# 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 or the data generating process, 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.

**Definition 0.2: Data Generating Process DGP**

The **data generating process (DGP)** is an algorithmic representation of a probabilistic model.

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 a 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, $\tau$, 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 $\theta$ is a set of scalar quantities for the model. Our goal is to now estimate $\theta$ given $y_1...y_n, x_1...x_n$.

First, we know $\theta$ is unknown. We must ask another question: "is $\theta$ 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 $\theta$ comes from, we classify $\theta$ as a constant. The 2x2 table for this problem reads as follows:

| | **Observed** | **Unobserved** |
|---|---|---|
| **Variable** | NA | NA |
| **Constant** | $x_1...x_n, y_1...y_n$ | $\theta$ |

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 $\tau$ 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 $\theta$ is a set of scalar quantities for the model. Similar to the earlier example, our goal is to estimate $\theta$ given $y_1...y_n, x_1...x_n$.

Now, suppose we posit the following model assumptions:

$$x_i \sim Normal(0, \sigma^2)$$
$$y_i = \theta x_i$$

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

come from a probability distribution, $\theta$ 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: $\sigma^2$ 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 $\theta$ is a constant. So, $y_i$ is an *observed variable* also.

| | **Observed** | **Unobserved** |
|---|---|---|
| **Variable** | $x_1...x_n, y_1...y_n$ | NA |
| **Constant** | $\sigma^2$ | $\theta$ |

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** | $z_1...z_n$ | $\theta, \alpha$ |

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:

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

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, $\theta$–a parameter describing the distribution for $y_i$–is not given and has no theoretical probability distribution, so it is an unknown constant. Likewise, $\alpha$ 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 | $\theta, \alpha$ |

**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:

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

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 | $\theta, \alpha$ |

**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 0.3: 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 $\theta$–the proper likelihood is $\mathbb{P}$(observed random variables| unknown constants) $= \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n|\theta)$.

**Definition 0.4: 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 $\theta$–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** | $\sigma^2$ | $\theta$ |

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

$$\text{Proper likelihood} = \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n, X_1 = x_1, \ldots, X_n = x_n | \theta).$$

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** | $z_1 ... z_n$ | $\theta, \alpha$ |

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:

$$\text{Proper likelihood} = \text{complete likelihood} = \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n | Z_1 = z_1, \ldots, Z_n = z_n).$$

**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 | $\theta, \alpha$ |

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 $\mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n | \theta, \alpha)$, 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:

$$\textit{Complete likelihood} = \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n, Z_1 = z_1, \ldots, Z_n = z_n | \alpha, \theta).$$

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

$$\begin{aligned}\text{Likelihood} &= \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n | \alpha, \theta) \\ &= \int \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n, X_1 = x_1, \ldots, X_n = x_n | \alpha, \theta) dx_i \\ &= \int \text{complete likelihood } dx_i.\end{aligned}$$

**Example 11.** Let us find the likelihood for the following, very simple data generating process laid out in Algorithm 1. Assume that we are given $y_1, \ldots, y_n$.

**for** $i = 1, \ldots, n$ **do**
  $Y_i \overset{\text{iid}}{\sim} Normal(\mu, \sigma^2)$.
**end**

**Algorithm 1:** Data Generating Process: Example 11

Then, because $Y_1, \ldots, Y_n$ come from a probability distribution and $\mu$ and $\sigma$ have no explicitly stated probability distributions, the 2x2 table for this problem would be given by

| | **Observed** | **Unobserved** |
|---|---|---|
| **Variable** | $Y_1, \ldots, Y_n$ | |
| **Constant** | | $\mu, \sigma^2$ |

Then, the likelihood would be given by

$$\text{Likelihood} = \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n | \mu, \sigma^2).$$

Because we assume that $Y_1, \ldots, Y_n$ are IID, we can rewrite the joint likelihood of $Y_1, \ldots, Y_n$ as the product of individual likelihoods:

$$\begin{aligned}\text{Likelihood} &= \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n | \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i | \mu, \sigma^2) \\ &= \prod_{i=1}^{n} Normal(Y_i = y_i | \mu, \sigma^2).\end{aligned}$$

Recall that the functional form of a Normally-distributed random variable, $Y_i$, with mean $\mu$ and variance $\sigma^2$ is given by

$$f_Y(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}}.$$

So,

$$\begin{aligned}\text{Likelihood} &= \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n | \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i | \mu, \sigma^2) \\ &= \prod_{i=1}^{n} Normal(Y_i = y_i | \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}.\end{aligned}$$

In modeling situations, we often see that there is a different process for one set of the population in comparison to another set of the population; for example, we can theorize that the distribution of house prices in Boston may be very different than the distribution of house prices in the Philadelphia suburbs. Because we have different processes for each subpopulation that we want to combine – or mix – into one process for the overall population, we refer to such models as mixture models.

```
ggplot(data = data.frame(x = c(-50, 100)), aes(x)) +
  stat_function(fun = dnorm, n = 101, args = list(mean = 50, sd = 10)) +
    stat_function(fun = dnorm, n = 101, args = list(mean = 10, sd = 20), linetype = 'dashed') +
  scale_y_continuous(breaks = NULL) +
  labs(x = NULL, y = NULL) +
  theme_bw()
```

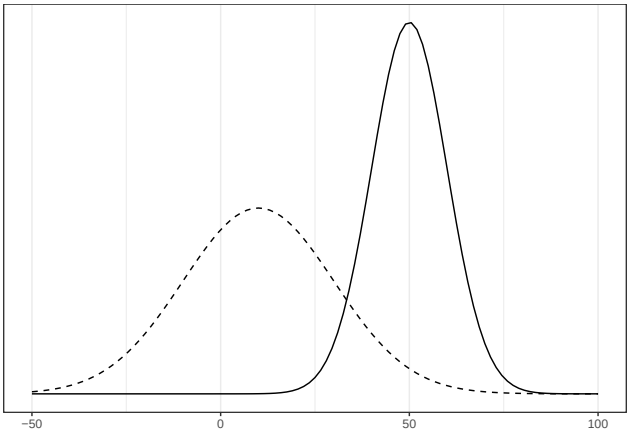


Figure 1: Mixture model example: dotted line represents distribution for one subset of the population, while solid represents the distribution for a different subset of the population.

**Definition 0.5: Mixture model**

A **mixture model** is a probabilistic model in which different components of our population follow different data generating processes.

Let us consider a few examples and illustrate how we would derive the likelihood for such a problem.

**Example 12.** Suppose we have a random sample of data from 43 students in a school, 20 of which are

graduate students and 23 of which are undergraduate students. We want to help administrators redesign curricula, for which they want to know how many courses the advisors should recommend that students take per semester.s

We assume that a student $i$ is a graduate student with probability $p$ or an undergraduate with probability $1-p$, which we represent by $Z_i$. Undergraduates are given 7 course offerings, while graduates are given 3 course offerings. Students can take all or none of the courses offered to them. We assume that a student decides to take each course independent of their decision to take another class; and the probability of taking a single class is given by $\theta_g$ for graduate students and $\theta_u$ for undergraduate students. The number of courses student $i$ takes is given by $Y_i$. Putting these together, we posit the following data generating process:

**for** $\underline{i = 1, \ldots, n}$ **do**

$$Z_i \sim Bernoulli(p)\ Y_i|Z_i = z_i \sim \begin{cases} Binomial(3, \theta_g), & z_i = 1 \\ Binomial(7, \theta_u), & z_i = 0 \end{cases}$$

**end**

**Algorithm 2:** DGP for Student Enrollments

We are given data on how many courses each student takes. However, we do not observe whether student is classified as a graduate or undergraduate student. Let us find the likelihood for this problem.

As usual, we will first begin by classifying our quantities in our 2x2 table. We know that because $Z_i, Y_i$ both come from a distribution, they are random variables. All other quantities – $p, \theta_g, \theta_u$ – are constants.

Additionally, because the only data we are given is $Y_1, \ldots, Y_{43}$, everything else is unobserved. So, our 2x2 would look as follows:

| | **Observed** | **Unobserved** |
|---|---|---|
| **Variable** | $Y_1, \ldots, Y_{43}$ | |
| **Constant** | $Z_1, \ldots, Z_{43}$ | $p, \theta_g, \theta_u$ |

Now, we can start calculating proper likelihood, which is the probability of realizing our observed data given the model's unknown constants:

$$\text{Likelihood} = \mathbb{P}(Y_1 = y_1, \ldots, Y_{43} = y_{43}|p, \theta_g, \theta_u).$$

However, here we run into a problem: the probability of realizing $Y_i$ depends on $Z_i$'s outcome, which is nowhere in our likelihood. To rectify this issue, we can take advantage of the idea of marginalizing distributions: for two sets $A$ and $B$, we can find the probability of a single set by integrating/summing "out" the other set: $\mathbb{P}(A) = \int_B \mathbb{P}(A, B)db$. In this case, we need to introduce $Z_1, \ldots, Z_{43}$; so, we can simply bring in this information by including the integrals around the complete likelihood:

$$\begin{aligned} \text{Likelihood} &= \mathbb{P}(Y_1 = y_1, \ldots, Y_{43} = y_{43}|p, \theta_g, \theta_u) \\ &= \int_{Z_1} \ldots \int_{Z_{43}} \mathbb{P}(Y_1 = y_1, \ldots, Y_{43} = y_{43}, Z_1 = z_1, \ldots, Z_{43} = z_{43}|p, \theta_g, \theta_u)dz_1 \ldots dz_{43}. \end{aligned}$$

Now, we can take advantage of the facts that all the $Z_i$'s are independent of one another and that all the $Y_i$'s

are independent of one another. So,

$$\begin{aligned}
\text{Likelihood} &= \mathbb{P}(Y_1 = y_1, \ldots, Y_{43} = y_{43} | p, \theta_g, \theta_u) \\
&= \int_{Z_1} \ldots \int_{Z_{43}} \mathbb{P}(Y_1 = y_1, \ldots, Y_{43} = y_{43}, Z_1 = z_1, \ldots, Z_{43} = z_{43} | p, \theta_g, \theta_u) dz_1 \ldots dz_{43} \\
&= \int_{Z_1} \ldots \int_{Z_{43}} \mathbb{P}(Y_1 = y_1, Z_1 = z_1 | p, \theta_g, \theta_u) \cdot \ldots \cdot \mathbb{P}(Y_{43} = y_{43}, Z_{43} = z_{43} | p, \theta_g, \theta_u) dz_1 \ldots dz_{43} \\
&= \int_{Z_1} \ldots \int_{Z_{43}} \prod_{i=1}^{43} \mathbb{P}(Y_i = y_i, Z_i = z_i | p, \theta_g, \theta_u) dz_1 \ldots dz_{43}.
\end{aligned}$$

Now comes another complication: we do not know the joint probability of $Y_i$ and $Z_i$. By taking advantage of Baye's Rule, we know that

$$\mathbb{P}(Y_i = y_i, Z_i = z_i) = \mathbb{P}(Y_i = y_i | Z_i = z_i) \mathbb{P}(Z_i = z_i).$$

Notice, we do know the quantities on the right hand side. From the data generating process,

$$\mathbb{P}(Y_i = y_i | Z_i = z_i) = \begin{cases} Binomial(3, \theta_g), & z_i = 1 \\ Binomial(7, \theta_u), & z_i = 0 \end{cases},$$

and

$$\mathbb{P}(Z_i = z_i) = Bernoulli(p).$$

We also know the functional form of Binomial and Bernoulli random variables, so

$$\mathbb{P}(Y_i = y_i | Z_i = z_i) = \begin{cases} \frac{3!}{(3-y_i)!(y_i)!} (\theta_g)^{y_i} (1-\theta_g)^{3-y_i}, & z_i = 1 \\ \frac{7!}{(7-y_i)!(y_i)!} (\theta_u)^{y_i} (1-\theta_u)^{7-y_i}, & z_i = 0 \end{cases},$$

and

$$\mathbb{P}(Z_i = z_i) = p^{z_i} (1-p)^{(}1 - z_i).$$

Now, let us simplify the problem and combine these different parts into the likelihood calculation for a single student $i$:

$$\text{Likelihood} = \int_{Z_i} \mathbb{P}(Y_i = y_i | Z_i = z_i, \theta_g, \theta_u, p) \mathbb{P}(Z_i = z_i) dz_i.$$

We know that if $Z_i = 1$, then $Y_i \sim Binomial(3, \theta_g)$, so

$$\begin{aligned}\text{Likelihood}|Z_i = 1 &= \int_{Z_i} \mathbb{P}(Y_i = y_i|Z_i = 1, \theta_g, \theta_u, p)\mathbb{P}(Z_i = 1)dz_i \\ &= \int_{Z_i} \mathbb{P}(Y_i = y_i|Z_i = 1, \theta_g, \theta_u, p)\mathbb{P}(Z_i = 1)dz_i \\ &= \int_{Z_i} \frac{3!}{(3-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{3-y_i}pdz_i.\end{aligned}$$

And similarly, we know that if $Z_i = 0$, then $Y_i \sim Binomial(7, \theta_u)$, so

$$\begin{aligned}\text{Likelihood}|Z_i = 0 &= \int_{Z_i} \mathbb{P}(Y_i = y_i|Z_i = 0, \theta_g, \theta_u, p)\mathbb{P}(Z_i = 0)dz_i \\ &= \int_{Z_i} \mathbb{P}(Y_i = y_i|Z_i = 0, \theta_g, \theta_u, p)\mathbb{P}(Z_i = 0)dz_i \\ &= \int_{Z_i} \frac{7!}{(7-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{7-y_i}(1-p)dz_i.\end{aligned}$$

However, we need to be able to consider the likelihoods for both cases together. So, we will take advantage of the facts that $x^0 = 1$ and $x^1 = x$ for any real number $x$. Then, we can combine these two cases together in the following way:

$$\begin{aligned}\text{Likelihood} &= \begin{cases} \int_{Z_i} \frac{3!}{(3-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{3-y_i}pdz_i, & z_i = 1 \\ \int_{Z_i} \frac{7!}{(7-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{7-y_i}(1-p)dz_i, & z_i = 0 \end{cases} \\ &= \left[\int_{Z_i} \frac{3!}{(3-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{3-y_i}pdz_i\right]^{z_i} \left[\int_{Z_i} \frac{7!}{(7-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{7-y_i}(1-p)dz_i\right]^{1-z_i}.\end{aligned}$$

Notice, when $Z_i = 1$, then

$$\begin{aligned}&\left[\int_{Z_i} \frac{3!}{(3-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{3-y_i}pdz_i\right]^{z_i} \left[\int_{Z_i} \frac{7!}{(7-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{7-y_i}(1-p)dz_i\right]^{1-z_i} \\ &= \left[\int_{Z_i} \frac{3!}{(3-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{3-y_i}pdz_i\right]^{1} \left[\int_{Z_i} \frac{7!}{(7-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{7-y_i}(1-p)dz_i\right]^{1-1} \\ &= \left[\int_{Z_i} \frac{3!}{(3-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{3-y_i}pdz_i\right] \left[\int_{Z_i} \frac{7!}{(7-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{7-y_i}(1-p)dz_i\right]^{0} \\ &= \left[\int_{Z_i} \frac{3!}{(3-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{3-y_i}pdz_i\right] \cdot 1 \\ &= \int_{Z_i} \frac{3!}{(3-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{3-y_i}pdz_i \\ &= \text{Likelihood}|Z_i = 1.\end{aligned}$$

And when $Z_i = 0$, then

$$\begin{aligned}
&\left[\int_{Z_i} \frac{3!}{(3-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{3-y_i}p dz_i\right]^{z_i}\left[\int_{Z_i} \frac{7!}{(7-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{7-y_i}(1-p) dz_i\right]^{1-z_i}\\
&= \left[\int_{Z_i} \frac{3!}{(3-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{3-y_i}p dz_i\right]^{0}\left[\int_{Z_i} \frac{7!}{(7-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{7-y_i}(1-p) dz_i\right]^{1-0}\\
&= 1 \cdot \left[\int_{Z_i} \frac{7!}{(7-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{7-y_i}(1-p) dz_i\right]\\
&= \int_{Z_i} \frac{7!}{(7-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{7-y_i}(1-p) dz_i\\
&= \text{Likelihood}|Z_i = 0.
\end{aligned}$$

So, we have found a way to combine the likelihoods of each case into one, non-piecewise form, that is also differentiable.

Now, we can put all the pieces together to find the joint likelihood of all 43 observations:

$$\begin{aligned}
\text{Likelihood} &= \mathbb{P}(Y_1 = y_1, \ldots, Y_{43} = y_{43}|p, \theta_g, \theta_u)\\
&= \int_{Z_1} \cdots \int_{Z_{43}} \left[\frac{3!}{(3-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{3-y_i}p\right]^{z_i}\left[\frac{7!}{(7-y_i)!(y_i)!}(\theta_g)^{y_i}(1-\theta_g)^{7-y_i}(1-p)\right]^{1-z_i} dz_1 \ldots dz_{43}.
\end{aligned}$$

## Using R for Probabilistic Modeling

Applied statistics and data science often deals with simulations to create synthetic data. We then use the simulated data to validate methods with clean data before working with real, messy data; or we use them to verify analytical solutions; or we use them to run experiments in a perfectly controlled world before implementing them in practice. In short, simulations are essential to the study of data science.

To actually put simulations into practice, we take advantage of the algorithmic construction of the data generating process of a model and translate the process to R and work with the data there. Let us consider a few use-cases of some simulations.

**Example 13.** Suppose we are working for a clothing manufacturing company and want to help them re-index their sizing charts. They have three sizes: small, medium, and large. They want about thirty-three percent of the population to lie within each of the sizes. The company has a sample of heights for 50 people in inches: `(58, 87, 64, 65, 48, 79, 77, 46, 50, 69, 35, 101, 77, 78, 88, 85, 71, 53, 77, 52, 103, 63, 74, 27, 25, 82, 42, 74, 52, 77, 84, 43, 49, 58, 91, 20, 101, 100, 46, 94, 86, 69, 83, 78, 35, 85, 43, 71, 34, 25)`

Using this information, we will need to identify at what heights the company should re-index the sizing charts.

**for** $i = 1, \ldots, 50$ **do**
  $Y_i \sim Normal(\mu, \sigma^2)$
**end**

**Algorithm 3:** DGP for Example 13

We are also given the data generating process laid out in Algorithm ??.

While we will more formally learn how to go through this process completely in later sections of this book, let us try to simulate some data and plot the likelihood for the given data and some possible values of $\mu$ and $\sigma^2$.

First, we need to write the 2x2 table and the likelihood for this problem. The only quantities we have are $Y_1, \ldots, Y_{50}$, $\mu$, and $\sigma^2$. Since we only observe $Y_1, \ldots, Y_{50}$, $Y_1, \ldots, Y_{50}$ are our only observed variables and $\mu, \sigma^2$ are unobserved constants:

| | **Observed** | **Unobserved** |
|---|---|---|
| **Variable** | $Y_1, \ldots, Y_{50}$ | |
| **Constant** | | $\mu, \sigma^2$ |

Notice, this 2x2 table and DGP is identical to the problem outlined in Example 11. So, we can reuse the likelihood we calculated there:

$$\text{Likelihood} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}.$$

Because we want to plot the likelihood of the given data, we need to calculate the likelihood itself, which we can do using `R`:

```
likelihood_calc <- function(data, mu, sigma) {
  # parameters:
    # data: a list of y_1, ..., y_50
    # mu: potential mean
    # sigma: potential standard deviation
  # returns:
    # numeric representing likelihood(mu, sigma | data)
  # initialize likelihood
  joint_likelihood <- 1
  for(y_i in data) {
    # calculate likelihood for individual i
    lik_i <- 1/(sqrt(2 * pi * sigma^2))*exp(-(1/(2 * sigma^2))*(y_i - mu)^2)
    # augment joint likelihood by multiplying by individual i's likelihood
    joint_likelihood <- lik_i * joint_likelihood
  }

  return(joint_likelihood)
}
```

Now, we can calculate the likelihood of the given data for a series of possible means and variances:

```
# save data
y_data <- c(58, 87, 64, 65, 48, 79, 77, 46, 50, 69, 35, 101, 77, 78, 88, 85, 71, 53, 77, 52, 103, 63, 74
25, 82, 42, 74, 52, 77, 84, 43, 49, 58, 91, 20, 101, 100, 46, 94, 86, 69, 83, 78, 35, 85, 43, 71,
34, 25)
```

```r
# potential values of mu and sigma
mu_list <- c(10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
sigma_list <- c(10, 20, 30, 40, 50, 60, 70, 80, 90, 100)

likelihoods <- c()
tried_mu <- c()
tried_sigma <- c()
for(mu in mu_list) {
  for(sigma in sigma_list) {
    tried_mu <- c(tried_mu, mu)
    tried_sigma <- c(tried_sigma, sigma)
    likelihoods <- c(likelihoods, likelihood_calc(data = y_data,
                                                  mu = mu,
                                                  sigma = sigma))
  }
}
```

Now, we can store this data in a dataframe and plot the data using ggplot2:

```r
# make the data frame
likelihood_data <- data.frame(mu = tried_mu,
                              sigma = tried_sigma,
                              likelihood = likelihoods)
# plot the data
library(ggplot2)
ggplot(likelihood_data) +
  geom_point(aes(x = mu, y = sigma, color = likelihood), size = 7) +
  labs(x = 'Potential mu',
       y = 'Potential sigma',
       color = 'Likelihood')
```

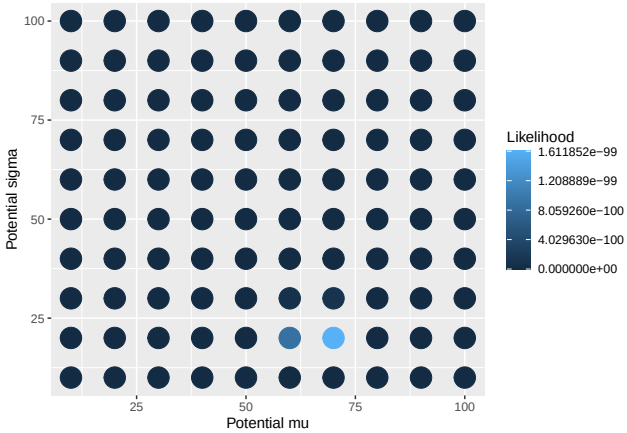


Figure 2: Likelihood of Potential mean and standard deviations

Figure 2 displays the potential values of $\mu$ on the x-axis and the potential values of $\sigma$ on the y-axis with each of the points colored by the likelihood; the higher the likelihood, the brighter the point's color. Recall, the likelihood quantifies the "fit" of the model's parameters taking on specific values; so, the higher the likelihood for specific parameter values, the greater the chances that the true model parameters are those values. Because we see the brightest blue point at (70,20), we can "guess" that the model's parameters are $\mu \approx 70$ and $\sigma^2 \approx 20^2 = 40$.

**Example 14.** Now, let us consider using `R` to study a mixture model case. Refer to Example 12 for the actual setup. In the problem, we want to estimate $\theta_u$ and $\theta_g$. A large complication in this problem is that we do not know the true values of $Z_1, \ldots, Z_{43}$: in other words, we do not know which students are graduate students and which are undergraduate.

A colleague suggests that to overcome this problem, we can assume $p = \frac{20}{43}$ (so, $p$ is now a known constant); and then we can randomly select 20 of our 43 students and claim they are graduate students. Now that we have "data" on who is a graduate student and who is not, we can simply calculate the sample success probability of the randomly classified graduate students and undergraduate students each. Doing so allows us to find $\hat{\theta}_u$ and $\hat{\theta}_g$: estimates of our unknown constants. To see if this approach works or not, we will run some simulations.

First, we will translate the DGP to `R` and simulate data for $\theta_u = 0.4$ and $\theta_g = 0.6$. Then, we will try to recover $\theta_u$ and $\theta_g$ by calculating the sample success probabilities.

In order to simulate data, we will translate the likelihood to `R` code. For this, we need to know how to draw samples from the Bernoulli distribution and the Binomial distribution, which we can accomplish using the function `rbinom`. `rbinom` takes three arguments:

1. `n`: the number of observations we want from the Binomial distribution
2. `size`: the number of independent Bernoulli trials
3. `prob`: the probability of success in each of the independent Bernoulli trials

Recall that the $Binomial(N, p)$ distribution represents the number of successes in $N$ independent Bernoulli trials. So, when $N = 1$, the Binomial variable only has one Bernoulli trial. So, $Bernoulli(p) = Binomial(1, p)$. Then, that means we can use `rbinom` with `size = 1` for our Bernoulli distribution.

So, we can translate $Z_1, \ldots, Z_{43} \overset{\text{iid}}{\sim} Bernoulli(\frac{20}{43})$ into `rbinom(n = 43, size = 1, prob = 20/43)`. And we can convert $Y_i \sim Binomial(7, \theta_u)$ into \texttt{rbinom(n = 1, size = 7, prob = theta_u)} for the code argument \texttt{theta_g}. Lastly, we can express $Y_i \sim Binomial(3, \theta_g)$ as \texttt{rbinom(n = 1, size = 3, prob = theta_g)}. So, our `R` representation of the DGP would be

```
# setting the seed ensures we get the same randomness to replicate results
set.seed(999)
dgp <- function(theta_u, theta_g) {
  # parameters:
    # theta_u: success probability of undergraduates taking a class
    # theta_g: success probability of graduates taking a class
  # returns:
    # data frame with four columns:
      # 1. values of z_1, ..., z_43
```

```
      # 2. values of y_1, ..., y_43
      # 3. value of theta_u
      # 4. value of theta_g
  # simulate 43 observations
  # first randomly assign each student's class: graduate (z_i = 1) or undergraduate (z_i = 0)
  z_vector <- rbinom(n = 43, size = 1, prob = 20/43)

  # based on each student's classification, sample the number of courses they take
  y_vector <- c()
  for(z_i in z_vector) {
    if(z_i == 1) y_i <- rbinom(n = 1, size = 3, prob = theta_g)
    if(z_i == 0) y_i <- rbinom(n = 1, size = 7, prob = theta_u)

    y_vector <- c(y_vector, y_i)
  }

  # make a data frame for Z_i, Y_i, and its corresponding parameter inputs
  return(data.frame(Z = z_vector,
                    Y = y_vector,
                    theta_u = theta_u,
                    theta_g = theta_g))
}
```

Notice that we return a data frame with four columns even though we only need data on $Y_1, \ldots, Y_{43}$. Including this extra information does not hurt our program and can help us account for other situations; for example, we can check how well using the sample success probabilities as estimates are when we know $Z_1, \ldots, Z_{43}$.

Now, let us write a function that randomly assigns each student in our observed data set to the graduate class with probability $\frac{20}{43}$:

```
estimate_thetas <- function(y_data) {
  # parameters:
    # y_data: a list of 43 elements representing the number of courses each of the 43 students take
  # returns:
    # estimates for theta_u and theta_g

  # classify each student as graduate (1) with probability 23/43
  z_vector <- rbinom(43, 1, 23/43)

  # calculate sample success probability for all students
  theta_g_hat <- 0
  theta_u_hat <- 0

  n_grads <- 0
```

```
  n_ugrad <- 0
  for(i in 1:43) {
    if(z_vector[i] == 1) {
      theta_g_hat <- theta_g_hat + y_data[i]/3
      n_grads <- n_grads + 1
    } else {
      theta_u_hat <- theta_u_hat + y_data[i]/7
      n_ugrad <- n_ugrad + 1
    }
  }

  theta_g_hat <- theta_g_hat/n_grads
  theta_u_hat <- theta_u_hat/n_ugrad
  return(list(theta_g_hat = theta_g_hat, theta_u_hat = theta_u_hat))
}
```

In order to validate this method, we will simulate data 1,000 times and run through this estimation strategy for each of the simulated times. By re-simulating data and re-estimating the unknown constants, we are able to understand how well this strategy works; when we simulate probabilistic data, there will be some data that is very representative data and also unrepresentative data that has realizations that take on extremes in the distribution. For example, it is possible to simulate data in which $Y_1 = \ldots = Y_{43} = 0$. But if $\theta_u$ and $\theta_g$ are much greater than 0, there is still a possibility of observing such data. So, we run the simulation a number of times so that we can ensure that our discoveries are not dependent on a single, potentially edge case.

For each set of simulated data, we will have $\hat{\theta}_u$ and $\hat{\theta}_g$ given by the sample success probability. The sample success probability is simply the number of successful trials (i.e., $y_i$) divided by the number of total trials in a Binomially distributed random variable.

Now, because we will have a total of 1,000 values, we can then build an empirical distribution for $\hat{\theta}_u$ and $\hat{\theta}_g$ and see if the estimated and actual values are close or not. We will specifically try this for $\theta_u = 0.4$ and $\theta_g = 0.6$.

From Figure 3, we see the relationship between the estimates for $\hat{\theta}_u$ and $\theta_u$, as well as the relationship between $\hat{\theta}_g$ and $\hat{\theta}_u$. Overall, we see the distribution is approximately centered around 0.35 for $\hat{\theta}_u$, which may or may not be a poor estimate. However, $\hat{theta}_g$ is centered around 0.8, which is a wide deviation from 0.6. While we will discuss how to evaluate the strength of an estimation strategy later, this basic framework of comparing the estimates to the actual outcome is a useful exercise that illustrates R's utility in probabilistic modeling.

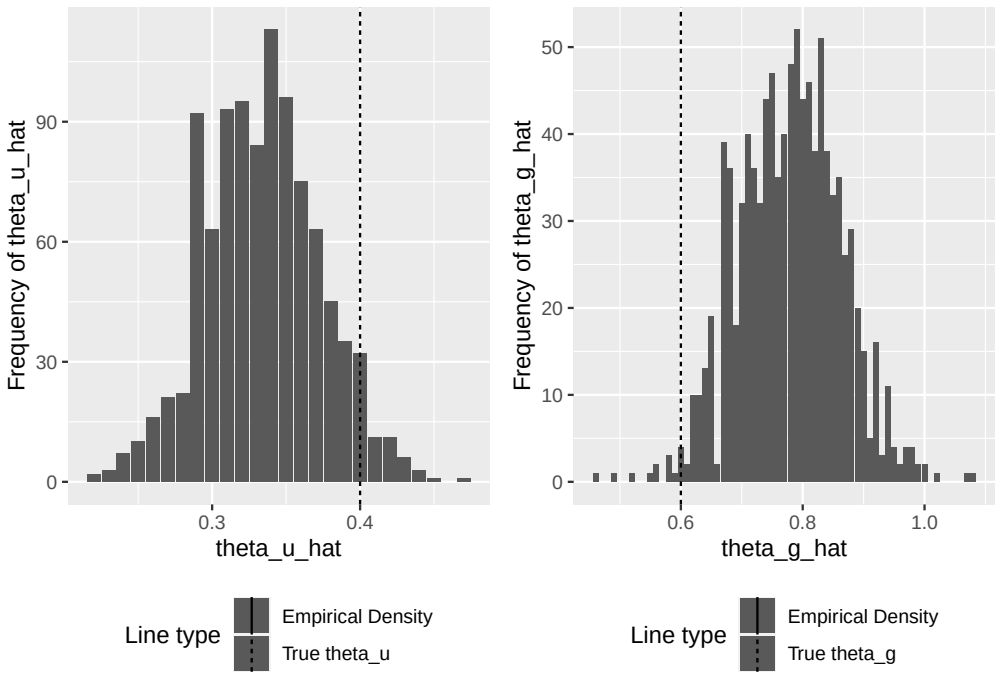


Figure 3: Simulation results for the proposed estimation strategy. Each figure compares the empirical distribution of the statistic estimate to the actual value of the statistic (dashed line). The closer the dashed line is to the "mode" of the empirical distribution, the better the proposed strategy for estmating the unknown constants.