoretical* variance for z_i, which will be greater than zero, it will be variable. Thus, z_i is an unknown variable. Other than that, it is the exact same 2x2 table as Example 5 with the exact same reasoning. The fact that z_i is not given has no bearing on the classification of other variables.

	Observed	Unobserved
Variable	$y_1...y_n$	$z_1...z_n$
Constant	NA	θ, α

Example 7. Consider the first problem/model statement from Example 3. We are given $y_1...y_n$, and our model assumptions are as follows:

$$y_i|z_i, \theta \sim Bernoulli(\theta + 2z_i) = \begin{cases} y_i|z_i = 1, \theta \sim Bernoulli(\theta + \alpha) \\ y_i|z_i = 0, \theta \sim Bernoulli(\theta) \end{cases}$$

Notice that this model statement is the exact same as in Example 4, but our model statement is now different: we do not know what $z_1...z_n$ are, so it is unknown. Additionally, if we were to calculate a *theoretical* variance for z_i, which we could by treating it as a random variable with probability 1, we would find that it has a variance of zero. So, it is a constant. Thus, it is an unknown constant. Besides that, the 2x2 table for this example and 4 are the exact same. The fact that $z_1...z_n$ is now unknown should not have any impact on our treatment of the other quantities.

	Observed	Unobserved
Variable	$y_1...y_n$	NA
Constant	NA	$\theta, \alpha, z_1...z_n$

Now that we know what known/unknown constants and variables are, we can start understanding the basics of the language of modeling between disciplines. These are the common terms that will represent each cell of our 2x2 table:

	Observed	Unobserved
Variable	Observed Random Variables	Latent Random Variables
Constant	Known Constants	Unknown Constants

Likelihood

A likelihood function is a function of the observed random variables–whatever they may be–given the constants from the problem and model statements.

Definition 0.2: Proper likelihood

The **proper likelihood** is a representation of how well the empirical and theoretical distributions align for our observed random variables. Mathematically–with observed random variables $Y_1, \ldots, Y_n$, realizations $y_1, \ldots y_n$, and unknown constants θ–the proper likelihood is $\mathbb{P}$(observed random variables| unknown constants) $= \mathbb{P}(y_1...y_n|\theta)$.

Definition 0.3: Complete likelihood

The **complete likelihood** is a representation of how well the empirical and theoretical distributions align for all our random variables, regardless of whether we observed them or not. Mathematically–with realizations of our observed random variable $y_1, \ldots y_n$, realizations of our unobserved random variable $x_1, \ldots, x_n$, and unknown constants θ–the complete likelihood is $\mathbb{P}$(observed random variables, latent random variables| unknown constants) $= \mathbb{P}(y_1, \ldots, y_n, x_1, \ldots, x_n|\theta)$.

So, if given a problem/model statement (or equivalently a 2x2 table), we should be able to provide the likelihood proper and the complete likelihood. If given latent random variables, then the complete likelihood is the simpler of the two. However, if we have latent random variables, we may need to integrate out the latent random variables to get the proper likelihood (i.e., the probability of our observed random variables given our constants).

One thing to note is that in probability theory, mathematicians often make a distinction between a random variable and the realization of a random variable; the random variable itself has variability while the realization is set and fixed. If we said the random variable X_1 was realized as x_1, which we write as $X_1 = x_1$, then x_1 is fixed. Because x_1 does not change, it is technically a constant. However, all realizations are always constant because they occurred in the past and are unchangeable. So, including both X_1 and x_1 in the 2x2 table is a little redundant because if X_1 is classified as an observed random variable, then we know its realizations are what is being observed and realizations are always constants; so we know their place in the 2x2 table because of X_1's location. Similarly, if we classified X_1 as an unobserved random variable, then we know its realizations are what is being unobserved; because realizations are constants, we know the realization x_1's place just through X_1's role. We can get rid of this redundancy by considering only the random variables instead of realizations in the 2x2 table.

Example 8. Reference Example 2 for the explicit problem and model statements. Here is the 2x2 table:

	Observed	**Unobserved**
Variable	$X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n$	NA
Constant	σ^2	θ

We have observed random variables, $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n$, and an unknown constant, θ. So,

$$\text{Proper likelihood} = \mathbb{P}(y_1 ... y_n, x_1 ... x_n | \theta).$$

Notice, there are no latent random variables, so the complete and proper likelihoods are equivalent.

Example 9. Refer to Example 4 for the explicit problem and model statements. Here is the 2x2 table:

	Observed	**Unobserved**
Variable	$y_1 ... y_n$	NA
Constant	$z_1 ... z_n$	θ, α

Notice, we have observed random variables $y_1 ... y_n$ and unobserved constants $z_1 ... z_n$ but no latent random variables. Very similar to Example 8, the proper and complete likelihoods are equivalent:

$$\text{Proper likelihood} = \text{complete likelihood} = \mathbb{P}(y_1 ... y_n | z_1 ... z_n).$$

Example 10. Refer to Example 6 for the explicit problem and model statements. Here is the 2x2 table:

	Observed	**Unobserved**
Variable	$y_1 ... y_n$	$z_1 ... z_n$
Constant	NA	θ, α

We now have observed random variables $y_1 ... y_n$, latent random variables $z_1 ... z_n$, and no observed constants. The proper likelihood will still be $\mathbb{P}(y_1 ... y_n | \theta, \alpha)$, but its calculation is somewhat difficult because we must account for . So, what we can do is find the *complete likelihood* and integrate out the latent variables:

$$\textit{Complete likelihood} = \mathbb{P}(y_1 ... y_n, z_1 ... z_n | \alpha, \theta).$$

Now, can get the likelihood proper from the complete likelihood:

$$\begin{aligned} \textit{likelihood} &= \mathbb{P}(y_1 ... y_n | \alpha, \theta) \\ &= \int \mathbb{P}(y_1 ... y_n, x_1 ... x_n | \alpha, \theta) dx_i \\ &= \int \text{complete likelihood } dx_i. \end{aligned}$$

Example 11. Let's find the likelihood for the following, very simple example:

$$y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2).$$

Then,

$$\begin{aligned} likelihood &= \prod_{i=1}^{n} \mathbb{P}(y_i|\mu, \sigma^2) \\ &= \prod_{i=1}^{n} \mathcal{N}(y_i|\mu, \sigma^2). \end{aligned}$$

We assume y_i to be observed random variables, but μ and σ are unknown constants.

We can just use the assumed distributions to find the likelihood, but rarely is real-world modeling ever so simple. We can identify a series of equations that are relevant to our solution, and we can rebuild the likelihood.