Probability Notes

Shiro Kuriwaki

Harvard University
kuriwaki@g.harvard.edu

Last updated September 21, 2018

> "Experts estimate a 35% chance of a U.S. civil war over the next ten to fifteen years. What do historians think? http://nyer.cm/MadCpII"
>
> Tweet by *The New Yorker* (2018)

(These notes are designed to accompany a social science statistics class. They emphasize important ideas and tries to connect them with verbal explanation and worked examples. However, they are not meant to be comprehensive, and they may contain my own errors, which I will fix as I find them. I rely on multiple sources for explanation and examples.[1] Thanks to the students of API-201Z (2017) for their feedback and Matt Blackwell for inspiration.)

WHERE ARE WE? WHERE ARE WE GOING?

An applied statistics class often starts with probability, a fundamental concept in quantitative reasoning about data.[2] Probability is especially important for policy makers and social scientists, who must make inferences and predictions about the social world. Probabilistic statements are intuitive and commonplace, but as the headline quoted above highlights, our intuitions can lead us awry. To think clearly requires some formal definitions, as well as the concept of about random variables and distributions. Probability and inference are mirror images of each other.

After probability the class moves on to inference. After that, we move on to models (namely linearly regression) for making inferences – the credibility of these models rests on statements about the distribution of the estimators used. Throughout, probability is fundamental.

[1] DeGroot and Schervish (2012), Blitzstein and Huang (2014), Blitzstein and Morris (unpublished book), and Imai (2017)

[2] But see Imai (2017) for a flipped order.

CONTENTS

CHECK YOUR UNDERSTANDING

- What is probability?
- What is conditional probability?
- What is another way to write the quantity $P(A \mid B)$?
- What is independence?
- What is a random variable?
- A random variable is a function. Then why is it called “random”?
- How do random variables and distributions relate to each other?
- How do we calculate the expectation and variance of a random variable in practice?

THE COUNTING DEFINITION OF PROBABILITY

Probability pervades our lives. We use words like “likely”, “I’m not sure”, “maybe”, “definitely” that convey probability. This leads to a intuitive definition of probability,

$$P(\text{event}) = \frac{\text{Number of events}}{\text{Number of possibilities}}$$

I will refer to this as the “counting definition” of probability. It is not technically correct but it is simple: We count up the number of events that we care about, we count up the number of all possible outcomes, and then we divide the former by the latter. This is how we often think when asked problems like "What is the probability that I roll a 2 from a die?".

This is a straightforward way of thinking about uncertainty and probability. Straightforward doesn't mean easy: counting sequences of complex events can get taxing quickly. For example, we count when we sample "with replacement" or "without replacement", and that requires different measures of counting. Problems involving complex counting schemes are a type of combinatorics problem.[3]

Example 1 (n choose k)**.** To fulfill a degree, a student can choose 4 out of 8 courses, with the constraint the at least 1 of the 4 chosen courses must be a statistics course. Suppose that 3 out of the 8 possible courses are statistics courses. How many permissible choices of a course schedule are there?

On the other hand, the counting definition of probability can only get us so far. Consider the following question: What is the probability that aliens exist? A naive version of the counting definition would give the answer $\frac{1}{2}$:

$$\frac{\{\text{Aliens exist}\}}{\{\text{Aliens exist, Aliesn don't exist}\}}$$

The number $\frac{1}{2}$ looks implausible. Further, what if one were asked instead the probability that intelligent life exists outside earth? The naive definition of probability would *also* give $\frac{1}{2}$

$$\frac{\{\text{Intelligent life exists}\}}{\{\text{Intelligent life exiss, Intelligent life don't exist}\}}$$

This makes the setup even more suspect, because we'd think that the probability of intelligent life existing should be *smaller* than the probability that any life exists.

The toy example[4] aims to show mainly two shortcomings of the counting version of probability. First, counting often treats each outcome equally likely even when in reality they are not. Second, counting can only count finite sample spaces[5].

A MORE RIGOROUS DEFINITION OF PROBABILITY

Definition 1 (Probability)**.** Probability is a function that maps events to a real number, obeying the axioms of probability. ■

I will get to the "axioms" shortly, but first, why is this definition useful? One role of probability is to serve as a transparent way to quantify events. As analysts of the social world, we usually

[3] Here's a hard one: "Suppose a class has 20 students, and there are exactly 20 seats in the classroom. Everyone attends both of the first two days. On both days, the students choose their seats completely randomly, one student per seat. What is the probability that no one sits in the same seat on both days?" (Blitzstein and Huang, 1.50)

[4] Drawn from Blitzstein, Stat 110

[5] It might seem unnecessary to think about infinite sample spaces for practical purposes. However, many things we care about indeed have infinite sample spaces. Any continuous variable, such as age, can take on infinite values if we treat it as a real number: 50, 50.1, 50.11, 50.111, 50.1111, ...

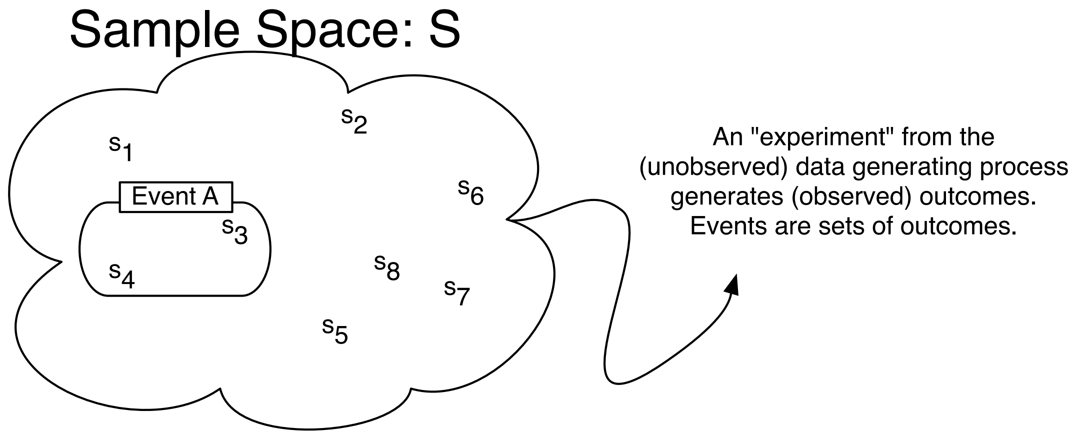
Sample Space: S
s1
s2
s3
s4
s5
s6
s7
s8
Event A
An "experiment" from the (unobserved) data generating process generates (observed) outcomes. Events are sets of outcomes.

Figure 1: Probability Measures Events

care about events – e.g., the event that a citizen has a certain opinion, or the event that another has a disease. However, the real world is messy; numbers are simpler. Numbers are also more transparent and universal: two analysts on the other side of the world may not understand an event in the same way (being people), but numbers have universal meaning.

Another thing probability handles is uncertainty of events. Fundamental to the setup of how probability is defined is the sample space S (some books use Ω). Formally, this is the *set of all possible outcomes of some experiment*. An experiment in the statistics sense (not to be confused as an experiment in psychology or a randomized control trial) is the process by which a possible outcome becomes a realized outcome. When a pollster asks a voter about his voting preferences, she might support the Democrat, support the Republican, not respond, respond not being sure, and so on. In probability we conceive of "an event A happening" in the intuitive sense as the event that the *observed outcome is in the set A*. That's the jump we make from events to set notation, and the reason we use sets in probability. With this footing, the additivity axiom of probability is a principled way to assign numbers to new events.

Definition 2 (The Countable Additivity Axiom of Probability)**.** For any probability function P,

$$P(A_1 \cup A_2) = P(A_1) + P(B_2) \text{ for disjoint} A, B$$

The extension of this is to take the union of more than two events,

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$$

■

The intuition for this axiom is that $A \cup B$ is adding the areas (measures, or likeli-ness) of A and B together. If the two events are disjoint (i.e. no overlap), the addition is simple and we don't need to do anything else.

We often add many, or even infinite things. So will this sum ever become infinity? No, because we have a *separate* axiom of probability that limits the sum of such probabilities of large unions to 1.

This might sound trivial — how else beside adding would you combine two events? Maybe so, but this linkage between unions and additions of probabilities will be handy with complex problems. Also, the important link this axiom is making is that *probabilities turn events into numbers.* It's thanks to this that we can easily think of "adding" things in the mathematical sense. Otherwise, how would you add events that are not numbers?

Common Error 1 (Category Error of P)**.** $P(A)$ is always a **number** between 0 and 1, not an event. A on the other hand is always a **event**, not a number. Confusing these is a type of Category Error. ■

CONDITIONAL PROBABILITY

Probability $P(A)$ quantifies the uncertainty that the event A occurs. But as we know from experience, context matters. That is in different contexts we enjoy different sets of information that change our uncertainty about the occurrence of events. For example, an analyst might tabulate the weather in Boston for the past 50 years and find that proportion of rainy days is 0.3, thus the probability that it rains on any future day (if weather is climate unchanged) is 0.3.

This claim sounds solid on the one hand, but on the other hand we would think that we would come to a different conclusion *given* other information. For example, if the future date in question was in the death of winter, the probability of rain (not snow) might be lower. And if the date in question preceded another rainy day, the probability of rain might be higher.

Conditional probability resolves this uneasiness. If A and B are events, the conditional probability of A given B is the probability that event A occurs given the condition that event B has already occurred. Mathematically we write this as:

Definition 3 (Conditional Probability)**.** The conditional probability of A given B is denoted $P(A \mid B)$ and this probability is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

■

The right-hand side of this formula has a natural English translation. The event A occurring given B already occurring is considering in a loose sense that both A and B occur. However, the event B is on the condition, so we need to re-scale our probability by the base rate of B (the condition) happening at all. Thus we divide by $P(B)$.

Rearranging the terms (bringing the denominator on the left-hand side to the right-hand side by multiplying both sides), we get the multiplication rule.

$$P(A \cap B) = P(A \mid B)P(B)$$

This also has a English translation if we think of the condition as coming temporally prior. First, B occurs (with probability $P(B)$). Then the probability of A and B both occurring is the probability that you get A conditional on B ($P(A \mid B)$) after getting B. Thus we multiply the two.

Common Error 2 (Mistaking $A \mid B$ as an event)**.** Conditioning can only happen once, and it is a statement that is distinct from set operations. So when we say $P(A \mid B)$, that is *not* the probability of an event "$A \mid B$" – that is not an event (category error). We stick to the original definition; as the event that A occurs given that B has occurs. ■

In naive probability, often with using cards and dice, we are used to writing $P(A_1 \cap A_2 \cap ...A_3)$ as the product $P(A_1)P(A_2)P(A_3)$. But we should be cautious in doing this switch – this only holds when all events are independent together (and in the social world, not many things are). Note that we could have written the conditional probability with a different quantity, $P(B \mid A) = \frac{P(B \cap A)}{P(A)}$ and then get $P(A \cap B) = P(B \mid A)P(A)$. Recall that $P(A \cap B)$ and $P(B \cap A)$ are equivalent.

LAW OF TOTAL PROBABILITY

Conditional probability brought in a second event distinct from the first one, B, to our discussion of events. It might seem like bringing in another event into the discussion complicates things. But in practice we rely on conditional probability to make computations easier.
An intuition that is useful to have about conditional probability is that it breaks up the real world to manageable pieces. Take the beginning rain example.

$$P(\text{Rain})$$

may at first seem simple to calculate, but as we think more about it, capturing such a general event requires us to think about various conditions — only under those conditions do we really have a good intuition about the likelihood. The Law of Total Probability captures this desire to partition things into manageable pieces. First, let's define what a partition is.

Definition 4 (Partition)**.** A set of events are said to *partition* a space when they are mutually exclusive and together union covers the entire that said space. Almost all of the time, the space in question is a sample space. If mutually exclusive events $B_1, B_2, \dots B_k$ partition a space, then $\bigcup_{j=1}^{k} B_j = S$. ■

Definition 5 (Law of Total Probability)**.** Any event A can be re-written as a total of conditional probabilities by conditioning on a partition of the sample space. Refer to this partition as the set of events $B_1, B_2, \cdots B_k$. Then

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_k)P(B_k)$$

Or in shorthand,

$$P(A) = \sum_{j=1}^{k} P(A \mid B_j)P(B)$$

■

The Law of Total Probability may be easy to see in this way:

$$P(A) = \sum_{j=1}^{k} P(A \cap B_j)$$

which is equivalent to the above. As long as we "capture" the entire space in which A lives exhaustively without any overcounting, we can cobble together the "measure" of A by combining $P(A \cap B_j) = P(A \mid B_j)P(B_j)$ for all $j = 1, 2, ..., k$. Remember this only works because the events $B_1, ...B_k$ is a partition: they are disjoint from one another and cover the whole sample space.

INDEPENDENCE

There is a special name for a relationship where conditioning does not seem to change our views of uncertainty. This is independence, which also has a wide every-day use and is the crux of many statistical analyses.

Definition 6 (Independence)**.** Two events are **independent** when knowing one thing (say event B) does not change the likelihood (probability) of the other thing (say event A).

$$P(A \mid B) = P(A)$$

We can also use the definition of conditional probability to get the definition of independence presented in lecture:

$$\frac{P(A \cap B)}{P(B)} = P(A)$$
$$P(A \cap B) = P(A)P(B)$$

■

In English, when you're told for example two events (e.g. two political scandals) developed independently, the first image that might come to mind is that the two events are "separate". However, this impression can be misleading!

Common Error 3 (Independence and Disjointness)**.** Independence and Disjointness are two different concepts. The former does not imply the latter and the latter does not imply the former. In the simplest example, consider a single coin flip. There are two events, Heads or Tails. These are *disjoint* (both cannot happen at the same time) but they are clearly *not* independent: If you know a flip is Heads, it is definitely not Tails! Conversely, consider two independent coin flips. The event of the first one being heads is *independent* with the second one being heads, but the two events are clearly *not* disjoint because the event that both events happen is clearly possible.
■

The clearest and safest way to understand independence might be through again relying on conditional probability. Two events are independent if knowing information about one event does not *add* or *detract* information about another.

To establish something is independent, you may need to rearrange the terms to recover the definitional form of independence. When dealing with abstract events, you'd need to rely on the basic rules to get to an expression you want. A lot of solving problems like this is pattern recognition.

Example 2 (Independence of Complements)**.** Suppose events A and B are independent. Show that A and B^c are also independent.

Answer We can show independence if $P(A \cap B^c) = P(A)P(B^c)$. We will try to rewrite the left-hand side as the right-hand side taking advantage of the fact that A and B are independent. To use this condition, we'd want an expression with A and B. Notice

$$P(A \cap B^c) = P(A) - P(A \cap B)$$

This can be reasoned verbally or through a Venn diagram. Then

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) \quad \because \text{condition provided} \\ &= P(A)(1 - P(B)) \\ &= P(A)P(B^c) \end{aligned}$$

APPLICATIONS WITH CONDITIONAL PROBABILITY

From some of these definitions of probability we can get some useful ways to find probabilities of interesting events we'd otherwise not be able to estimate.

Bayes Rule is one of the most well-known of such formulations.

Definition 7 (Bayes Rule)**.**

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_k)P(A_k)}$$

For a partition of A denoted as $A_1, A_2, \cdots A_k$ ■

But the intermediary formulations we derive from Bayes Rule from the definition of conditional probability are equally important

$$
\begin{aligned}
P(A \mid B) &= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(B \mid A)P(A)}{P(B)} \\
&= \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)} \text{ or} \\
&= \frac{P(B \mid A)P(A)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_k)P(A_k)} \\
&\text{For a partitions of } A : A_1, A_2, \cdots A_k
\end{aligned}
$$

We end up relying on these relationships when we are asked for a conditional probability we do not observe directly.

Example 3 (The Lie Detector)**.** A Lie Detector gives one of two results: “Honest” or “Liar” and is 80 percent accurate. That is, it detects Honest people correctly 80 percent of the time, and it detects Liars correctly 80 percent of the time. Suppose 10 percent of the entire population are liars. Suppose we tested the lie detector on a representative sample of the population of 100 people. We then inspect closely the subset of people for whom the Lie Detector reported ”Liar”. Among this subset, what is the probability that any given member is truly a liar?
Answer We are asked for a conditional probability – the event that a person is a liar given that the lie detector reported “Liar”. We set this up and try to express it in terms that we have numbers for:

$$
\begin{aligned}
&P(\text{Liar} \mid \text{Reported Liar}) \\
&= \frac{P(\text{Reported Liar} \mid P(\text{Liar})P(\text{Liar})}{P(\text{Reported Liar})} \\
&= \frac{P(\text{Reported Liar} \mid \text{Liar})P(\text{Liar})}{P(\text{Reported Liar} \mid \text{Liar})P(\text{Liar}) + P(\text{Reported Liar} \mid P(\text{Honest})P(\text{Honest})} \\
&= \frac{0.80 \times 0.10}{0.80 \times 0.10 + 0.20 \times 0.90} \\
&\approx 0.30
\end{aligned}
$$

We could decompose the denominator in the above way because we know the events “Liar” and “Honest” are disjoint and partition the sample. Otherwise, we would not have been able to rewrite this probability in this way.
Naively, we might think that at least 80 percent of those who were reported as liars should be liars – after all the test is 80 percent accurate regardless of the truth. But once we ”condition on” a result, we see that we get a completely different picture.

Let’s take another, more conceptual example.

Example 4 (Prosecutor's Fallacy)**.** In a 1999 court case, Sally Clark was convicted of infanticide when two of sons died early at birth. On trial the prosecutor argued that Clark was most definitely likely of infanticide. The argument went as follows. First he cited a statistic that in the population, only 1 out of 8,500 young births result in death from natural causes. The probability that two successive deaths occur naturally, then, is (1/ 8500) times (1/8500): one in over 72 million. This is extremely unlikely. Thus, given the evidence of two successive deaths, the probability that they were natural causes is one in over 72 million. In other words, there is overwhelming evidence that Clark had committed infanticide.

Suppose that the 1 out of 8,500 number is true.[6] and applies to Clark. What are the problem(s) in this reasoning?

Answer 1 The first problem is the multiplication of two probabilities, because they do not deal with dependence. Let E_1 be the event (evidence) that Clark's first child dies of natural causes. Accept as true that $P(E_1) = 1/8500$. Let E_2 be the event that Clark's second child dies of natural causes. Then, from the statistic, we get $P(E_2) = 1/8500$ as well. However, the prosecutor is making claims about **both** children dying of natural causes. Let's call that event $E_{1,2}$.

Is it true that $P(E_1, E_2) = (1/8500)^2$? Not necessarily, because we can rewrite

$$P(E_{1,2}) = P(E_1 \cap E_2) = P(E_1)P(E_2 \mid E_1)$$

If E_2 was independent of E_1, then by definition of independence $P(E_2 \mid E_1) = P(E_2)$ and we do get the $(1/8500)^2$ number. However, this is probably not the case. Perhaps Clark has a genetic condition or is in a social condition that makes it more likely to have a string unfortunate miscarriages.[7] Then $P(E_2 \mid E_1)$ will be higher than $P(E_2)$, and so $P(E_{1,2})$ will be higher than $(1/8500)^2$.

Answer 2 There is also a second, arguably more important problem in the reasoning. The prosecutor's argument sounds like that of conditional probability: the probability of innocence given the evidence is very unlikely. Writing this out in terms of probability of events is helpful to clarify our thinking. Let I be the event that Clark is innocent. Then our main quantity of interest is

$$P(I \mid E_{1,2})$$

However, notice that the 1 in 8,500 statistic is really about the "opposite" quantity – the probability of seeing two deaths in the population (a proxy for innocent population). Is $P(I \mid E_{1,2})$ a similar number to $P(E_{1,2} \mid I)$? From Bayes Rule, we can rewrite this as

[6] What this statement being true means can be technically tricky. How were the numbers obtained and what population does it speak to? Remember that the probability is always defined with respect to a sample space S. What you define S to be *depends on the analyst* who is conceptualizing these probabilities; they differ case by case. Here it would be reasonable to think that the sample space is the life or death for every mother's children, among mothers who have children. If we think this rate is constant across history, then the time window should not matter.

[7] Whether or not $P(E_2 \mid E_1) > P(E_2)$ is always true is debatable, and is an empirical question. One might argue that conditional on the first death, the parents will take necessary precautions to decrease the likelihood the probability of another death in ways that others whose first child did not die would not. In which case, $P(E_2 \mid E_1) < P(E_2)$.

$$P(I \mid E_{1,2}) = \frac{P(E_{1,2} \mid I)P(I)}{P(E_{1,2})}$$

If $P(I \mid E_{1,2})$ is a small number, then Clark's innocence is unlikely and it is likely that she is guilty. On the numerator of the right-hand side we see $P(E_{1,2} \mid I)$, which may not be as small as 1 in 72 million as the prosecutor claimed but is at most 1 in 8,500 under our assumption, which is pretty small. However we multiply this by $P(I)$ – the marginal probability of Innocence – which can be quite large. Furthermore in the numerator we have $P(E_{1,2})$, which is a very small number. When we divide by a small number less than 1, we will get a larger number. Thus, all together, $P(I \mid E_{1,2})$ is likely quite large.

RANDOM VARIABLES

We saw that probabilities (technically, a function called the probability measure) assign numbers to individual events that capture the size of that event in the sample space. To do more with this concept, we need to introduce random variables, another fundamental concept in statistics.

Why do we need another mathematical concept? It turns out probability is useful, but describing real-world events with event notation gets unwieldy very quickly. For example, policy planners might want to know the *total number* of people who buy a product. We can write this event by a sequence of events defined at the individual level – for example, we can define $B_1, B_2, \cdots B_N$ as the events that person 1, 2, up to N participate. $B_1 \cap B_2 \cap B_N^c$ is the event that persons 1 and 2 show up but not N. There are N choose n combinations for the sequence of events. It would be too cumbersome to write out the numerous sets and intersections of these events just to get the total number of attendees.

The second issue with continuing with events is that events cannot be transformed by algebra. Suppose we want to know both the number of people that will buy a product *and* the total revenue from the purchase. We want to multiply the attendance by the price per attendance, but it is not obvious how to express this simply with event notation. Algebra (like multiplication) can be applied to numbers, but not to events.

Random variables solve both of these issues. The definition of a random variable is subtle and sounds abstract:

Definition 8 (Random variables)**.** A random variable, or r.v. (typically denoted by a capital letter, such as X) is a function that maps a given sample space S to a number (technically, the "real" number line). ■

Example 5 (Total Number of Occurrences)**.** Consider three binary outcomes, one for each patient recovering from a disease: R_i denotes the event in which patient i ($i = 1, 2, 3$) recovers from a disease. R_1, R_2, and R_3. How would we represent the total number of people who end up recovering from the disease?

Answer Define the random variable X be the total number of people (out of three) who recover from the disease. Random variables are functions, that take as an input a set of events (in the sample space S) and deterministically assigns them to a number of the analyst's choice.

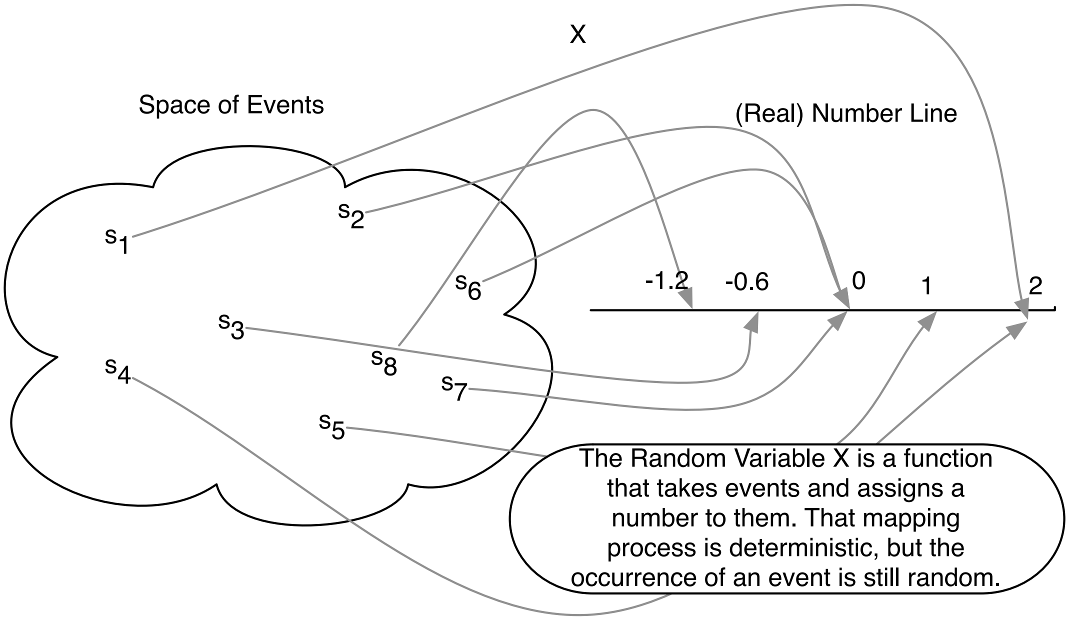
X
Space of Events
(Real) Number Line
s1
s2
s3
s4
s5
s6
s7
s8
-1.2
-0.6
0
1
2
The Random Variable X is a function that takes events and assigns a number to them. That mapping process is deterministic, but the occurrence of an event is still random.

Figure 2: A Random Variable is a Function that Maps Events to Numbers

Because we defined X as the total number of recoveries, it directly follows that X should obey:

$$X(\{R_1, R_2^c, R_3^c\}) = 1 \qquad \text{only patient 1 recovers, so one total recovery}$$
$$X(\{R_1, R_2, R_3^c\}) = 2 \qquad \text{patients 1 and 2 recover, so two total recoveries}$$
$$X(\{R_1, R_2, R_3\}) = 3 \qquad \text{all three patients recover, so three total recoveries}$$

and so on.

In practice, we almost always abbreviate the parentheses on the left-hand side, and we deal with statements like $X = 1$, $X = 2$, and so on.

Why do we call this function a random variable? There is nothing random (or stochastic) in the function that takes an event and gives a number, deterministically, for a particular event. Actually, the randomness comes from the experiment and the fact that the occurrence of the event in question is random. We will see this distinction reinforced when we think about distributions.

DISTRIBUTIONS

We now have two main concepts in this section – probability and random variables. Given a sample space S and the same experiment, both probability and random variables take events as their inputs. But they output different things (probabilities measure the "size" of events, random variables give a number in a way that the analyst chose to define the random variable). How do the two concepts relate?

The concept of distributions is the natural bridge between these two concepts.

Definition 9 (Distribution of a random variable)**.** A distribution of a random variable is a function that specifies the probabilities of all events associated with that random variable. There are several types of distributions: A probability mass function for a discrete random variable and probability density function for a continuous random variable. ■

Notice how the definition of distributions combines two ideas of random variables and probabilities of events. First, the distribution considers a random variable, call it X. X can take a number of possible numeric values. Recall that with each of these numerical values there is a class of *events*. In the previous example, for $X = 3$ there is one outcome (R_1, R_2, R_3) and for $X = 1$ there are multiple $(\{(R_1, R_2^c, R_3^c), (R_1^c, R_2, R_3^c), (R_1^c, R_2^c, R_3), \})$. Now, the thing to notice here is that each of these events naturally come with a probability associated with them. That is, $P(R_1, R_2, R_3)$ is a number from 0 to 1, as is $P(R_1, R_2^c, R_3^c)$. These all have probabilities because they are in the sample space S. The function that tells us these probabilities that are associated with a numerical value of a random variable is called a distribution.
In other words, a random variable X *induces a probability distribution* P (sometimes written P_X to emphasize that the probability density is about the r.v. X)
How we express a distribution depends on whether its random variable are discrete or continuous.

- Discrete r.v.'s are those whose possible values are finite and countable[8].
- Continuous r.v.'s are those whose possibles values are uncountably infinite.

DISCRETE WORLD

Most real-world things are discrete, because the physical world is countable. For example, the number of atoms that compose the planet earth is a very large number but countable. Yet many times we often opt to conceptualize a variable as a continuous one because making such an assumption will give us nice mathematical properties, which we will touch on later. Thinking some aspects of the social world – such as income and age – as continuous might also be natural.
For a discrete random variable, its distribution can be defined and written as the probability of any particular event (call it x) happening:

$$P(X = x)$$

The fact that X is discrete means that for any x whose associated events are in the sample space, the probability is not zero. This statement sounds trivial, but we will see it does not hold in the same way for continuous distributions. We refer to the distribution with the symbol f and treat it as a function

$$f(x) = P(X = x), X \text{ is discrete}$$

and we call this distribution f a probability mass function (pmf).
Another type of distribution is the *cumulative* distribution:

[8] or countably infinite

Definition 10 (Cumulative Distribution Function)**.** The Cumulative distribution function (cdf), often denoted F, is a function that takes a value of a random variable and outputs the probability that the random variable takes on a value x that is **at most** some particular value x.

$$F(x) = P(X \leq x)$$

■

By the additivity axiom of probability, for a discrete X we know that we can rewrite this as a sum of probabilities of each possible outcome of X that is less than x:

$$P(X \leq x) = \sum_{a \leq x} P(X = a)$$

where the summation symbol with subscript means to apply the sum for any a such that $a \leq x$. The letter a seems to have come from nowhere; this is sometimes called a "book-keeping" variable meant only to track which element of the possible values of X we are on. It is introduced to provide the math for the English, and has no important interpretation of its own.

The cumulative distribution function is useful when our probability of interest is cumulative. For example, we might be interested in the probability that a mother has 2 children or less. If X is the number of children a mother has and its cdf is called F, then our probability of interest can be expressed as a cdf $F(x = 3)$.

Example 6 (Grade Distribution)**.** One common example that fits the motivation of a r.v. well is grading in U.S. systems. The letter grade comes with a particular number. A is given a 4.0 and $A-$ is given a 3.7, and so on. How we got to these numbers is somewhat arbitrary, but transferring these events to numbers has a practical use: We can average over numbers to compare students, which would have been tedious to do with only piles of letter-grades.

Suppose a school defines the numerical summary of a grade in the following way, and also determines a grading distribution as follows:[9]

Event	X(Event)	$P(X = x)$
A	4.0	c
A-	3.7	0.25
B+	3.3	0.10
B	3	0.05
B-	2.7	0.04
C+	2.3	0.02
C	2	0.02
C	1.7	0.02

1. Suppose the value of c is hidden from you. What is c?

[9] Distribution inspired by "Substantiating Fears of Grade Inflation, Dean Says Median Grade at Harvard College Is A-, Most Common Grade Is A". *Harvard Crimson*, December 3, 2013. http://www.thecrimson.com/article/2013/12/3/grade-inflation-mode-a/

Answer Probabilities in a PMF must sum to 1, so c is 1 minus the other probabilities: $c = 0.5$
2. If $X \sim F_X$, where F_X is X's CDF, what is $F(3.5)$?
Answer $F(3.5) = P(X \leq 3.5) = P(X = 3.3) + P(X = 3) + ...P(X = 0.7)$. Note that we can consider values of X that are not on the table but their probabilities would be 0 so it wouldn't affect the value of $F(3.5)$, which is 0.25.
3. What is the expected value of X, or $E(X)$?
Answer The expectation $E(X)$ is a weighted average. Which means

$$E(X) = \sum_x P(X = x) \times x = 0.7 \times 0.02 + 1 \times 0.02 + \cdots 0.5 \times 5 = 3.633$$

For known distributions their is a simpler formula for $E(X)$. But for any discrete r.v., we could calculate $E(X)$ in this manner.
Final note: This example is a case in which the probabilities associated with an event is only expressible in a table. There is no one-line formula, at least at first glance, to get from the values of $X(s)$ to $P(X = x)$. The distributions that we will see in the next section have specific formulas.

continuous world

The cdf is also useful as scaffolding to introduce the distribution of a *continuous* random variable. We cannot define the distribution induced by a continuous random variable in the same way we did for discrete random variables. This is because that for any particular x, the quantity $P(X = x)$ is always zero due to the fact that the sample space of a continuous random variable is infinite (intuitively, because the values the r.v. takes are infinitely many).
But the intuition for trying $P(X = x)$ is not far off. We want to convey that when we use the following definition of a distribution function for a continuous r.v.:

Definition 11 (Probability density function)**.** The probability density function (pdf), often denoted f, satisfies the property

$$P(a < X < b) = F(b) - F(a) = \int_a^b f(x)dx$$

where F is the CDF of X. Another definition of f is as the derivative of F,

$$f(x) = F'(x)$$

where the dash or "prime" symbol denotes derivation with respect to X. The two equations will be equivalent due to the fundamental theorems of calculus. ■

Common Error 4 (PDF)**.** The distribution of a continuous r.v. X is not $P(X = x)$. In fact, $P(X = x) = 0$ everywhere. Instead, the pdf $f(x)$ is $f(x) = F'(x)$ where F is the cdf, i.e. $F(x) = P(X \leq x)$. However, like a probability $f(x)$ does indicate the likelihood of a realization of the r.v. ("density"). ■

This is bit of a mouthful, but we should still keep the intuition that for a continuous r.v., the pdf f indicates some measure of the likelihood of a particular event happening. In calculus terms, the area under F between endpoints a and b will give the probability. For any single value of X the probability $P(X = x) = 0$, but $f(x)$ will not necessarily be 0. Only when x is not a possible value of X then would $f(x) = 0$, otherwise it would give some non-negative number. We should be careful not to call $f(x)$ a probability, but we call it a density to preserve its meaning of likelihood and also to connect it with the area under the curve conceptualization.
For a continuous r.v. the cumulative version can be defined similarly,

Definition 12 (Cumulative Density function)**.** For a continuous r.v. X, the cumulative density function is the probability that X takes on value that is at most some particular value x

$$F(x) = P(X \leq x)$$

now, unlike the discrete case, P is an integral, so F can be also be written as

$$F(x) = \int_{\infty}^{x} f(u)du$$

where u is simply a book-keeping variable. ■

Finally, we use the tilde symbol to denote that a particular r.v. follows a particular pdf/pmf, i.e. "is distributed" according to that distribution:

$$X \sim f$$

Why bother going through all these definitions? The nice thing about distributions is that some types of functions come up in the natural and social world over and over again. By picking up on these common distributions and studying its mathematical properties, analysts can make statements about any event associated with that random variable. Additionally, some summary statistics of random variables (such as the mean, median, mode) depend only on the *distribution* of the random variable. So handling distributions can simplify the problem without losing any useful information.

TYPES OF DISTRIBUTIONS

These common distributions include the following:
For discrete r.v.'s:

- Bernoulli: If $X \sim$ Bernoulli(p), $f(x = 1) = p$, $f(x = 0) = 1 - p$.
- Binomial: If $X \sim$ Bin(n, p), $f(x) = \binom{n}{x} p^x (1 - p)^{n-x}$.
- Poisson: If $X \sim$ Poisson(λ), $f(x) = e^{-\lambda} \frac{\lambda^x}{x!}$

For continuous r.v.'s

- Uniform: If $X \sim$ Unif(a, b), $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, $f(x) = 0$ for all other values of x.
- Normal: If $X \sim N(\mu, \sigma^2)$, $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Let's work on some with examples.

BERNOULLI AND BINOMIAL

A Binomial is a sum of Bernoullis. A Bernoulli with parameter is a single coin toss, where the "success" occurs with probability p. A Binomial is different from a Bernoulli in two ways: (1) it involves any number of Bernoullis, and (2) it counts the *total* number of successes.

Example 7 (Total Delays)**.** In a class of 30 students, each student has 0.1 probability of arriving late to class. X is the total students who arrive late on a given date.

1. What is the distribution of X?
Answer By the setup, we realize this is a Binomial. There are 30 "trials", so the n parameter is 30. Each trial's success rate is 0.1, so the p parameter is 0.1.

$$X \sim \text{Binomial}(n = 30, p = 0.1)$$

2. What is the PMF of X?
Answer A PMF is a type of distribution but explicitly asks you for a function.

$$P(X = x) = \binom{30}{x}(0.1)^x(0.9)^{30-x}$$

3. Define a new r.v., X_1, which is 1 if the Sam (a student) is late and 0 if he is not late. What is the distribution of X_1? What is the PMF of X_1?
Answer $X_1 \sim \text{Bernoulli}(0.1)$ and $P(X_1 = x) = (0.1)^x(0.9)^{1-x}$. It is important to note that a Binomial is simply a sum of n Bernoullis.

3. What is $E(X)$?
Answer You *could* compute $\sum_{x=0,1,\ldots,30}\binom{30}{x}(0.1)^x(0.9)^{30-x}$. But the expectation of known r.v.'s such as the Binomial is derived for you, and in this case has an intuitive feel to it. $E(X) = np = 3$.

4. What is the probability that more than 27 people arrive *on time*?
Answer That's the probability that either 0, 1, or 2 people arrive late. The probability of this union is the sum of probabilities, because events are disjoint. So

$$P(\text{27 or more people are on time}) = \sum_{x=0}^{2}\binom{30}{x}(0.1)^x(0.9)^{30-x} = 0.41$$

Using R, \

```
pbinom(q = 2, size = 30, prob = 0.1)
```

```
## [1] 0.4113512
```

■

POISSON

A Poisson r.v. has the PMF,

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

λ is a constant that quantifies the rate of change. It could be 0.1, 1, or 1000, etc..e is Euler's constant, 2.718282...

Example 8 (Poisson PMF)**.** Let $X \sim$ Poisson(λ). Fill in the following cells.

x	$P(X = x)$	$P(X \leq x)$
0		
1		
2		

Answer This is largely a plug-in and simplify exercise.

x	$P(X = x)$	$P(X \leq x)$
0	$\frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}$	$e^{-\lambda}$
1	$\frac{e^{-\lambda}\lambda^1}{1!} = e^{-\lambda}\lambda$	$e^{-\lambda}(1 + \lambda)$
2	$\frac{e^{-\lambda}\lambda^2}{2!} = \frac{1}{2}e^{-\lambda}\lambda^2$	$e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2}\right)$

The interesting point about a Poisson is that x is countably infinite. That is we can compute quantities like $P(X = 99999999)$ and keep on going. In a Binomial, we had to stop at n. If X can take on infinitely many values and this probability table will go on for ever, you might doubt whether the PMF will sum to 1. After all, we are summing an infinite amount of things. However, it does not. The sum of this sequence actually converges to 1.

■

UNIFORM DISTRIBUTION

When we say things like "completely random" or "completely arbitrary", we are often talking about a Uniform distribution. How do we represent this mathematically? The way to do this is to have the density be constant:

Definition 13 (Uniform)**.** If $X \sim$ Uniform(a, b), where a is the beginning point of the possible range and b is the end point of the possible range, then its PDF $f(x)$ is

$$f(x) = \frac{1}{b - a}\text{for } x \text{ between } a \text{ and } b.$$

■

Perhaps an easier way to remember this is that the PDF of a Uniform variable is always a rectangle, because the height across a through b is constant. The density is the height of this rectangle. What else do we know about this rectangle? Well, the width of the rectangle is $b - a$, because of the range. And the area of this rectangle has to be 1, like all densities. Thus, the height is the area (1) divided by the width $(b - a)$.

Example 9. A r.v. $V \sim \text{Uniform}(-2, 8)$. What is $P(0 < V < 7)$?

Answer 7/10, because we take the slice of the rectangle, so to speak, where the values are $0 < v < 7$. The entire width of the rectangle is from -2 to 8, so 10. The probability is 7 over 10, then. The fact that the probability range is not inclusive of the endpoints 0 and 7 should not affect your answer, because continuous r.v.'s have Probability 1 for any particular value (i.e., $P(V = 0) = 0, P(V = 7) = 0$, so $P(0 < V < 7) = P(0 \leq V \leq 7) = 0$).

■

NORMAL DISTRIBUTION

The Normal distribution has an intimidating PDF:

Definition 14 (Normal)**.** If $X \sim \text{Normal}(\mu, \sigma^2)$, where μ is the mean and σ^2 is the variance, its PDF $f(x)$ is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

A special type of Normal distribution is one with mean 0 and variance 1. We often refer to this as Z:

$$Z \sim N(\mu = 0, \sigma^2 = 1)$$

■

As the naming of the parameter suggests, we can show that $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$. For now, you can take these as given.

Even though its PDF looks daunting, Normal distributions come up everywhere. In inference, we will start with the Central Limit Theorem, which posits that the sum of most random variables will be distributed as a Normal distribution as the number of elements increase. This means that many phenomena we observe in real life have approximately a Normal distribution. The densities of a Normal PDF vary by their parameters, i.e. μ and σ^2. However, a useful fact is that any Normal r.v. has a straightforward transformation to our favorite type of Normal, a standard Normal. To get this, subtract μ from X and then divide that difference by σ.

$$\text{For any } X \sim \text{Normal}(\mu, \sigma^2), \frac{X - \mu}{\sigma} \sim Z$$

This standardization is useful when we want to compare two Normal distributions with different parameters.

Example 10 (SATs and ACTs)**.** The Table shows the mean and standard deviation for total scores on the SAT and ACT.

	SAT	ACT
Mean	1500	21
Variance	90000	25

The distribution of SAT and ACT scores are both nearly normal.
1. Suppose Ann scored 1800 on her SAT and Tom scored 24 on his ACT. Who performed "better"?
2. Determine the proportion of SAT test takers who scored better than Ann on the SAT.
Answer Standardize the two to the same distribution.

$$\frac{X-1500}{300} \sim Z, \; \frac{Y-21}{5} \sim Z$$

Then Ann's standardized score, z_A is $\frac{1800-1500}{300} = 1$, Tom's standardized score z_B is $\frac{24-21}{5} = \frac{3}{5}$. Because both are on the same scale, we can now compare the raw numbers. Ann's score is higher when standardized. In other words, if we put the two students along a standard Normal distribution, Ann would be placed to the right of Tom on the x-axis.
The second part asks us to find the probability $P(X > 1800)$. So let's start from the equation and re-write in terms of Z:

$$\begin{aligned}
&= P(X > 1800) \\
&= P\left(\frac{X-1500}{300} > \frac{1800-1500}{300}\right) \\
&= P\left(Z > 1\right) \\
&= P(Z < -1)
\end{aligned}$$

We could look up the Z-score table for standard Normal. Another rule of thumb is that in a Normal, one standard deviation below and above the man capture 68 percent of the area. Because the standard deviation of Z is 1, we could approximate $P(Z < -1) = (1 - 0.68)/2$. Alternatively, with a statistical program we could compute this value (which is a CDF evaluated at -1) directly:

```r
pnorm(q = -1, mean = 0, sd = 1)
```

```
## [1] 0.1586553
```

Either way, the answer is approximately 0.16.

■

EXPECTATION

A distribution tells us the probability that a random variable will fall in any given set, but sometimes we want just one numerical summary. This is the motivation for dealing with the mean, or expected value of a random variable.

Definition 15 (Expectation of a r.v.)**.** An expectation of a random variable is the weighted mean of all its possible values, where the weight for each value is set to the probability of the r.v. taking that particular value. It is denoted with E:
For a discrete r.v.,

$$E(X) = \sum_x \underbrace{x}_{\text{value}} \underbrace{P(X = x)}_{\text{weight}}$$

in other words, the weight given to a value is the pmf evaluated at each value.
In the continuous case, we need to think of densities rather than probabilities

$$E(X) = \int_{-\infty}^{\infty} \underbrace{x}_{\text{value}} \underbrace{f(x)}_{\text{weight}} \, dx$$

■

Intuitively the expected value is the center of mass of a distribution.
Expectation, as a function, has several properties due to the way it is built. The fundamental property is that *expecation is linear*, that is to say the expectation of a linear combination (i.e., addition and subtraction of scalar products) is the linear combination of expectations. Concretely, the simplest example is

$$E(X + X) = E(X) + E(X)$$

Similarly we can continue on to say that $E(X+X+X+X) = E(X)+E(X)+E(X)+E(X)$, and because $E(X) + E(X)$ is clearly $2E(X)$, we can say that for any constant c,

$$E(cX) = cE(X)$$

we may refer to this procedure as taking the constant "out" of the expectation. Often we are told what $E(X)$ is but don't know what the expectation of some complicated function of X is; taking out terms out of the expectation simplifies our question.

Common Error 5 (Category Error for E)**.** Note that the expectation is a number, not a random variable (remember: a r.v. is a function, which in turn gives a number). There is variability on what a random variable X can take on, but there is no variability of $E(X)$ because we have summed over all the possibilities and generated just one number. ■

VARIANCE

Two random variables might have the same expectation, or central tendency, yet have very different properties. In particular, one might have very "similar" values occur with high probability, and the other may have less similar values occur with decent probability. In other words, there is more variability in the experiment, from the perspective of our random variable, in the latter. Measuring this uncertainty so we can compare the reliability of a r.v.'s properties is the motivation for variance.

Definition 16 (Variance of a r.v.)**.** The variance, denoted Var, of a random variable is

$$\text{Var}(X) = E\left[(X - E(X))^2\right]$$

■

This definition is consistent with our conception of "spread", because the inner term $X - E(X)$ measures how much a value of a r.v. is far from its central tendency, squaring the distance transforms everything into positive values like distance[10], and then we summarize that variability by an expectation.
Sometimes evaluating $E\left[(X - E(X))^2\right]$ mathematically by each step is not possible. There are many ways to rewrite this variance.

Example 11 (Alternative formulation for variance)**.** Show that $\text{Var}(X) = E\left(X^2\right) - (E(X))^2$. Starting from the definition of variance,

$$\begin{aligned}
\text{Var}(X) &= E\left[(X - E(X))^2\right] \\
&= E\left[X^2 - 2XE(X) + E(X)^2\right] \\
&= E\left(X^2\right) - E[2XE(X)] + E[E(X)^2] \because \text{Linearity of expectation} \\
&= E\left(X^2\right) - 2E(X)E(X) + (E(X))^2 \because \text{Expectation is a constant} \\
&= E(X^2) - (E(X))^2
\end{aligned}$$

■

Common Error 6 (Category Error for Var)**.** Similar to expectation, the variance of a random variable is a number, not a random variable. ■

Common Error 7 (Subtracting two different variables does not reduce its variance)**.** It might sound intuitive that if you take one r.v. (that is noisy) and subtract a second r.v. (also noisy) from it, the difference will be less noisy than the individual parts, i.e., the variance decreases. This is wrong – combining two random variables either by addition or subtraction will still increase the variance (if the two are sufficiently independent). Thus,

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$$

The correct intuition is that subtracting a noisy variable from another noisy variable does not magically erase the noise.

[10] there are a couple of reasons why we chose not to take the absolute value rather than squaring it. One of them is that the quantity $E(X^2)$ has nice theoretical properties that allow analysts to conduct more reliable inference.

But notice that there is a covariance term in this formulation, which might affect your overall results. As Cov(X, Y) becomes large, the Var$(X - Y)$ will get smaller and smaller relative to Var(X) + Var(Y). At the extreme, if X and Y are the same variable, they will be perfectly correlated, so

$$\begin{aligned}\text{Var}(X - X) &= \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, X)\\ &= 0 \ \because \text{Cov}(X, X) = \text{Var}(X).\end{aligned}$$

■

Means and Variances allow us to expands our horizon by allowing us to speak to different functions of our simple random variables, as in this example.

Example 12 (Sums of r.v.'s)**.** A family on vacation in the US will pay on average $247 per day, with a standard deviation of $60 per day. Assume this is normally distributed. The amount a family on vacation in Europe will pay per day is also normally distributed with a mean of $240 and a standard deviation of $50 per day. What is the probability that a five-day vacation in Europe would cost a family less than a five-day vacation in the US?

Answer

Let X_1 = cost per US day. Define X_2, X_3, X_4, and X_5 similarly for days 2 through 5. Same with Y_i but for Europe.

$$\begin{aligned}X_i &\sim \text{Normal}(\mu = 247, \sigma = 60) \text{ for } i = 1, 2, 3, 4, 5\\ Y_i &\sim \text{Normal}(\mu = 240, \sigma = 50) \text{ for } i = 1, 2, 3, 4, 5\end{aligned}$$

It is often useful to introduce your own notation for quantities of interest. Even though this adds more letters to your solution, it reduces your keystrokes and simplifies ideas. Let U = total for a five-day vacation in the US., V for Europe.

$$\begin{aligned}U &= X_1 + X_2 + X_3 + X_4 + X_5\\ V &= Y_1 + Y_2 + Y_3 + Y_4 + Y_5\end{aligned}$$

To find $P(V < U)$, characterize the distributions of V and U. Assume that (1) the costs day to day are independent of other days, and (2) accept as truth that a linear combinations of Normals is also Normal.

$$X_i \sim \text{Normal}(\mu = 247, \sigma = 60, Y_i \sim \text{Normal}(\mu = 240, \sigma = 50)$$

Expectation is linear, so

$$\begin{aligned}E(U) &= E(X_1 + X_2 + X_3 + X_4 + X_5)\\ &= 5 \cdot E(X_i]\\ &= 5 \cdot 247\\ &= 1235\end{aligned}$$

Variance is not linear, but here $X_1, X_2, ..., X_5$ are independent so the variance of sum is sum of variance.

$$\begin{aligned}\text{Var}(U) &= \text{Var}(X_1 + X_2 + X_3 + X_4 + X_5)\\ &= \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + \text{Var}(X_4) + \text{Var}(X_5)\\ &= 5 \cdot \text{Var}[X_1] \quad \because \text{all } X_i\text{'s have the same variance}\\ &= 5 \cdot 60^2\end{aligned}$$

So

$$E(U) = 1235, SD(U) = 134.16$$

Similarly, doing calculations will give

$$E(V) = 1200, SD(V) = 111.80$$

Now, because U and V are linear combinations of normal variables, they themselves are normally distributed:

$$\begin{aligned}U &\sim \text{Normal}(\mu = 1235, \sigma = 134.16)\\ V &\sim \text{Normal}(\mu = 1200, \sigma = 111.80)\end{aligned}$$

Remember, we are trying to find $P(V < U)$. Let's define a new variable to help us do this: Define $D = V - U$. If we can characterize the distribution of D, then we can calculate $P(D < 0) = P(V < U)$.
So, let's find the expected value and standard deviation of D. Remember that U and V are independent.
$E(D) = E(V - U) = E(V) - E(U) = 1200 - 1235 = -35$

$$\begin{aligned}\text{Var}(D) &= \text{Var}(V - U)\\ &= \text{Var}(V - U)\\ &= \text{Var}(U) + \text{Var}(V)\\ &= 134.16^2 + 111.80^2\end{aligned}$$

We now know that

$$D \sim \text{Normal}(\mu = -35, \sigma = 174.64).$$

Now we are finally ready to calculate $P(V < U)$

$$\begin{aligned}P(V < U) &= P(D < 0)\\ &= P\left(Z < \frac{0 + 35}{174.64}\right)\\ &= P(Z < 0.20)\\ &= 0.579\end{aligned}$$

So, there is a 57.9% chance that a five-day vacation in Europe will cost less than a five-day vacation in the US.

■

applications: markov chains (extra)

Many real-world phenomena are dependent in systematic ways – most obviously, present status depends in some way on past status. Considering problems where data comes from an independently draw sequence of random variables. Dependent events are hard to model, because the joint probability of a set of events happening is no longer the product of the single components. Intuitively, recall how the variance of the sum of two random variables was not just the sum of each of the variances, but includes a covariance term. Complexity adds up when we start to look at dependent variables jointly.

A Markov chain is an attempt to capture that complexity, but with enough simplicity so that we can make general probabilistic claims.

Definition 17 (Markov Property and Markov Chains)**.** A sequence of random variables, $X_1, ...X_n$ is said to have the Markov property if for each of random variables, the "future" state (the subsequent X_i) depends only on the present state, i.e. is independent of past states conditional on the past states:

That is for a given n,

$$\begin{aligned} &P(X_{n+1} = j \mid X_1 = i_1, ...X_{n-1} = i_{n-1}, X_n = i) \\ =&P(X_{n+1} = j \mid X_n = i) \quad \text{i.e. the values from } 1, ...n-1 \text{ don't add any new information} \end{aligned}$$

A sequence of random variables that have the Markov property is simply called a Markov Chain.

■

That is, we allow for one level of dependency in a sequence of random variables. This is probably not enough to be a true representation of the world — past events from varying time influence our current state — but it is much better than the completely independent world we would have dealt with otherwise.

The complexity introduced is still simple enough that we can express our probabilities neatly into writing. Because we need to worry about one step across two time periods, we deal with $k \times k$ probabilities, where k is the number of possible states. This leads us to organize them into a table, or matrix:

Definition 18 (Transition probability matrix)**.** For a Markov Chain $X_1, ...X_n$, define the probability

$$P_{i,j} = P(X_{n+1} = j \mid X_n = i)$$

that is, $P_{i,j}$ is the probability that an observation that is in state i at time n is in state j in the next time period. Notice that we don't need to consider any further history before n because of the Markov Property.

Then the matrix P is a way to store each of the possible $P_{i,j}$ terms, for every combination of i and j:

$$\mathbf{P} = \begin{pmatrix} P_{1,1} & P_{1,2} & P_{1,3} & \cdots \\ P_{2,1} & P_{2,2} & P_{2,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

■

The transition probability matrix contains information about the movement in *one* time step. But it can be easily expanded to quantify movement across *multiple* time steps, by matrix multiplication.

$$\text{Probabilities for exactly } t \text{ steps} = \underbrace{\mathbf{PP}\cdots\mathbf{P}}_{\text{multiply } t \text{ times}}$$

After many times, the product will eventually become something like:

$$\begin{pmatrix} \pi_1 & \pi_2 & \pi_3 & \cdots \\ \pi_1 & \pi_2 & \pi_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

That is, the values of the probabilities in each column will eventually become the same number. Intuitively, after many turns of systematic status change, your original position at $t = 0$ becomes irrelevant.

The probabilities $\pi_1, ...\pi_k$ is a quantity of interest; it quantifies the probability that an observation in the Markov Chain ends up in a statuses $1, ...k$ after a number of changes. It is possible to compute what the values of these are just from the probability transition matrix (which contains information about one step).

In a simple two-state example, the probabilities are

$$\mathbf{P} = \begin{pmatrix} P_{1,1} & P_{1,2} \\ P_{2,1} & P_{2,2} \end{pmatrix}, \ \text{Steady state} = \begin{pmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{pmatrix}$$

One more iteration of the Markov chain should not change the probabilities, so

$$\begin{pmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{pmatrix} \begin{pmatrix} P_{1,1} & P_{1,2} \\ P_{2,1} & P_{2,2} \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{pmatrix}$$

From matrix algebra, this equation indicates a set of qualities, with two unknowns (π_1, π_2) and know values $(P_{1,1}, P_{1,2}, P_{2,1}, P_{2,2})$

$$\begin{aligned} P_{1,1}\pi_1 + P_{2,1}\pi_2 &= \pi_1 \\ P_{2,1}\pi_1 + P_{2,2}\pi_2 &= \pi_2 \\ \pi_1 + \pi_2 &= 1 \end{aligned}$$

and re-arranging terms we can figure out what π_1 and π_2 need to be in order for these equations to hold.

Example 13. We construct a simple model of class mobility by setting up a Markov Chain. In this model, each citizen is in one of three classes – lower, middle, and upper. Each time point is generational turnover. Suppose economic mobility works in the following way:

- If a citizen is in the lower class (L), with probability 0.8 he stays in the lower class in the next period, with probability 0.1 he moves up to the middle class in the next period, and with probability 0.1 he moves up to the upper class in the next period.
- If a citizen is in the middle class (M), with probability 0.6 she stays in the middle class in the next period, with probability 0.2 she moves up to the upper class in the next period, and with probability 0.2 she drops down to the lower class in the next period.
- If a citizen is in the upper class (U), with probability 0.5 he stays in the upper class in the next period, with probability 0.4 he drops down to the middle class in the next period, and with probability 0.1 he drops down to the lower class in the next period.
- Conditional on the citizen's current class, the probability that she is at any given class in the next step is independent of her past history.

(a) Write down the probability transition matrix of this Markov chain. Recall that a probability transition matrix is a matrix where the cell ij (row i column j) holds the value

$$P_{ij} = P(\text{moving from state } i \text{ to state } j)$$

.

Answer

$$\mathbf{P} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.4 & 0.5 \end{pmatrix}$$

(b) Denote as π_i the probability that a citizen will end up in the class i after many generational turnovers (the steady-state probability). Write the system of equations that will yield the steady-state probabilities of a citizen being in the lower class (π_L) , in the middle class (π_M) and in the upper class (π_U).

Answer

$$\begin{aligned} \pi_L &= 0.8\pi_L + 0.2\pi_M + 0.1\pi_U \\ \pi_M &= 0.1\pi_L + 0.6\pi_M + 0.4\pi_U \\ \pi_U &= 0.1\pi_L + 0.2\pi_M + 0.5\pi_U \\ \pi_L + \pi_M + \pi_U &= 1 \end{aligned}$$

(c) Compute steady-state probabilities of a citizen being in the lower class(π_L), in the middle class (π_M) or in the upper class (π_U). Remember that the steady-state probabilities sum to 1, $\pi_L + \pi_M + \pi_U = 1$. Your final answer should be in simple fractions.)

Answer Starting with

$$\begin{cases} \pi_L &= 0.8\pi_L + 0.2\pi_M + 0.1\pi_U \\ \pi_M &= 0.1\pi_L + 0.6\pi_M + 0.4\pi_U \\ \pi_U &= 0.1\pi_L + 0.2\pi_M + 0.5\pi_U \end{cases}$$

Aggregate the terms,

$$-0.2\pi_L + 0.2\pi_M + 0.1\pi_U = 0 \tag{1}$$

$$0.1\pi_L - 0.4\pi_M + 0.4\pi_U = 0 \tag{2}$$

$$0.1\pi_L + 0.2\pi_M - 0.5\pi_U = 0 \tag{3}$$

subtracting (2) and (3) give

$$-0.6\pi_M + 0.9\pi_U = 0$$

This means that

$$\pi_U = \frac{0.6}{0.9}\pi_M = \frac{2}{3}\pi_M \tag{4}$$

To use the information in (1), we want to reduce it in terms of π_M and π_U so we can cancel out one of them out with (4). So use the fact that the probabilities sum to 1 and rewrite (1) as

$$\begin{aligned} -0.2(1 - \pi_M - \pi_U) + 0.2\pi_M + 0.1\pi_U = 0 \\ \Rightarrow 0.4\pi_M + 0.3\pi_U - 0.2 = 0 \end{aligned}$$

Combined with (4), we now know that

$$\begin{aligned} 0.4\pi_M + 0.30\frac{2}{3}\pi_M - 0.2 = 0 \\ \Rightarrow 0.6\pi_M = 0.2 \\ \Rightarrow \pi_M = \frac{1}{3} \end{aligned}$$

By (4), $\pi_U = \frac{2}{9}$, and because the probabilities sum to 1,

$$\pi_L = 1 - \frac{1}{3} - \frac{2}{9} = \frac{4}{9}$$

So solving the system of equations gives

$$\pi_L = \frac{4}{9}, \;\; \pi_M = \frac{3}{9}, \;\; \pi_U = \frac{2}{9}$$

This is the predicted steady state income distribution – the plurality of people in the lower class, and about a third in the middle class.

■