STAT 3503/8109 Lecture 2 Notes

Edoardo Airoldi

Scribe: Srikar Katta[*]

Fall 2020: August 31, 2020

1 Introduction

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 graphical models. Using, this information, we can construct the data science pipeline:

1. Data + model $\implies$ *find* likelihood function
2. Maximize likelihood to find unknown quantities
3. Use best guess for unknown quantities to estimate our objective

2 Language of Modeling

2.1 Variables

Even between literature and professionals, there may exist a disconnect in the statistical/data science/machine learning language, particularly with the use of buzzwords. For example,

[*]Please share any comments or suggestions with Srikar Katta at srikar@temple.edu

parameter, a quite popular "buzzword," can be associated with a typically unobserved quantity that lacks variability (we will learn that we can classify this as an unobserved constant). However, from a Bayesian perspective, a *parameter* is a variable with unknown variability (we will learn that this is an unobserved latent variable). Other words like this include fixed effects, random effects, causal effects, and mixture weights - they all change from discipline to discipline. As data scientists, it is our responsibility to cross over this language barrier and ensure that we can perrform inference properly.

Due to this difficulty in communication, we must first classify our variables, for which we have four groups:

1. Known constants
2. Unknown constants
3. "Latent" variables (sometimes referred to as omitted variables)
4. Observations/data/random variables

2.2 Probability Distributions

Other relevant language of data science includes probability distributions. Two elements are going to help us classify any two quantities into our groups:

1. Problem setting: what are we hoping to estimate?
2. Assumptions/"model" that you think is reasonable

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

	Observed	**Unobserved**
Variable		
Constant		

Any problem in the world is typically is meant to estimate some quantity, call it τ, and we will need to have some notation to describe this quantity. Given a bunch of other quantities, I must estimate τ. The assumptions/model from our problem statement is going to tell us

if something is variable or not (i.e., what follows a distribution). Using this information, we will estimate our quantities.

Example 2.1. *Suppose we want to estimate some quantity, $\hat{\tau}(y_1...y_n, x_1...x_n, \theta)$, 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. Goal: estimate θ given $y_1...y_n, x_1...x_n$.*

First, we know θ is unknown since that is what we want to quantify. We must ask another question, "is θ an unknown constant or an unknown variable?"

Notice, we got our problem statement from the example. What we also now need is to come up with our set of assumptions/model. Suppose the model to be as simple can be:

$$y_i = \theta x_i, i = 1...n$$

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 you don't make any assumptions about variables having a distribution. Obviously, θ is unobserved and is not variable because we have not made any assumptions about x_i and y_i being random variables that might suggest that θ itself has a distribution.

	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 2.2. *Suppose we want to estimate some quantity, $\hat{\tau}(y_1...y_n, x_1...x_n, \theta, \sigma^2)$, 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. Goal: estimate θ given $y_1...y_n, x_1...x_n$.*

What do we want to estimate?

$$\hat{\tau}(y_1...y_n, x_1...x_n, \theta)$$

Now, let's say that we agree on the following model assumptions:

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

While most is the sample as example 2.1, we also know that x_i is a random quantity that has a normal distribution with population variance σ^2 , the distinction from the earlier problem. Let us classify our variables as before. We know that θ is still an unknown variable. It is *still* a constant because we don't make any assumptions about its distributions.

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. This is really an observed value, and we don't make any assumptions about its distribution. So, it is an observed constant. We can also include σ^2 as an input for τ.

Classifying our y_i is a little bit more complicated than before, but we will learn later on in the course 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

2.2.1 Theoretical vs Empirical Probability 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, a statistician/data scientist/researcher 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. This is a perfectly likely 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 2x2 framework. More information can be found in Appendix 4.1.

2.3 Different Problem and Model Statements with the Same Objective

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

Example 2.3.

Consider the following problem/model statement:

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

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

- **Example 2.3.1**: 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 2.3.1. *Consider the first problem/model statement from Example 2.3. You are given $y_1...y_n, z_1...z_n$, and your 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: $\mathbb{V}(y_i) = \mathbb{E}(\mathbb{V}(y_i|z_i)) + \mathbb{V}(\mathbb{E}(y_i|z_i))$. 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 2.3.2. *Consider the second problem/model statement from Example 2.3. You are given $y_1...y_n, z_1...z_n$. But now your 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 2.3.1, 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 it is now classified 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 2.3.3. *Consider the third problem/model statement from Example 2.3. You are given $y_1...y_n$, and your 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 2.3.2, 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 2.3.2 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 2.3.4. *Consider the first problem/model statement from Example 2.3. You are given $y_1...y_n$, and your 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 2.3.1, 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 2.3.1 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	$\theta, \alpha, z_1...z_n$

2.4 Comments on Quantity Classification

Classifying something as *known* or *unknown* depends only on whether something is given to you or not. Classifying something as *variable* or *constant* depends entirely on your assumptions of the model and the distributions of your quantity - not based on what you actually see in the data. When discussing inference, a researcher does not necessarily even need the problem and/or model statement though - this 2x2 table contains all the information necessary. However, nobody explicitly shows the table. They state the problem statement, possibly the model statement, the goals of the inference (i.e., what are they estimating?), and their methods. Ideally, any researcher can backtrack from a 2x2 table to identify the problem and model statements.

2.5 Standardizing Notation and Language for Statistical Modeling

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

3 Likelihood

A likelihood function is a function of the observed random variables, whatever they may be, given the constants that a researcher finds from the problem and model statements, or equivalently that 2x2 table.

1. Likelihood proper: $\mathbb{P}(\text{observed random variables(rv)}|\text{ unknown constants}) = \mathbb{P}(y_1...y_n|\theta) \iff$ probability of your observations and random variables, given observed constants
2. Complete likelihood: $\mathbb{P}(\text{observed rv, latent rv} \mid \text{unknown constants}) = \mathbb{P}(y_1...y_n, x_1...x_n|\theta) \iff$ Joint probability of latent/observed rvs, given observed constants

So, if given a problem/model statement (or equivalently a 2x2 table), a researcher 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 you have latent random variables, you may have to integrate out the latent rv to get the likelihood proper, which is the probability of your observations and random variables.

Example 3.1. *Reference Example 2.2 for the explicit problem and model statements. Here is the 2x2 table:*

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

We have observed random variables, $x_1...x_n, y_1...y_n$, and an unknown constant, θ. So,

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

Notice, there are no latent random variables, so there is no notion of complete likelihood.

Example 3.2. *Refer to Example 2.3.1 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 3.1, no complete likelihood exists. But we can find the likelihood proper:

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

Example 3.3. *Refer to Example 2.3.3 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 likelihood will still be $\mathbb{P}(y_1...y_n|\theta,\alpha)$. But calculating the likelihood proper is somewhat difficult. 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:

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

This will be covered more in later lectures.

Example 3.4. *Let's find the likelihood for the following, very simple example:*

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

Then,

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

We assume y_i to be observed random variables, but μ,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.

4 Appendix

4.1 Constant vs variable

In Example 2.2, we know what y_i and x_i are, even if x_i is distributed - so shouldn't we be able to "backtrack" and estimate θ? Suppose $x_i \sim \mathcal{N}(\mu, \sigma^2)$. We know that x_i is a random variable.

Consider the following two scenarios:

1. a 6-sided die with 6 on **all sides** $\implies$ $x_i = 6$ with probability 1
   - $\mathbb{E}(x_i) = 6, \mathbb{V}(x_i) = 0$
   - Flip x_i: 6, 6, 6, 6, 6, 6 ...
   - The empirical and theoretical distributions are the same **always**: expectation and variance of empirical distributions (i.e., what actually happens) are also 6 and 0 respectively
2. a 6-sided die with $1, 2, ..., 6,$, so each outcome has probability $\frac{1}{6}$, so $\mathbb{P}(y_i) = \frac{1}{6}$
   - $\mathbb{E}(y_i) = 3.5, \mathbb{V}(y_i) \geq 0$
   - Flip y_i: 1, 1, 1, 1, 1, 1 ...is *one possible sample*
   - Flip y_i: 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6 ...is also *one possible sample*
   - Theoretical and empirical distributions do not have to be the same, *necessarily*

The first example is a probabilistic representation of a **constant**. So, x_i is just one number. On the other hand, in the second situation, x_i is not guaranteed to be the same outcome each time *empirically*.

Example 4.1. *Problem statement: estimate the number of days that some user i made some purchase*
Model: $y_i|\theta$ *where* y_i *is independently and identically distributed (i.i.d.)* $Binomial(365, \theta)$

We know that y_i, given θ follows a Binomial, so what are $\mathbb{V}(\theta), \mathbb{V}(y_i)$

$$\mathbb{V}(y_i) = \theta(1-\theta) > 0 \implies y_i \textbf{ is variable}$$

But what if I give you a sample as follows: $y_1 = 5, y_2 = 5, y_3 = 5$ but the estimated variance is 0!! So the distinction between the empirical and theoretical variability comes from the sample outcome versus the expected outcome.

As an exercise, what is the two-by-two table? Notice, y_i comes from a Binomial, so, *theoretically* it is variable. Also, θ is given, and we make no assumptions on its distribution, so it is constant.

	Observed	**Unobserved**
Variable	y_1, y_2, y_3	NA
Constant	NA	θ

As another exercise, try finding the likelihood for this problem and model statement.

There are no latent random variables, so there is no complete likelihood. But the likelihood proper can be defined as follows:

$$likelihood = \mathbb{P}(y_1, y_2, y_3|\theta).$$