MODELS AND SIMULATION

ETHAN LEVIEN

Contents

1. Statistical models

A central concept to this course is that of a **model**. Broadly speaking, models are simplified representations of the world[1] There are many ways to represent models, but in science (and life), we often use **mathematical models**. For example, you might be familiar with Newton's equation:

$$F = ma \tag{1}$$

which related the force (F) acceleration (a) and mass (m) of a particle. This a mathematical model of the motion of a (non-relativistic) particle in a force field. While it isn't always true, it holds for such a wide range of applications that we call it a *fundamental* law of nature, and it turns out to be extremely powerful. We can combine Newton's equation with other fundamental lows build models of more complicated systems involving many particles. For example, to build a model of planetary motion, we can combine Newton's equation with his other law of gravity, which states that the gravitation force between two objects of masses m_1 and m_2 a distance r from each other is

$$F = G\frac{m_1 m_2}{r^2} \tag{2}$$

where G is a constant. We refer to these types of models – that is, those which are built upon fundamental laws of nature – as **mechanistic** models.

In statistics, we are often interested in problems where nothing remotely close to a fundamental laws exist (yet). Instead, they are based directly on observations (data) or our intuitions. We will call these **phenomenological models**. Such models still be quite useful provided we are aware of their limitations. Of course, there is not sharp distinction between mechanistic and phenomenological models, but the distinction is helpful nonetheless.

Often our models cannot make exact predictions about the value of a variable (this is true in both mechanistic and phenomenological models, but especially the latter). Instead, they only tell us the probability that a variable has a certain value, or falls within a range of values.

For example, if we were to construct a model of the (y) of the height a randomly selected pine tree in New Hampshire as a function of its age (x), we might begin by searching a relationship of the form

$$y = f(x) \tag{3}$$

Such a model might be relevant for conservation efforts, since it would be important to understand how trees develop over time and influence their surrounding environment. However, if we take a **sample** of the population, meaning we go out and measure the height of some trees for which we know the age, we will quickly find that trees of the same age can have different heights. This variation will be

Date: April 2022.

[1]Model is sometimes used interchangeably with **theory**, although I usually think of models as being smaller in scope.

as result of many variables which are not in our model, e.g. the surrounding, the specific subspecies of pine, genetic variation within a subspecies to name just a few. In principle, we could construct a more complex model which includes, say, the DNA sequence of the tree (g):

$$y = f(x, g). \tag{4}$$

This model would presumably have less variation. That is, if we collect a sample of trees for which we know the height and DNA sequence, we would find that the variation between trees with the same DNA and height is less than the variation between trees with only the same height. Yet variation will remain unless we include all the variables which effect tree height. However, it is not so useful to include these variables, since we can't actually measure most of them and many will have only a small influence on the tree height. One approach to dealing with these "hidden" sources of variation is to define a **statistical model**, where y is not the same for each tree of the same height, but is instead a **random variable**. On type of a statistical model is a regression model

$$y = f(x) + \epsilon \tag{5}$$

where ϵ is a random variable (usually one that is zero, on average). Models of this sort are the topic of this course.

2. Random variables and distributions

The rigorous mathematical theory for random variables is very useful, but requires certain machinery which is beyond the scope of these notes. Fortunately, we go a long way without such formalism. For our purposes, a random variable can be understood as a variable which we cannot predict prior to an observation, regardless of how much information we have. We can define the space of **outcomes** as all the possible values that a random variable may take on. The outcomes for the roll of a dice are $1, 2, \ldots, 6$ for the dice, or positive numbers for the height of a tree. Usually the outcomes are numbers, even if we use a number to represent a non-numerical quantity (e.g. someone's gender). In probability theory, one distinguishes between outcomes and **events** – the latter are subsets of outcomes. For example, we might refer to the event that the roll of a die is grater than 2. It's good to be aware of these definitions, but you don't need to memorize them.

We can describe a characterize a random variable using a **probability distribution**, which maps a set of possible events to real numbers between 0 and 1. For example, suppose we ask a random student in the college whether they were born in the US. A probability distribution $P(Y)$ which *models* their answer is the **Bernoulli distribution**,

$$P(Y) = \begin{cases} q & Y = \text{YES} \\ 1 - q & Y = \text{NO} \end{cases} \tag{6}$$

where q is the fraction of students in the college who were born in the US. More compactly, we can represent YES with 1 and NO with 0, and write

$$P(Y) = q^Y (1 - q)^{1-Y}. \tag{7}$$

We say that q is a **parameter** in our model because regardless of it's value our model is still a Bernoulli distribution. It is very important that the sum of $P(Y)$ over all possible outcomes is 1 – this is simply saying that we are certain one of the outcomes will happen. We will use $P(1)$ to mean "the probability that $Y = 1$", or, in there is some ambiguity in which variable we are referring to, we might write $P(Y = 1)$.

In this context, q has a clear interpretation: it is the fraction of students in the college who were born in the US. The Bernoulli distribution is our default model for any variable that can take two possible outcomes, usually abstracted as 0 or 1. In order to state that a Bernoulli distribution is a model for some random variable Y, we write

$$Y \sim \text{Bernoulli}(q). \tag{8}$$

We might also say "Y *follows* as Bernoulli distribution" or "Y is a Bernoulli random variable". More generally, we say that a variable in a model follows a given distribution by writing

$$\text{Variable} \sim \text{Distribution}(\text{parameters}). \tag{9}$$

We will sometimes use θ to denote the parameters.

Turning back to the example of our survey, let's suppose we don't have information about every students in the college. Rather, a survey of five students from this class is conducted, finding 4 yeses and 1 no. What is our best prediction of

the total fraction of students in the college who answered YES? What assumption do we make when we answer this question? The process of answering this question is **statistical inference**. More generally, we use statistical inference to make predictions about things we don't observe based one what we do observe (data).

3. Python as a tool for statistical modeling

When we generate samples using a computer we call them **simulations**. We will use python to perform simulations, and it is therefore important to have a basic understanding of the python language. It is assumed that you will go through the separate python tutorial notebook. For convenience, we will cover some basic tasks in this Notebook

Exercise 1: *Working with for loops*

3.1. Simulations.

Here, we will focus on tools relevant for statistics. In Python, we can simulate random variables using the `numpy` library:

```
> import numpy as np
> q = 0.5
> y = np.random.choice(range(2),p=[q,1-q])
```

We can generate multiple samples using a for loop

```
> n_samples = 100
> y = np.zeros(n_samples) # makes an empty list (i.e. array) of n_samples zeros.
> for k in np.range(n_samples):
>   y[k] = np.random.choice(range(2),p=[q,1-q])
```

A simpler way of doing this is

```
> y = np.random.choice(range(2),n_samples,p=[q,1-q])
```

The more general form of this command is

```
> y = np.random.choice(range(k),n_samples,p=[q_1,q_2,...,q_k])
```

where $q_1 + \cdots q_k = 1$. This will generate a sample from

Exercise 2: *Building a probability model*

We can also generate simulations of more complex random variables using simple ones. In this case, it is useful to define a function in Python which generates samples of our new random variable. For, example:

Example 1. *Writing a function to run simulations of coin flips.*

Exercise 3: *Modifying existing code*

3.2. Visualization.

An important tool for visualizing samples is a histogram. In python, we would write:

```
> plt.hist(samples,100,density=true)
```

The histogram shows us the frequency of different outcomes. Histograms are discussed here

3.3. Working with tabular data.

Frequently, we will work with data in tabular form. We can do this using Numpy (hopefully you read about this in the python tutorial), e.g.

```
>  # imagine we have an array of times and corresponding temperature measurements:
> times = np.array([1,2,3,4,5])
> temps = np.array([72,71,75,75,73])
> # we can make a 2d numpy array
> data = np.transpose(np.array([times,temps]))
> data
```

The pandas package in python provides some additional functionality:

```
> # the pandas library provides a convenient way organize this data
> import pandas as pd
> df = pd.DataFrame(data,columns = ["time","tempature"])
```

Examples from class can be found here here.

3.4. **Monte Carlo.** Often, we run many simulations of a model in order to say something about the distribution without performing any analytical calculations. We call these **Monte Carlo** simulations. Monte Carlo simulations make use of the fact that we can always conceptualize probabilities as fraction of things. That is, if we have n samples of a variable Y and we want to estimate $P(Y = y)$, then we can count the number for which $Y = y$ – we denote this as $n(Y = y)$, and divide by the total number: $P(Y = y) \approx n(Y = y)/n$.

Example 2. *Running Monte Carlo simulations*

Questions concerning how many samples we need to generate to obtain meaningful estimates from Monte Carlo simulations will be addressed later on.

3.5. **Means, variances, etc.** There are ways in which we summarize attributes of random variables. If we have many samples $Y_1, Y_2, \ldots, Y_n$ of a random variable Y (e.g. answers to a survey question), the **sample mean** is defined as

$$\bar{Y} = \frac{1}{n}\sum_i Y_i \tag{10}$$

Often it is useful to quantify the deviations from the mean. Suppose each Y_i can take on outcomes $Y = 1, 2, 3, \ldots, m$. If n is large, then the fraction of samples for which $Y_1 = y$ will be $P(Y_1 = y)$, thus the sample mean converges to the true mean:

$$\bar{y} = \frac{1}{n}\sum_{y=1}^{m} y n_i = \sum_{y=1}^{m} y\frac{n_i}{n} \approx \sum_{y=1}^{m} yP(Y = y) \tag{11}$$

Sometimes we write $P(Y_i = y_i)$ and sometimes we write The expression on the right is the definition of the mean, or **expectation**, often denoted $\mathbb{E}[Y]$.

For this, we have the **sample variance**

$$\sigma^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2. \tag{12}$$

We will see why this makes sense as a measure of how spread our a distribution is later on when we talk about inference. For large n, this converges to the square root of the variance

$$\mathrm{Var}(Y) = \sum (y_i - \mathbb{E}[Y])^2 P(Y_i = y_i). \tag{13}$$

We are often interested in the standard deviation

$$\sigma = \sqrt{\frac{1}{n-1}\sum_i (y_i - \bar{y})^2}. \tag{14}$$

In python, functions for implementing the mean and standard deviation are follows:

```
> np.mean(y)
> np.std(y)
```

By default, the standard deviation function divides the sum by the number of samples, we can fix this

```
> np.std(y,ddof=1)
```

Example 3. *Verifying an analytical formula with simulations*

Exercise 4: *Verifying an analytical formula with simulations*

We can always calculate the mean and standard deviation and mean of a sample regardless of the distribution it has been drawn from. However, we need to be careful, as the results may not be so meaningful. For example, the mean of a Bernoulli random variable is q, but (unless $q = 0$ or $q = 1$), the variable will never actually take on this value. For example, it might be more meaningful to think of the **mode**, which is the value that occurs most frequently.

4. Joint probabilities, Independence

We introduce, very briefly, the concepts of independence and conditioning. Just as we have considered single random variables, we can consider multiple random variables within the same model. Suppose we have two Bernoulli random variables

Y_A and Y_B which model whether a person has mutations at two different locations on their genome. In this case, we need a model of both variables together:

$$\mathbb{P}(Y_A, Y_B) = \begin{cases} q_{00} & \text{if } Y_A = 0 \text{ and } Y_B = 0 \\ q_{01} & \text{if } Y_A = 0 \text{ and } Y_B = 1 \\ q_{10} & \text{if } Y_A = 1 \text{ and } Y_B = 0 \\ q_{11} & \text{if } Y_A = 1 \text{ and } Y_B = 1 \end{cases} \tag{15}$$

The probability distribution $P(Y_A, Y_B)$ tells us the probabilities for observing *both* variables together, e.g. observing a person with both mutations. It does not directly tell us the probabilities of observing e.g. someone with only one mutation. This can be obtained via marginalization; that is, summing over the other variable:

$$P(Y_A) = \sum_y P(Y_A, y) = P(Y_A, Y_B = 0) + P(Y_A, Y_A = 1) \tag{16}$$

where in the general the sum is taken over all possible outcomes for the second variable. $\mathbb{P}(Y_1)$ is defined similarly. For example,

$$P(Y_A = 1) = q_{10} + q_{11}. \tag{17}$$

This means that

$$Y_A \sim \text{Bernoulli}(q_{10} + q_{11}). \tag{18}$$

This is the distribution of Y_A absent any knowledge of Y_B. What if we are interested in the chance that someone has a mutation in gene A and we know they do not have a mutation in gene B? In this case, we introduce the **conditional probability** $P(Y_A = 1|Y_B = 0)$. This is defined as the chance that gene A has a mutation in a person if we know there is no mutation at gene B. If we want to think about this in terms of population averages, it is the fraction of mutations in gene A among only those people without mutations in gene gene B.

How do we calculate this? Using N to denote the number of individuals in a population with a given gene configuration,

$$P(Y_A = 1|Y_B = 0) = \frac{N(Y_A = 1, Y_A = 0)}{N(Y_B = 0)} = \frac{N(Y_A = 1, Y_A = 0)/n}{N(Y_B = 0)/n} \tag{19}$$

$$= \frac{P(Y_A = 1, Y_B = 0)}{P(Y_B = 0)} \tag{20}$$

This is a specific instance of Bayes' formula:

$$P(Y|X) = \frac{P(Y, X)}{P(X)} \tag{21}$$

Two variables are said to be **independent** if $P(Y|X) = P(Y)$. This is also true for events. Can you see why X being independent of Y implies Y is independent of X.

For our purposes, it is important to understand the process of conditioning with data.

Example 4. *Showing independence*

Example 5. *Conditional averages*

Exercise 5: *Estimating conditional probability of dice*

Exercise 6: *Conditioning in gene model*

5. Additional exercises

Exercise 7: *Working with homocide data*

Exercise 8: *Simulating covid*