Probability Lecture 1

MATH 113

Fall 2017

1 First Lecture

Definition 1.1. We think of a **sample space** S as the set of all possible outcomes of an 'experiment' or observation. An **outcome** is an element of the sample space.

Example 1.2. If we are rolling a 6-sided die, $S = \{1, 2, 3, 4, 5, 6\}$. If we are flipping a coin two times, $S = \{HH, HT, TH, TT\}$. If we are playing Minesweeper, $S = \{Die, LiveDie, LiveLiveDie, LiveLiveLiveDie, \ldots\}$.

Definition 1.3. An **event** E is a subset of the sample space, thought of as a collection of outcomes.

Example 1.4. When we are rolling a 6-sided die, if E is rolling an even number, then $E = \{2, 4, 6\}$. If $H = \{4, 5, 6\}$, then one way to describe H is 'rolling higher than 3.'

There may be more than one way to describe the same event, and the same description might correspond to different events if the sample space is different.
Since events are sets, we can do the usual things to them.

Definition 1.5. The **union** of two events A, B is the event $A \cup B$, which can be described as 'A or B.' The **intersection** of A, B is $A \cap B$, 'A and B.' The **complement** of A is A^c, 'not A,' or 'A doesn't happen.'

Definition 1.6. $\emptyset$ is called the **null event** (it never happens) and S is the **certain event** (it always happens).

Definition 1.7. Two events A, B are called **mutually exclusive** if $A \cap B = \emptyset$.

Definition 1.8. Given a sample space S, a **probability distribution** is a map

$$P : \{events\} \longrightarrow [0, 1]$$

such that

i) $P(S) = 1$, $P(\emptyset) = 0$

ii) If A and B are mutually exclusive, $P(A \cup B) = P(A) + P(B)$

We will usually call $P(E)$ the probability of E.

Definition 1.9. A sample space along with a probability distribution is called a **probability space**. If every outcome is equally likely, it is called a **uniform probability space**. In a uniform probability space where $|S| < \infty$, $P(E) = |E|/|S|$.

Some properties of probability distributions follow directly from set theory!

Proposition 1.10. *i) If $A \subseteq B$, $P(A) \leq P(B)$.*

ii) $P(A) = 1 - P(A^c)$

iii) $P(A \cup B) = P(A) + P(B) + P(A \cap B)$

iv) $P(A \cup B) + P(A^c \cap B^c) = 1$

v) $P(A \cap B) + P(A^c \cup B^c) = 1$

Here is how we might prove i):

Proof. Suppose $A \subseteq B$. Then $A \cap B \backslash A = \emptyset$. So $P(B) = P(A) + P(B \backslash A)$. But $P(B \backslash A) \geq 0$, so $P(B) \geq P(A)$. □

Example 1.11. Suppose we have a standard deck of 52 cards, with 13 cards of each suit: hearts ♡ and diamonds ♢ (both red), and clubs ♣ and spades ♠ (both black). Suppose we have shuffled the deck so that the cards are in random order, and we pick two cards off the top. What is the probability that the first two cards are both red?

Let's call R the event that the first two cards are red. The order of the cards is random, so any pair of cards is equally likely. Therefore $P(R) = |R|/|S|$. Here are two different ways to solve this problem; there are doubtless many more.

There are 52 possible first cards, and then 51 possible second cards, so the total number of outcomes is $52 \cdot 51$. There are 26 red cards, so there are $26 \cdot 25$ outcomes in R and $P(R) = \frac{26 \cdot 25}{52 \cdot 51} = \frac{25}{102}$.

Alternatively, there are $\binom{52}{2}$ ways to pick two distinct cards out of the deck. There are $\binom{26}{2}$ ways to pick red cards, so $P(R) = \frac{\binom{52}{2}}{\binom{26}{2}} = \frac{\frac{26!}{2!24!}}{\frac{52!}{2!50!}} = \frac{26 \cdot 25}{52 \cdot 51} = \frac{25}{102}$.

Just as in combinatorics, if you want to check your work, count it in two different ways and see if you get the same answer.

2 Second Lecture

Let's start with an example.

Example 2.1. Alisha and Bachir each sit in a row of 7 chairs, choosing their seats at random. What is the probability that they don't sit next to each other?

There are $7 \cdot 6$ ways to sit. We could count all the different ways to sit so that there is at least one seat in between them. If A is in the first or last spot, B has 5 choices for where to sit. Otherwise B has only 4 choices, since A plus one seat on each side takes away 3 out of the 7 spots. Therefore there are $2 \cdot 5 + 5 \cdot 4 = 30$ different ways for the pair to sit not next to each other, and the probability of them not sitting next to each other is $\frac{30}{7 \cdot 6} = \frac{5}{7}$.

Alternatively, it's perhaps easier to count the different ways for them to sit together and then take the complement. In this case, there are 6 ways we can choose a spot for the pair and 2 ways they can sit in that spot (AB or BA) so the probability we want is $1 - \frac{6 \cdot 2}{7 \cdot 6} = 1 - \frac{2}{7} = \frac{5}{7}$.

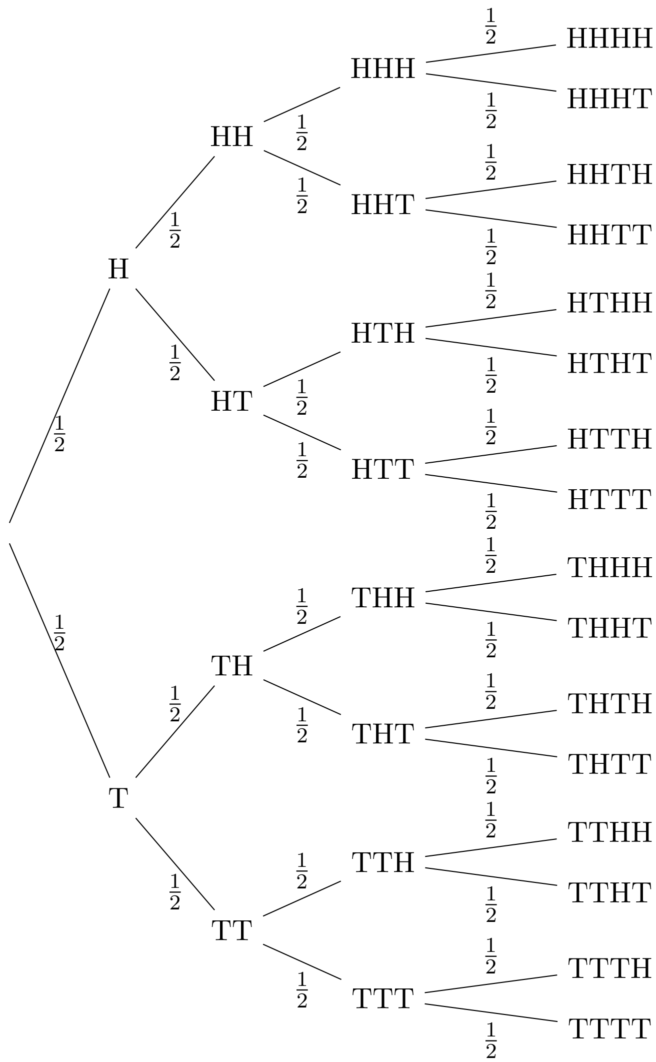
Keep in mind that it's sometimes easier to count a complement. This can be a good way to check your answer.

Definition 2.2. If $P(A \cap B) = P(A) \cdot P(B)$, then we call A and B **independent relative to** P. This is different than the colloquial meaning of independent. Unless independence is explicitly given in the problem, you have to prove it. Be suspicious of your intuition, because it is often wrong!

Example 2.3. Suppose we have an unfair coin, so the probability of flipping heads is always 0.75. What is the probability of getting 4 heads in a row? 4 tails in a row? exactly 2 heads out of 4 flips?

Notice this is NOT a uniform probability space. However, each flip has the same probability of being heads as the flip before it. Effectively, the problem as stated is asserting that flipping heads on the first, second, third, or fourth flip are all independent of each other.

We can model this with what I call a probability tree. This is just a visual organizer, not a mathematical object. It's not the only way to solve this, but I like it! Each level in the tree will be an independent event, with branches labelled with probability. To calculate, find the right leaves, multiply the probabilities that go down to those leaves, and add them all up.



$$P(HHHH) = \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} = \left(\frac{3}{4}\right)^4 = \frac{81}{256}$$

$$P(TTTT) = \left(\frac{1}{4}\right)^4 = \frac{1}{256}$$

$$\begin{aligned} P(HHTT) + P(HTHT) + P(HTTH) + P(THHT) + P(THTH) + P(TTHH) &= \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^2 \cdot 6 \\ &= \frac{54}{256} = \frac{27}{128} \end{aligned}$$

Notice that we could also compute the last probability by $\binom{4}{2}\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^2$.

Example 2.4. Suppose that we draw a number from the set $\{1, 2, \ldots, 49\}$ at random. Let F be 'picking a number divisible by 5' and let E be 'picking an even number.' Are these events independent?

We can construct a uniform probability space to solve this, where $F = \{5, 10, \ldots, 45\}$ and $E = \{2, 4, \ldots, 48\}$. Then $|S| = 49$, $|F| = \lfloor \frac{49}{5} \rfloor = 9$, $|E| = \lfloor \frac{49}{2} \rfloor = 24$, and $|F \cap E| = \lfloor \frac{49}{10} \rfloor = 4$, so $P(F) = \frac{9}{49}$, $P(E) = \frac{24}{49}$, $P(F) \cdot P(E) = \frac{9 \cdot 24}{49^2}$ and $P(F \cap E) = \frac{4}{49}$. But $\frac{4}{49} \neq \frac{9 \cdot 24}{49^2} = \frac{216}{2401}$, so these events are NOT independent.

This is how you prove that things are independent!

3 Third Lecture

Recall that events $A, B \subseteq S$ are *independent* when $P(A)P(B) = P(A \cap B)$. What's happening when events are *not* independent?

Definition 3.1. Let $A, B \subseteq S$ be events and assume $P(B) > 0$. Define $P(A|B) := P(A \cap B)/P(B)$. Then $P(A|B)$ is called a *conditional probability* and read "the probability of A given B."

Note that A and B with $P(B) > 0$ are independent if and only if $P(A|B) = P(A)$. Since $P(A|B)$ is the probability that A happens given that B happens, we see that A and B are independent when the occurrence of B does not make the occurrence of A any more or less likely.

Example 3.2. We toss a fair coin four times. We don't see the results, but someone who does truthfully tells us that at least two of the tosses were heads. What is the probability that all four tosses were heads?
To answer this question, we must find $P(A|B)$ where A is the event "all four tosses are heads" and B is the event "at least two tosses are heads." Note that $A \cap B = A$, so $P(A|B) = P(A)/P(B)$. Of course, $P(A) = (1/2)^4 = 1/16$. Meanwhile, B is the disjoint union of the events "exactly two heads," "exactly three heads," and A. Thus

$$P(B) = \frac{\binom{4}{2}}{16} + \frac{\binom{4}{3}}{16} + \frac{1}{16} = \frac{11}{16}.$$

We conclude that $P(A|B) = 1/11$.

Example 3.3. Let $\underline{n} = \{1, 2, \ldots, n\}$ and let $\pi : \underline{n} \to \underline{n}$ be a randomly selected permutation. Let A be the event that $\pi(1) > \pi(2)$. Let B be the event that $\pi(2) > \pi(3)$. What is $P(A|B)$? Are A and B independent events?
Clearly $P(A) = P(B) = 1/2$. Note that $A \cap B$ is the event that $\pi(1) > \pi(2) > \pi(3)$. Since there are $3! = 6$ orderings of 3 numbers, $P(A \cap B) = 1/6$. Thus $P(A|B) = P(A \cap B)/P(B) = 1/3$. Since $P(A) = 1/2 \neq 1/3$, we conclude that A and B are not independent.

It is relatively intuitive that the events of Example 3 are not independent. After all, if $\pi(2) > \pi(3)$, then $\pi(2)$ is "on the big side," so it will be harder for it to be smaller than $\pi(1)$. But be careful in applying this sort of reasoning. Intuition can easily lead us astray in probability theory, as the following example demonstrates.

Example 3.4. During the 2016 Renn Fayre softball tournament, Professor Fim Jix had a higher batting average than Professor Pavid Derkinson. The same is true of their batting averages during the 2017 tournament. Does it follow that Jix's cumulative 2016–17 batting average is higher than Derkinson's?
Counterintuitively – but unsurprisingly given the setup – the answer is NO, not necessarily. Indeed, consider the following statistics.

		2016	2017	2016–17
	hits	3	24	27
Jix	at bats	10	60	70
	average	.300	.400	.386
	hits	10	3	13
Derkinson	at bats	30	5	35
	average	.333	.600	.371

We see that Derkinson has higher batting averages each season, but Jix has the higher cumulative batting average!

This counterintuitive phenomenon is pervasive and important enough to merit a name: *Simpson's paradox.* Note that there is no real paradox here, only something that goes against our intuition. In order to put a finer point on how and why Simpson's paradox arises, we turn to Bayes' Theorem.

Theorem 3.5 (Bayes' Theorem)**.** *Let A and B be mutually exclusive events ($A \cap B = \varnothing$) such that $A \cup B = S$ and $P(A)P(B) > 0$. Then for any event C,*

$$P(C) = P(C|A)P(A) + P(C|B)P(B).$$

We can interpret this theorem as saying that the probability of C is the weighted average of its conditional probabilities. (Here $P(A)$ and $P(B)$ are the weights. Note that the hypotheses imply that $P(A) + P(B) = 1$, so this really makes sense as a weighted average.)

Proof. Note that $A \cap C$ and $B \cap C$ are disjoint and $(A \cap C) \cup (B \cap C) = C$. Thus $P(C) = P(C \cap A) + P(C \cap B)$. Meanwhile,

$$\begin{aligned} P(C|A)P(A) + P(C|B)P(B) &= \frac{P(C \cap A)}{P(A)}P(A) + \frac{P(C \cap B)}{P(B)}P(B) \\ &= P(C \cap A) + P(C \cap B). \end{aligned}$$

We conclude that the two quantities are equal. □

In the case of Example 5, we get the following clearer picture of our softball heroes' batting averages. Let Hit_J be the event of Jix getting a hit in 2016 or 2017 and similarly define Hit_D to be the event of Derkinson geting a hit in either season. Let J_{2016} denote Jix's at bats in 2016, and let J_{2017} denote his bats in 2017. Let D_{2016} denote Derkinson's at bats in 2016, similarly define D_{2017}. Then by Bayes' Theorem (moral exercise: check that the hypotheses hold!),

$$\begin{aligned} P(\mathrm{Hit}_J) &= P(\mathrm{Hit}_J\,|J_{2016})P(J_{2016}) + P(\mathrm{Hit}_J\,|J_{2017})P(J_{2017}), \text{ and} \\ P(\mathrm{Hit}_D) &= P(\mathrm{Hit}_D\,|D_{2016})P(D_{2016}) + P(\mathrm{Hit}_D\,|D_{2017})P(D_{2017}). \end{aligned}$$

In the setup of Example 5, we know that all of the "J" conditional probabilities are smaller than their matching "D" conditional probabilities, but we have no control over how the "weights" $P(J_{2016})$, *etc.* compare. It turns out they can spoil our intuition and result in the "paradox" of $P(\mathrm{Hit}_J) > P(\mathrm{Hit}_D)$.
We conclude this lecture by considering how to generalize independence and Bayes' Theorem when there are more than two events. For independence, the right generalization is the maximally strong one.

Definition 3.6. Events $A_1, \dots, A_n$ are *independent* if for any nonempty set $I = \{i_1, \dots, i_k\} \subseteq \{1, \dots, n\}$,

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}).$$

We get the following generalization of Bayes' Theorem via a completely analogous proof. (Moral exercise: check the details.)

Theorem 3.7. *Let $A_1, \dots, A_n$ be events in the same sample space S such that $A_1 \cup \cdots \cup A_n = S$ and $A_i \cap A_j = \varnothing$ for all $i \neq j$. Let $C \subseteq S$ be any event. Then*

$$P(C) = P(C|A_1)P(A_1) + \cdots + P(C|A_n)P(A_n).$$

4 Fourth Lecture

In this lecture, we will study *random variables* and *expected value*. By the end of it, we should be able to precisely formulate and answer questions such as "How much can I expect to win if I play the lottery?" and "What is the expected number of fixed points for a random permutation?" Throughout, P is a probability distribution on a finite sample space S.

Definition 4.1. A *random variable* is a function $X : S \to \mathbb{R}$.

In other words, a random variable is some way of assigning numbers to elements of a sample space. Note that we can add and multiply random variables X, Y on the same sample space, and we can also scale random variables by a real number. For $s \in S$ and $c \in \mathbb{R}$ these operations are given by the rules

$$\begin{aligned}(X+Y)(s) &= X(s) + Y(s),\\ (XY)(s) &= X(s)Y(s),\\ (cX)(s) &= c \cdot (X(s)).\end{aligned}$$

We can also assign an expected value (also called expectation, average value, or mean) to every random variable.

Definition 4.2. Let $X : S \to \mathbb{R}$ be a random variable and let $X(S) = \{X(s) \mid s \in S\}$ denote the image of X. Then the number

$$E(X) := \sum_{y \in X(S)} y \cdot P(X = y)$$

is called the *expected value* of X on S. Here $P(X = y)$ is shorthand for the probability of the event $\{s \in S \mid X(s) = y\}$, *i.e.* the event that random variable X takes the value y.

In other words, $E(X)$ is the weighted average of the values X takes, with weights given by the probability that X takes the corresponding value.

Example 4.3. A lottery offers \$1 tickets on which you choose six distinct numbers between 1 and 48, inclusive. The lottery announces winning numbers and if your ticket matches all the winning numbers (irrespective of order) you get \$1,000,000; otherwise you get nothing. Expected value allows us to at least partially answer the question "Should you play this lottery?"
Let S be the sample space of 6-element subsets of $\underline{48} = \{1, 2, \dots, 48\}$. Define $X : S \to \mathbb{R}$ such that $X(s) = -1$ is s is not the winning ticket (because you've then lost your \$1 investment) and $X(s) = 999\,999$ if s is the winning ticket (the million dollar prize minus the ticket cost). Then $X(S) = \{-1, 999\,999\}$ and the expected value of X is

$$E(X) = -1 \cdot \frac{\binom{48}{6} - 1}{\binom{48}{6}} + 999\,999 \cdot \frac{1}{\binom{48}{6}} \approx -0.918.$$

This means that if you play this lottery many many times, then in the long run you can expect to lose about 92 cents each time you play, so it's not a good investment.

Expected value has an unexpected property: *linearity*. For those who have experience with linear algebra, this literally means that E, as a function from the $\mathbb{R}$-vector space of random variables to $\mathbb{R}$, is a linear transformation. If you don't speak that language yet, consider the following simply stated theorem as a definition of the term.

Theorem 4.4. *Let $X, Y : S \to \mathbb{R}$ be random variables and let $c \in \mathbb{R}$. Then*

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

and

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

Linearity of expected value is an extremely powerful tool. We will use it to give a simple proof of the following remarkable fact.

Theorem 4.5. *The expected value of the number of fixed points in a randomly selected permutation of $\underline{n} = \{1, 2, \dots, n\}$ is 1.*

Proof. Recall that a permutation π has i as a fixed point if $\pi(i) = i$. For $1 \leq i \leq n$ and π a permutation of $\underline{n}$, let $X_i(\pi) = 1$ if $\pi(i) = i$ and let $X_i(\pi) = 0$ otherwise. Define $X := X_1 + X_2 + \cdots + X_n$. Then $X(\pi)$ is equal to the number of fixed points of π and we are trying to find $E(X)$. By linearity, it suffices to find $E(X_i)$ for each i and then add up the values.
For a random permutation π of $\underline{n}$, $\pi(i)$ is equally likely to take any of the values in $\underline{n}$. Thus $P(X_i = 1) = 1/n$ and $P(X_i = 0) = (n-1)/n$. As such,

$$E(X_i) = 1 \cdot \frac{1}{n} + 0 \cdot \frac{n-1}{n} = \frac{1}{n}$$

for each $1 \leq i \leq n$. Thus

$$E(X) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \frac{1}{n} = n \cdot \frac{1}{n} = 1.$$

□

Note that Theorem 6 holds for any natural number n, so we say that the expected number of fixed points of a permutation of a finite set is 1.

5 Fourth Lecture Supplement

In this supplement to Lecture 4, we'll look at another application of linearity of expectation, and then provide the promised proof of linearity. Recall that linearity is the following statement.

Theorem 5.1. *Let $X, Y : S \to \mathbb{R}$ be random variables and let $c \in \mathbb{R}$. Then*

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

and

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

Example 5.2. Consider the sample space $S = \underline{6} \times \underline{6}$ of two rolls of a fair 6-sided die. Define the random variable $X : S \to \mathbb{R}$ to be the sum of the two rolls. We will compute the expected value of X in two ways: first, via the definition of expectation, then via linearity of expectation.
The sum of two rolls is any integer between 2 and 12, inclusive, so $X(S) = \{2, 3, \dots, 12\}$. We need to compute $P(X = 2)$, $P(X = 3)$, $\dots$, $P(X = 12)$.

We can only have $X = 2$ if both rolls take the value 1, so $P(X = 2) = 1/6^2 = 1/36$. We can get $X = 3$ only with rolls $(1, 2)$ and $(2, 1)$, so $P(X = 3) = 2/36$. For $X = 4$ we have rolls $(1, 3)$, $(2, 2)$, $(3, 1)$, so $P(X = 4) = 3/36$. For $X = 5$ we have rolls $(1, 4)$, $(2, 3)$, $(3, 2)$, $(4, 1)$, so $P(X = 5) = 4/36$. For $X = 6$ we have rolls $(1, 5)$, $(2, 4)$, $(3, 3)$, $(4, 2)$, $(5, 1)$, so $P(X = 6) = 5/36$. For $X = 7$ we have rolls $(1, 6)$, $(2, 5)$, $(3, 4)$, $(4, 3)$, $(5, 2)$, $(6, 1)$, so $P(X = 7) = 6/36$. For $X = 8$ (now things get interesting), we have rolls $(2, 6)$, $(3, 5)$, $(4, 4)$, $(5, 3)$, $(6, 2)$, so $P(X = 8) = 5/36$. For $X = 9$ we have rolls $(3, 6)$, $(4, 5)$, $(5, 4)$, $(6, 3)$, so $P(X = 9) = 4/36$. For $X = 10$ we have rolls $(4, 6)$, $(5, 5)$, $(6, 4)$, so $P(X = 10) = 3/36$. For $X = 11$ we have rolls $(5, 6)$ and $(6, 5)$, so $P(X = 11) = 2/36$. Finally, for $X = 12$ we have the single roll $(6, 6)$ so $P(X = 12) = 1/36$. We conclude that

$$\begin{aligned}
E(X) &= 2\frac{1}{36} + 3\frac{2}{36} + 4\frac{3}{36} + 5\frac{4}{36} + 6\frac{5}{36} + 7\frac{6}{36} + 8\frac{5}{36} + 9\frac{4}{36} + 10\frac{3}{36} + 11\frac{2}{36} + 12\frac{1}{36} \\
&= \frac{252}{36} \\
&= 7.
\end{aligned}$$

Linearity provides a much less labor intensive way to compute the expected value of X. Define $X_1 : S \to \mathbb{R}$ to be the value of the first roll, and X_2 to be the value of the second role. Then $X = X_1 + X_2$, so $E(X) = E(X_1) + E(X_2)$. Since each roll is no different from the other, we have $E(X_1) = E(X_2)$, and thus $E(X) = 2E(X_1)$. Now it is quite easy to compute $E(X_1)$ since $P(X_1 = 1) = P(X_1 = 2) = \cdots = P(X_1 = 6) = 1/6$. Thus

$$\begin{aligned}
E(X_1) &= 1\frac{1}{6} + 2\frac{1}{6} + \cdots + 6\frac{1}{6} \\
&= \frac{1 + 2 + \cdots + 6}{6} \\
&= \frac{6 \cdot 7/2}{6} \\
&= \frac{7}{2}.
\end{aligned}$$

We conclude that $E(X) = 2 \cdot 7/2 = 7$.

We now proceed to the proof of Theorem 1 for which we will need the following equivalent formulation of expected value.

Lemma 5.3. *If $X : S \to \mathbb{R}$ is a random variable, then*

$$E(X) = \sum_{s \in S} X(s)P(s).$$

(Here we are abusing notation and writing $P(s)$ for $P(\{s\})$.)

Proof. For each $y \in X(S)$, let $X^{-1}y := \{s \in S \mid X(s) = y\}$. Then

$$\begin{aligned}
\sum_{s \in S} X(s)P(s) &= \sum_{y \in X(S)} \sum_{s \in X^{-1}y} X(s)P(s) && \text{(grouping like terms)} \\
&= \sum_{y \in X(S)} \sum_{s \in X^{-1}y} yP(s) && \text{(since } X(s) = y \text{ for } s \in X^{-1}y) \\
&= \sum_{y \in X(S)} y \sum_{s \in X^{-1}y} P(s) && \text{(factoring).}
\end{aligned}$$

It remains to show that $\sum_{s \in X^{-1}y} P(s) = P(X = y)$, but this follows from the axioms for a probability distribution since $\bigcup_{s \in X^{-1}y} \{s\}$ is a partition of the event $\{s \in S \mid X(s) = y\}$. $\square$

Proof of Theorem 1. Given the lemma, the proof is an exercise is tracing through definitions. We will prove the first statement and leave the second one as a moral exercise for the reader.
We have

$$\begin{aligned} E(X+Y) &= \sum_{s\in S}(X+Y)(s)P(s) && \text{(Lemma 3)} \\ &= \sum_{s\in S} X(s)P(s) + \sum_{s\in S} Y(s)P(s) && \text{(definition of } X+Y \text{ and distribution)} \\ &= E(X)+E(Y) && \text{(Lemma 3 twice)}, \end{aligned}$$

as desired. □

6 Fifth Lecture

Remember that a random variable $X : S \to \mathbb{R}$ assigns a value to each outcome in a sample space. Say we're running an experiment, and all we care about is whether it succeeds or not. We can model this with a **Bernoulli random variable** X, where $X = 1$ if the experiment is a success and $X = 0$ otherwise. In this case $P(X = 1)$ is usally denoted p and $P(X = 0)$ as $q = 1 - p$.
If we do a sequence of independent experiments, each of which results in success with probability p and failure with probability $q = 1 - p$, and we are interested in the number of successes we can model this with a **binomial random variable**.

Example 6.1. We have a (possibly unfair) coin, which lands on heads with probability p and tails with probability q. If I flip the coin 3 times, what is the probability of getting exactly two heads? Let X be the number of heads out of 3 flips. Then

$$P(X = 2) = p\cdot p\cdot q + p\cdot q\cdot p + q\cdot p\cdot p = \binom{3}{2}p^2q.$$

This is why X is called a binomial random variable. If instead I flip the coin n times, the probability of getting exactly k heads is

$$P(X = k) = \binom{n}{k}p^kq^{n-k}.$$

Additionally, notice that

$$\sum_{k=0}^{n} P(X = k) = \sum_{k=0}^{n}\binom{n}{k}p^kq^{n-k} = (p + q)^n = 1$$

by the Binomial Coefficient Theorem, so all the probabilities sum to 1 as we expect.
To find the expected number of heads after n flips, we can make our lives easier by using the linearity trick. $X = I_1 + I_2 + \ldots + I_n$ where

$$I_j = \begin{cases} 1 & \text{if the coin is heads on the } j\text{th flip} \\ 0 & \text{otherwise} \end{cases}$$

These I_j are called **indicator random variables** because they indicate when a certain condition is met. Then for any j,

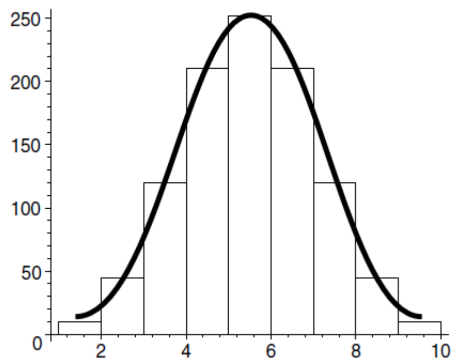
$$E[I_j] = 0\cdot P(I_j = 0) + 1\cdot P(I_j = 1) = p$$

so

$$E[X] = E[I_1] + E[I_2] + \ldots + E[I_n] = np.$$

If we graph the probabilities associated with a binomial random variable, they have a particular shape.

Example 6.2. If $n = 10$ and $p = \frac{1}{2}$, then $P(X = k) = \binom{10}{k}\left(\frac{1}{2}\right)^{10}$.



Graph of $\binom{10}{k} = P(X = k) \cdot 2^{10}$

As n gets bigger, this approaches a bell curve, or Gaussian distribution. It is appropriate to approximate the probability distribution of a binomial random variable with a Gaussian distribution if n is large enough (usually when np and nq are both significantly larger than 10).

If we again run a series of independent experiments, but we are interested in the number of attempts needed to obtain the first success, we can model this with a **geometric random variable** X, where $X = k$ means that it takes k trials for the first success. Since succeeding for the first time on the kth try means failing on all tries up to $k-1$, $P(X = k) = q^{k-1}p$. Do all these probabilities still sum to 1?
You may have seem the trick in 112

$$\sum_{i=0}^{\infty} r^i = 1 + r + r^2 + r^3 + \ldots = \frac{1}{1-r} \qquad \text{if } |r| < 1.$$

Notice that

$$\sum_{k=1}^{\infty} P(X = k) = \sum_{k=1}^{\infty} q^{k-1}p = p + qp + q^2p + \ldots = p(1 + q + q^2 + \ldots) = p\left(\frac{1}{1-q}\right) = \frac{p}{p} = 1$$

Example 6.3. We have a fair twenty-sided die. What is the probability that I roll a critical hit (20 on the die) within 6 rolls?
This is $P(X \leq 6)$ where $p = 1/20$ and $q = 19/20$. Then

$$\begin{aligned} P(X \leq 6) &= P(X = 1) + P(X = 2) + \ldots + P(X = 6) \\ &= \frac{1}{20} + \frac{19}{20} \cdot \frac{1}{20} + \left(\frac{19}{20}\right)^2 \cdot \frac{1}{20} + \left(\frac{19}{20}\right)^3 \cdot \frac{1}{20} + \left(\frac{19}{20}\right)^4 \cdot \frac{1}{20} + \left(\frac{19}{20}\right)^5 \cdot \frac{1}{20} \\ &\approx 0.265 \end{aligned}$$

What is the expected number of rolls before I roll a 20? Intuition says that if I have a 1/20 chance, then I'll probably roll one every 20 rolls. Through a similar infinite series trick to the one above,

$$E[X] = \sum_{k=1}^{\infty} k \cdot P(X = k) = p + 2q \cdot p + 3q^2 \cdot p + 4q^3 \cdot + \ldots = p(1 + 2q + 3q^2 + 4q^3 \cdot + \ldots) = p\left(\frac{1}{(1-q)^2}\right) = \frac{1}{p}$$

so in this case the math confirms our intuition.