Axioms of Probability
Math 217 Probability and Statistics

Prof. D. Joyce, Fall 2014

Probability as a model. We'll use probability to understand some aspect of the real world. Typically something will occur that could have a number of different outcomes. We'll associate probabilities to those outcomes. In some situations we'll know what those probabilities are. For example, when there are n different outcomes and there's some symmetry to the situation which indicates that all the outcomes should have the same probability, then we'll assign all the outcomes a probability of $1/n$. In other situations we won't know what the probabilities are and our job is to estimate them.

Besides individual outcomes we'll want to assign probabilities to sets of outcomes. We'll use the term *event* to indicates a subset of all the possible outcomes.

Discrete probability distributions. We'll start with the case of discrete probabilities and then go to the general case.

Definition. A *discrete probability distribution* on a set Ω is a function $P : \Omega \to [0,1]$, defined on a finite or countably infinite set Ω called the *sample space*. The elements x of Ω are called *outcomes*. Each outcome is assigned a number $P(x)$, called the *probability* of x with $0 \leq P(x) \leq 1$. It is required that the sum of all the values $P(x)$ equals 1. Furthermore, for each subset E of Ω, called an *event*, we define the *probability* of E, denoted $P(E)$, to be the sum of all the values $P(x)$ for s in E.

It's useful to have a special symbol like Ω to denote a sample space. That way when you see it, you can tell right away what it is.

Note that we're assigning to the event which is the singleton set $\{x\}$ the same probability we assign to the outcome x, that is $P(\{x\}) = P(x)$. If nothing else, this allows us to dispence with excess parentheses.

Most of our examples of discrete probabilities will have finite sample spaces, in which case the discrete probability is called a *finite probability*.

An example that we've already looked at is rolling a fair die. The sample space is $\Omega = \{1,2,3,4,5,6\}$. The probability of each of the six outcomes is $\frac{1}{6}$. The probability of an event E depends on the number of outcomes in it. If E has k elements, then $P(E) = k/6$.

More generally, whenever you have uniform discrete probability with n possible outcomes, each outcome has probability $1/n$, and the probability of and event E is equal to its cardinality divided by n. So $P(E) = |E|/n$.

Here's an example of an infinite discrete probability. Toss a fair coin repeatedly until an H shows up. Record the number of tosses required. The sample space is $\Omega = \{1,2,3,4,\ldots\}$. We can easily determine the probabilities on Ω. With probability $\frac{1}{2}$, H shows up on the first toss, so $P(1) = \frac{1}{2}$. Otherwise, we get T and toss the coin again. Since on the second toss we'll get H with probability $\frac{1}{2}$, but we only reach the second toss with probability $\frac{1}{2}$, therefore $P(2) = \frac{1}{4}$. Likewise, $P(3) = \frac{1}{8}$, and in general, $P(n) = 1/2^n$. Now, this function satisfies the condition to be a discrete probability since the sum of all the values $P(x)$ equals 1. That's because

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n} + \cdots = 1.$$

Recall that infinite sums are called *series*. This particular series is an example of a geometric series.

Continuous probabilities. Discrete distributions are fairly easy to understand whether they're finite or infinite. All you need to know for a such a distribution is the probablilities of each individual outcome x, which we've denoted $P(x)$. Then the probability of an event is just the sum of the probabilities of all the outcomes in the event.

Besides discrete distributions, there are also continuous distributions. They are just as important as the discrete ones, but they're a little more complicated to understand, so first we'll study the discrete case in depth. Even thought we're putting off a study of the continuous case, we should at least survey what they are.

In a continuous distribution there are uncountably many outcomes, and the probability of each outcome x is 0. Therefore, the probability of any event that has only finitely many outcomes in it will also be 0. Furthermore, the probability of any event that even has countably infinitely many outcomes in it will also be 0. But there will be events that have uncountably many outcomes, and those events may have positive probability. We'll need to look at examples to see how this works.

(Recall that an infinite set is *countable* if its elements can be listed, that is, it is in one-to-one correspondence with the positive integers $1, 2, 3, \ldots$. But many useful infinite sets are uncountable. The set of real numbers is uncountable as are intervals of real numbers.)

A continuous probability example. Our example is the uniform continuous distribution on the unit interval $\Omega = [0, 3]$. The outcome of an experiment with this distribution is some number X that takes a value between 0 and 3. We want to capture the notion of uniformity. We'll do this by assuming that the probability that X takes a value in any subinterval $[c, d]$ only depends on the length of the interval. So for instance, $P(X \in [0, 1.5]) = P(X \in [1.5, 3])$ so each equals $\frac{1}{2}$. Likewise, $P(X \in [0, 1]) = P(X \in [1, 2]) = P(X \in [2, 3]) = \frac{1}{3}$. In fact, the probability that X lies in $[c, d]$ is proportional to the length $d - c$ of the subinterval. Thus, $P(X \in [c, d]) = \frac{1}{3}(d - c)$.

We can still assign positive probabilities to many events. For instance, we want $P(X \leq \frac{1}{2})$ to be $\frac{1}{2}$, and we want $P(X \geq \frac{1}{2})$ to be $\frac{1}{2}$, too. In fact, if $[a, b)$ is any subinterval of $[0, 1)$ we want $P(a \leq X < b)$ to be $b - a$, the length of the interval.

It turns out, as we'll soon see, that an entire continuous distribution is determined by the values of a function F defined by $F(x) = P(X \leq x)$ for all numbers x. Note how we're using the symbol X for a real random variable, that is, the numerical outcome of an experiment, but we're using the symbol x to denote particular real numbers. The function F is called a *cumulative distribution function*, abbreviated c.d.f., or, more simply, a *distribution function.*

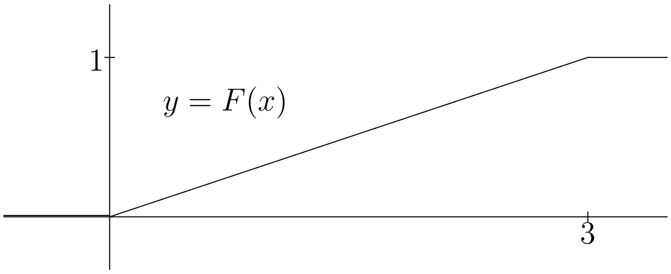
We can use F to determine probabilities of intervals as follows

$$P(X \in [a, b]) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

For a uniformly continuous random variable on $[0, 3]$ the c.d.f. is defined by

$$F(X) = \begin{cases} 0 & \text{if} \quad x \leq 0 \\ x/3 & \text{if} \quad 0 \leq x \leq 3 \\ 1 & \text{if} \quad 3 \leq x \end{cases}$$

Its graph is



More generally, a uniformly continuous random variable on $[a, b]$ has a c.d.f. which is 0 for $x \leq a$, 1 for $x \geq b$, and increases linearly from $x = a$ to $x = b$ with derivative $\dfrac{1}{b - a}$ on $[a, b]$.

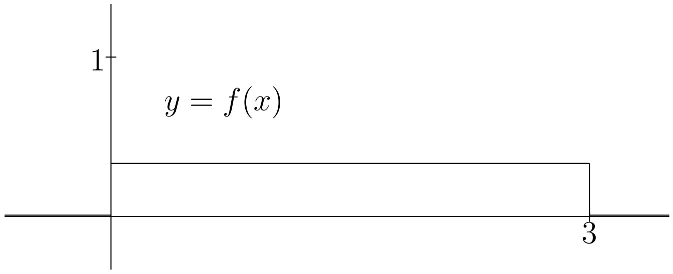
The probability density function $f(x)$. Although the c.d.f. is enough to understand continuous random variables (even when they're not uniform), a related function, the density function, carries just as much information, and its graph visually displays the situation better. In general, densities are derivatives, and in this case, the probability density function $f(x)$ is the derivative of the c.d.f. $F(x)$. Conversely, the c.d.f. is the integral of the density function.

$$f(x) = F'(x) \quad \text{and} \quad F(x) = \int_{-\infty}^{x} f(t)\, dt$$

For our example of the uniformly continuous random variable on $[0,3]$ the probability density function is defined by

$$f(X) = \begin{cases} 0 & \text{if} \quad x \le 0 \\ 1/3 & \text{if} \quad 0 \le x \le 3 \\ & \text{if} \quad 3 \le x \end{cases}$$

Its graph is



More generally, a uniformly continuous random variable on $[a,b]$ has a density function which is 0 except on the interval $[a,b]$ where it is the reciprocal of the length of the interval, $\dfrac{1}{b-a}$.

We'll come back to continuous distributions later, and when we do, there will be many to consider besides uniform distributions on intervals.

Next we'll look at axioms for probability. These are designed to work for both discrete and continuous probabilities. They also work for mixtures of discrete and continuous, and they even work when the sample space is infinite dimensional, but we won't look at that in this course.

Kolmogorov's axioms for probability distributions. It's not so easy to define probability in the nondiscrete case. The probabilities of the events (subsets of the sample space Ω) are not determined by the probabilities of the outcomes (elements of the sample space). A number of mathematicians including Émile Borel (1871–1956), Henri Lebesgue (1875–1941), and others worked on the problem in the first part of the 1900s and developed the concept of a measure space. As Andrey Kolmogorov (1903–1987) described it in 1933, a probability distribution on a sample space, which is just a measure space where the measure of the entire space is 1. When the total measure is 1, the measure is also called a probability distribution.

Definition. A *probability distribution* P on a set Ω, called the *sample space* consists of

- a collection of subsets of Ω, each subset called an *event*, so that the collection is closed under countably many set operations (union, intersection, and complement),
- for each event E, a real number $P(E)$, called the *probability* of E, with $0 \le P(E) \le 1$, such that
- $P(\Omega) = 1$, and
- P is countably additive, that is, for every countable set of pairwise disjoint events $\{E_i\}$,

$$P\left(\bigcup_i E_i\right) = \sum_i P(E_i)$$

The requirement that the sets be closed under countably many set operations is that given countably many events, their union and intersection are also events; also the complement of an event is an event.

The term *pairwise disjoint* in the last item means that $E_i \cap E_j = \emptyset$ for all $i \ne j$.

The last condition applies to any countable number of events, even 2. In that case it says that if E and F are disjoint events, then $P(E \cup F) = P(E) + P(F)$.

This definition extends the definition for discrete probability distributions given above, and it allows for continuous probabilities. Although there are technical aspects to this definition, since this is just an introduction to probability, we won't dwell on them.

Properties of probability. The definition doesn't list all the properties we want for probability, but we can prove what we want from what it does list.

For example, it doesn't say that $P(\emptyset) = 0$, but we can show that. Since Ω and $\emptyset$ are disjoint (the empty set is disjoint from every set), therefore $P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset)$. But $\Omega \cup \emptyset = \Omega$, so $1 = 1 + P(\emptyset)$. Thus, $P(\emptyset) = 0$.

We'll prove the following theorem in class with the help of Venn diagrams to give us direction.

Theorem. Let Ω be a sample space for a discrete probability distribution, and let E and F be events.

- If $E \subseteq F$, then $P(E) \leq P(F)$.
- $P(E^{\mathsf{c}}) = 1 - P(E)$ for every event E.
- $P(E) = P(E \cap F) + P(E \cap F^{\mathsf{c}})$.
- $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

That last property is the *principle of inclusion and exclusion* for two events. There is a version of it for any finite number of sets. For three sets E, F, and G, it looks like

$$\begin{aligned} & P(E \cup F \cup G) \\ = \; & P(E) + P(F) + P(G) \\ & - P(E \cap F) - P(E \cap G) - P(F \cap G) \\ & + P(E \cap F \cap G) \end{aligned}$$

Partitions. We can generalize the property $P(E) = P(E \cap F) + P(E \cap F^{\mathsf{c}})$ mentioned above in a useful way.

A set S is said to be *partitioned* into subsets $A_1, A_2, \ldots, A_n$ when each element of S belongs to exactly one of the subsets $A_1, A_2, \ldots, A_n$. That's logically equivalent to saying that S is the disjoint union of the $A_1, A_2, \ldots, A_n$. We'll have that situation from time to time, and we can use it to our advantage to compute probabilities.

If the sample space Ω is partitioned into events $A_1, A_2, \ldots, A_n$, and E is any event, then E is partitioned into $E \cap A_1, E \cap A_2, \ldots, E \cap A_n$. Then by the last axiom for probability, we have

$$P(E) = P(E \cap A_1) + P(E \cap A_2) + \cdots + (E \cap A_n).$$

Math 217 Home Page at `http://math.clarku.edu/~djoyce/ma217/`