

# MODELS AND SIMULATION

ETHAN LEVIEN

## CONTENTS

1. Statistical models	1
2. Random variables and distributions	2
3. Python as a tool for statistical modeling	3
3.1. Simulations	3
3.2. Visualization	3
3.3. Monte Carlo	3
3.4. Means and variances	3
4. Joint probabilities, Independence	4
5. More random variables	4
5.1. Binomial: Distribution	4
5.2. Normal distribution and the central limit theorem	5

## 1. STATISTICAL MODELS

A central concept to this course is that of a **model**. Broadly speaking, models are simplified representations of the world<sup>1</sup> 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:

(1) 
$$F = ma$$

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 laws 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

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

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. Instead, they are based directly on observations (data) or our intuitions. In economics, for example, models are built in numerous assumptions about the human nature, many of which may seem obvious or intuitive. However, our intuition about economics is itself shaped by the particular economic system we experience, and in reality these assumptions regularly break down as economic conditions change. 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 all the pine trees in New Hampshire as a function of their age ( $x$ ), we might begin by searching for a function  $f$  such that

(3) 
$$y = f(x)$$

---

Date: April 2022.

<sup>1</sup>Model is sometimes used interchangeably with **theory**, although I usually think of models as being smaller in scope.

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 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. We could construct a more sophisticated model which includes, say, the DNA sequence of the tree ( $g$ ):

$$(4) \quad y = f(x, g).$$

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. This is of course impossible, as there are million, perhaps billions of such variables and their effects are far beyond our understanding of biology. Moreover, it is not so useful to include these variables, since we can actually measure most of them and many of them will have small effects. 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

$$(5) \quad y = f(x) + \epsilon$$

where  $\epsilon$  is a random variable (usually one that is zero, on average).

## 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  $1, 2, \dots, 6$  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**,

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

where  $q$  is the fraction of students in the college who were born in the US. 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.

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

$$(7) \quad Y \sim \text{Bernoulli}(q).$$

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

$$(8) \quad \text{Variable} \sim \text{Distribution}(\text{parameters}).$$

We will sometimes use  $\theta$  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. It seems like our best estimate of  $q$  should be the fraction of  $4/5$ . This is an example of **statistical inference**. More generally, we use statistical inference to make predictions about things we don’t observe based one what we do observe (data).

# MODELS AND SIMULATION

3

## 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.arange(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  $q_1 + \dots + q_k = 1$ . 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. 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

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

The histogram shows us the frequency of different outcomes.

**3.3. 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.

**Example 2.** *Running Monte Carlo simulations*

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

$$(9) \quad \bar{y} = \frac{1}{n} \sum_i y_i$$

Often it is useful to quantify the deviations from the mean. Suppose  $y_i$  can take on outcomes  $y_1, \dots, y_m$ . If  $n$  is large, then the fraction of samples for which  $Y_1 = y_1$  will be  $P(Y_1 = y_1)$ .

$$(10) \quad \bar{y} \approx \frac{1}{n} \sum_{\text{outcomes}} y_i n_i = \sum y_i P(Y_i = y_i)$$

The expression on the right is the definition of expected value, often denoted  $\mathbb{E}[Y]$ .

For this, we have the **sample standard deviation**

$$(11) \quad \sigma = \sqrt{\frac{1}{n-1} \sum_i (y_i - \bar{y})^2}.$$

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

$$(12) \quad \text{Var}(Y) = \sum (y_i - \mathbb{E}[Y])^2 P(Y_i = y_i).$$

Importantly, we can always calculate the mean and standard deviation 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.

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

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

**Example 3.** *Verifying an analytical formula with simulations*

**Exercise 4:** *Verifying an analytical formula with simulations*

#### 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  $Y_1$  and  $Y_2$  which model whether a person has mutations at two different genomes. In this case, we need a model of both variables together:

$$(13) \quad \mathbb{P}(Y_1, Y_2) = \begin{cases} q_{00} & \text{if } Y_1 = 0 \text{ and } Y_2 = 0 \\ q_{01} & \text{if } Y_1 = 0 \text{ and } Y_2 = 1 \\ q_{10} & \text{if } Y_1 = 1 \text{ and } Y_2 = 0 \\ q_{11} & \text{if } Y_1 = 1 \text{ and } Y_2 = 1 \end{cases}$$

The probability distribution  $P(Y_1, Y_2)$  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:

$$(14) \quad P(Y_1) = \sum_y P(Y_1, y) = P(Y_1, 0) + P(Y_1, 1)$$

where in the general the sum is taken over all possible outcomes for the second variable.  $\mathbb{P}(Y_1)$  is defined similarly. Sometimes, we are interested in the value of a random variable given, or **conditioned on** another random variable. The probability of that  $Y_1 = 1$  if we know  $Y_2 = 0$ , denoted  $P(Y_1 = 1 | Y_2 = 0)$  would be the chance that gene 1 has a mutation in a person if we know there is no mutation at gene 2. Said another way, we are looking at the chance of someone having a mutation at gene 1 among only those people that do not have a mutation at a gene 2.

How do we calculate this? Bayes theorem tell us that for two random variables  $X$  and  $Y$

$$(15) \quad \mathbb{P}(Y|X) = \frac{P(Y, X)}{P(X)}$$

Two variables are said to be **independent** if  $\mathbb{P}(Y|X) = P(Y)$ . This is also true for events.

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*

#### 5. MORE RANDOM VARIABLES

**5.1. Binomial: Distribution.** A situation that often arises is that we take many, say  $n$ , independent samples from a Bernoulli distribution. Now let  $Y$  be the number of 1s. Then  $Y$  follows **binomial distribution**:

$$(16) \quad Y \sim \text{Binomial}(n, q)$$

The binomial distribution has two parameters,  $k$  and  $p$ , and the probability distribution is

$$(17) \quad \mathbb{P}(Y) = \binom{n}{Y} q^Y (1 - q)^{n-Y}$$

The binomial distribution has a mean  $qn$  and variance  $nq(1 - q)$ . We can draw samples from a binomial distribution

```
> y = np.random.binomial(n,p,n_samples)
```

# MODELS AND SIMULATION

5

**5.2. Normal distribution and the central limit theorem.** In the previous example, we say that if we take the average of many Bernoulli random variables, we get a histogram that looks a lot like a “bell curve” with a standard deviation that scales as  $1/\sqrt{n}$ .

It turns out this is true when we add up *any* sequence of independent and independent distributed random variables which are not too pathological (actually it is also true for many sequences of random variables which are not independent). Since the “bell curve” arises in the limit where we sum or average many random variables. This motivates us to define the Normal distribution. Unlike previous random variables we’ve seen, a normal random variable, denoted

$$(18) \quad Y \sim \text{Normal}(\mu, \sigma)$$

can take on any number, positive or negative, decimal or integer. We can generate Normal random variables in python with

```
> np.random.normal(0,1)
```

We want to describe this random variable in terms of a probability distribution, but the probability for  $Y$  to equal any given value of  $Y$  is zero. This is because the probabilities must sum to one and that won’t work out due to math stuff we don’t worry about. The key point is that the probability for  $Y$  to fall in a certain range should depend on how wide that range is. For example, the chance that someone is six feet tall plus  $h$  inches is going to decrease as  $h$  gets very small, and the chance that someone is *exactly* 6 feet tall is zero (although it won’t appear that way in data to to imprecision in our measurements). Therefore, instead of defining our model in terms of the chance that someone is exactly  $Y$  feet tall, we define it in terms of a density,  $f$ , such that for very small  $h$

$$(19) \quad P(y \leq Y \leq y + h) \approx f(y)h.$$

We can think of  $f$  as the probability *per unit of the random variable*, e.g. probability/feet. This means if we want to know what an arbitrary probability is, we can sum over  $f$ :

$$(20) \quad P(y \leq Y \leq y + h) \approx \sum_j f(y + j/Lh)h/L$$

(If know calculate, you will recognize this becomes an integral) For the Normal distribution,

$$(21) \quad f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

This is the classic bell curve shown in Figure 1.

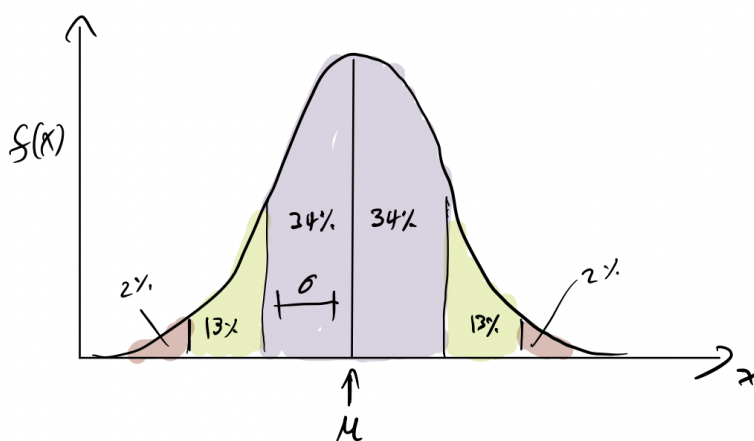


FIGURE 1. Probabilities in the Normal distribution

**Exercise 6:** Comparing histograms

**Example 6.** Working with Normal random variables

**Exercise 7:** More with Normal random variables

Going forward, you will need to know the following properties of Normally distributed random variables. Let

$$(22) \quad Y_1 \sim \text{Normal}(0, 1), \quad Y_2 \sim \text{Normal}(0, 1)$$

Then

$$(23) \quad aY_1 + b \sim \text{Normal}(b, a)$$

$$(24) \quad aY_1 + bY_2 \sim \text{Normal}(0, \sqrt{a^2 + b^2})$$

We now return to the connection to sums of random variables and the Binomial distribution. The **central limit theorem** tells us that for a set of random variables  $x_1, \dots, x_n$  with each have  $\mathbb{E}[x_i] = \mu_x$  and  $\text{var}(x_i) = \sigma_x^2$ ,

$$(25) \quad Y = \sum x_i$$

is approximately Normal with mean  $\mu$  and variance  $\sigma/\sqrt{n}$ . In other words Normal random variables emerge when we add up many small sources of randomness.

**Exercise 8:** *Learning about a new distribution using simulations*