PROBABILITY CONCEPTS

ETHAN LEVIEN

Contents

1. Statistical models and and random variables

I key concept in the course is that of a **model**. Before doing any mathematics, let's reflect on the notion of a model in the abstract.

Exercise 1. *Reflect on the different meanings of the word model. What do fashion models, architectural models, mathematical models and other sorts of models have in common? What are some of the reasons we utilize "models" of things?*

Broadly speaking, models are ways of simplifying aspects of the world in order to draw conclusions. We we might be familiar with the idea of a mathematical model, perhaps from physics, as a mathematical description of the relationship between variables. For example, in Newton's law we relate force to mass and acceleration. This is useful, because if you observe a particle we can measure its mass and acceleration (e.g. a distant star), we can perhaps say something about the force being exerted on it. Newton's laws are what we might call *fundamental*, in the sense that they hold true all throughout nature and other theories are built upon them. In statistics, we rarely have something like Newton's laws. In economics for example, models are built in numerous assumptions about the human nature and often fail. Such models can still be quite useful provided we have a firm understanding of their range of applicability and they can at least make some predictions (this is never true in economics).

A **Statistical model** is simply a model in which at least some of the variables are **random variables**. To my knowledge, the term **probabilistic model** is sometimes used to mean the same thing as a statistical model. Returning to Newton's example, we could imagine the force exerted on a particle is random (as is the case for molecules inside cells), and then (assuming the mass is fixed) the acceleration would also be random. Now imagine an example of more relevance to this course: Consider the variable Y representing whether a randomly selected student from Dartmouth identifies as a republican or not. Y is a random variable because we cannot predict it until we ask a student.

Exercise 2. *What are some variables in nature where it makes sense to use non-statistical, or* **deterministic** *models, and some where it makes more sense to use statistical models?*

We can define a random variable more general as a variable we can not predict exactly prior to an observation of the variable, no matter how much information we have (e.g. the roll of a dice). When we observe a random variable, it will take on a value from a set of possible outcomes ($1, 2, \dots, 6$ for the dice). We can describe a characterize a random variable using a **probability distribution**, which maps a set of possible outcomes to to real numbers between 0 and 1. Usually the outcomes are numbers, even if we use a number to represent a non-numerical quantity (e.g. whether someone is a smoker). In statistics, we often refer to observations of random variables as **samples**. Often when we generate samples using a computer we call them **simulations**.

Let consider a concrete example: A random variable X taking on the possible outcomes in $\{0, 1\}$ is said to be a Bernoulli random variable. In this case we can write the probability distribution as

$$P(X) = \left\{ \begin{array}{cc} q & X = 0 \\ 1-q & X = 1 \end{array} \right. \tag{1}$$

Returning to the example of the Dartmouth Survay: Well, it has to be a Bernoulli distribution, since there are only two outcomes. In order to say that we are modeling Y with a Bernoulli distribution symbolically, we write

$$Y \sim \text{Bernoulli}(q). \tag{2}$$

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

$$\text{Variable} \sim \text{Distribution}. \tag{3}$$

Turning back to the model, we notice there is a missing piece: the value of q. Without this, we can't make any predictions at all. The process is determining q based on data (e.g. a survey of Dartmouth students or conversations we have had) is known as **statistical inference**.

Date: April 2022.

Exercise 3. *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. Use Monte Carlo simulations to find the average time it takes to get a* 1.

2. More distributions

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

2.2. **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 something that looks a lot like a bell curve. It turns out this is true when we add up *any* random variables which are sufficiently independent (we will make this precise soon).

2.3. **Other distributions.** Read the wikipedia page for the following random variables

- Binomial
- Chi-Squared

2.4. **Independence and correlation.** [POLISH] We introduce, very briefly, the concepts of independence and conditioning. Given two random variables, X and Y, to condition Y on X (denoted $Y|X$) means we are looking at Y for a fixed value of X. Two random variables are independent if conditioning does not change their probability distribution.