# Lecture 6 - Random Variables and Parameterized Sample Spaces

6.042 - February 25, 2003

We've used probablity to model a variety of experiments, games, and tests. Throughout, we have tried to compute probabilities of *events*. We asked, for example, what is the probability of the event that you win the Monty Hall game? What is the probability of the event that it rains, given that the weatherman carried his umbrella today? What is the probability of the event that you have a rare disease, given that you tested positive?

But one can ask more general questions about an experiment. *How hard* will it rain? *How long* will this illness last? *How much* will I lose playing 6.042 games all day? These questions are fundamentally different and not easily phrased in terms of events. The problem is that an event either does or does not happen: you win or lose, it rains or doesn't, you're sick or not. But these questions are about matters of degree: how much, how hard, how long? To approach these questions, we need a new tool: *random variables*.[1]

## 1 Random Variables

Let's begin with an example. Consider the experiment of tossing three independent, unbiased coins. Let $C$ be the number of heads that appear. Let $M = 1$ if the three coins come up all heads or all tails, and let $M = 0$ otherwise. Now every outcome of the three coin flips uniquely determines the values of $C$ and $M$. For example, if we flip heads, tails, heads, then $C = 2$ and $M = 0$. If we flip tails, tails, tails, then $C = 0$ and $M = 1$. In effect, $C$ counts the number of heads, and $M$ indicates whether all the coins match.

Since each outcome uniquely determines $C$ and $M$, we can regard them as functions mapping outcomes to numbers. For this experiment, the sample space, the set of all possible outcomes, is:

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

The random variable $C$ is then a function that maps each outcome in the sample space to a number as follows:

$$\begin{array}{rcl@{\qquad}rcl}
C(HHH) & = & 3 & C(THH) & = & 2 \\
C(HHT) & = & 2 & C(THT) & = & 1 \\
C(HTH) & = & 2 & C(TTH) & = & 1 \\
C(HTT) & = & 1 & C(TTT) & = & 0
\end{array}$$



[1]Yeah, yeah. Get your mind out of the gutter.

Similarly, $M$ is a function mapping each outcome another way:

$$\begin{array}{rclcrcl} M(HHH) & = & 1 & \qquad & M(THH) & = & 0 \\ M(HHT) & = & 0 & & M(THT) & = & 0 \\ M(HTH) & = & 0 & & M(TTH) & = & 0 \\ M(HTT) & = & 0 & & M(TTT) & = & 1 \end{array}$$

Sure enough, $C$ and $M$ are examples of random variables. The general definition is as follows:

**Definition 1** *A* random variable *is a function whose domain is the sample space.*

Usually, we'll consider random variables whose range is some subset of the real numbers, as is the case with $C$ and $M$. Notice that "random variable" is a misnomer; random variables are actually functions!

## 1.1 Indicator Random Variables

One type of random variable is particularly common and simple:

**Definition 2** *An* indicator random variable *(or simply an* indicator*) is a random variable that maps every outcome to either 0 or 1.*

The random variable $M$ is an example. If all three coins match, then $M = 1$; otherwise, $M = 0$.

Indicator random variables are closely related to events. In particular, an indicator partitions the sample space into those outcomes mapped to 1 and those outcomes mapped to 0. For example, the indicator $M$ partitions the sample space into two blocks as follows:

$$\underbrace{HHH \quad TTT}_{M=1} \quad \underbrace{HHT \quad HTH \quad HTT \quad THH \quad THT \quad TTH}_{M=0}$$

In the same way, an event partitions the sample space into those outcomes in the event and those outcomes not in the event. Therefore, each event is naturally associated with a certain indicator random variable and vice versa:

**Definition 3** *An* indicator for an event $E$ *is an indicator random variable that is 1 for all outcomes in* $E$ *and* 0 *for all outcomes not in* $E$*.*

Thus, $M$ is an indicator random variable for the event that all three coins match.

## 1.2 Random Variables and Events

There is a strong relationship between events and more general random variables as well. A random variable that takes on several values partitions the sample space into several blocks. For example, $C$ partitions the sample space as follows:

$$\underbrace{TTT}_{C=0} \quad \underbrace{TTH \quad THT \quad HTT}_{C=1} \quad \underbrace{THH \quad HTH \quad HHT}_{C=2} \quad \underbrace{HHH}_{C=3}$$

Each block is a subset of the sample space and is therefore an event. Thus, we can regard an equation or inequality involving a random variable as an event. For example, the event that $C = 2$ consists of the outcomes $THH$, $HTH$, and $HHT$. The event $C \leq 1$ consists of the outcomes $TTT$, $TTH$, $THT$, and $HTT$.

Naturally enough, we can talk about the probability of events defined by equations involving random variables. For example:

$$\begin{aligned}
\Pr\{C = 2\} &= \Pr\{THH\} + \Pr\{HTH\} + \Pr\{HHT\} \\
&= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\
&= \frac{3}{8}
\end{aligned}$$

As another example:

$$\begin{aligned}
\Pr\{M = 1\} &= \Pr\{TTT\} + \Pr\{HHH\} \\
&= \frac{1}{8} + \frac{1}{8} \\
&= \frac{1}{4}
\end{aligned}$$

## 1.3 Conditional Probability

Mixing conditional probabilities and events involving random variables creates no new difficulties. For example, $\Pr\{C \geq 2 \mid M = 0\}$ is the probability that at least two coins are heads ($C \geq 2$), given that not all three coins are the same ($M = 0$). We can compute this probability using the only definition of conditional probability:

$$
\begin{aligned}
\Pr\{C \geq 2 \mid M = 0\} &= \frac{\Pr\{C \geq 2 \cap M = 0\}}{\Pr\{M = 0\}} \\
&= \frac{\Pr\{\{THH, HTH, HHT\}\}}{\Pr\{\{THH, HTH, HHT, HTT, THT, TTH\}\}} \\
&= \frac{3/8}{6/8} \\
&= \frac{1}{2}
\end{aligned}
$$

The expression $C \geq 2 \cap M = 0$ on the first like may look odd; what is the set operation $\cap$ doing between an inequality and an equality? But recall that, in this context, $C \geq 2$ and $M = 0$ are events, *sets* of outcomes!

## 1.4 Independence

The notion of independence carries over from events to random variables as well.

**Definition 4** *Random variables $R_1$ and $R_2$ are independent if for all $x_1, x_2 \in \mathbb{R}$, we have:*

$$
\Pr\{R_1 = x_1 \cap R_2 = x_2\} = \Pr\{R_1 = x_1\} \cdot \Pr\{R_2 = x_2\}
$$

As for events, we can formulate independence for random variables in an equivalent and perhaps more intuitive way:

**Definition 5** *Random variables $R_1$ and $R_2$ are independent if for all $x_1, x_2 \in \mathbb{R}$ such that $\Pr\{R_2 = x_2\} > 0$, we have:*

$$
\Pr\{R_1 = x_1 \mid R_2 = x_2\} = \Pr\{R_1 = x_1\}
$$

In words, the probability that $R_1$ takes on a particular value is unaffected by the value of $R_2$.

As an example, are $C$ and $M$ independent? Intuitively, the answer should be “no”. The number of heads, $C$, completely determines whether all three coins match; that is, whether $M = 1$. But, to verify this intuition, we must find some $x_1, x_2 \in \mathbb{R}$ such that:

$$
\Pr\{C = x_1 \cap M = x_2\} \neq \Pr\{C = x_1\} \cdot \Pr\{M = x_2\}
$$

One appropriate choice of values is $x_1 = 2$ and $x_2 = 1$. In this case, we have:

$$\Pr\{C = 2 \cap M = 1\} = 0 \quad \text{but} \quad \Pr\{C = 2\} \cdot \Pr\{M = 1\} = \tfrac{3}{8} \cdot \tfrac{1}{4} \neq 0$$

The first probability is zero because we never have exactly two heads ($C = 2$) when all three coins match ($M = 1$). The other two probabilities were computed earlier.

The notion of independence generalizes to a set of random variables as follows:

**Definition 6** *Random variables $R_1, R_2, \ldots, R_n$ are mutually independent if for all $x_1, x_2, \ldots, x_n \in \mathbb{R}$ we have:*

$$\begin{aligned}\Pr\{R_1 = x_1 \cap R_2 = x_2 \cap \ldots \cap R_n = x_n\} \\ = \Pr\{R_1 = x_1\} \cdot \Pr\{R_2 = x_2\} \cdots \Pr\{R_n = x_n\}\end{aligned}$$

A consequence of this definition of mutual independence is that the probability of an assignment to a *subset* of the variables is equal to the product of the probabilites of the individual assignments. Thus, for example, if $R_1, R_2, \ldots, R_{100}$ are mutually independent random variables and $x_1, x_2, \ldots, x_{100}$ are arbitrary real numbers, then it follows that:

$$\Pr\{R_1 = x_1 \cap R_7 = x_7 \cap R_{23} = x_{23}\} = \Pr\{R_1 = x_1\} \cdot \Pr\{R_7 = x_7\} \cdot \Pr\{R_{23} = x_{23}\}$$

# 2 A Choice of Two Dice Games

You can win a prize by successfully playing one of two possible games:

- Pick one number between 1 and 6. Roll a fair die four times. If your number ever comes up, you win.
- Pick two different numbers between 1 and 6. Roll a fair die two times. If either of your numbers ever comes up, you win.

Assume that the outcomes of the die rolls are mutually independent. Which game should you play to have the best chance of winning the prize?

We can resolve this question by analyzing a more general game that subsumes these two. In this more general game, a player chooses a set of numbers $X \subseteq \{1, 2, \ldots, 6\}$. She then rolls the die $n$ times. If a number in the set $X$ comes up, she wins the prize. With what probability does this happen?

Let the random variables $R_1, R_2, \ldots, R_n$ be the numbers that come up on the die. In order to win the game, the player must do one of the following:

- Win on the first roll.
- Lose on the first roll, but win on the second.
- Lose on the first two rolls, but win on the third.
- ...
- Lose on the first $n-1$ rolls, but win on the $n$-th.

Note that these are disjoint events. Therefore, according to the sum rule, we have:

$$\begin{aligned}
\Pr\{\text{win}\} &= \Pr\{R_1 \in X\} \\
&\quad + \Pr\{(R_1 \notin X) \cap (R_2 \in X)\} \\
&\quad + \Pr\{(R_1 \notin X) \cap (R_2 \notin X) \cap (R_3 \in X)\} \\
&\quad + \ldots \\
&\quad + \Pr\{(R_1 \notin X) \cap (R_2 \notin X) \cap \ldots \cap (R_{n-1} \notin X) \cap (R_n \in X)\} \\
&= \Pr\{R_1 \in X\} \\
&\quad + \Pr\{R_1 \notin X\} \cdot \Pr\{R_2 \in X\} \\
&\quad + \Pr\{R_1 \notin X\} \cdot \Pr\{R_2 \notin X\} \cdot \Pr\{R_3 \in X\} \\
&\quad + \ldots \\
&\quad + \Pr\{R_1 \notin X\} \cdot \Pr\{R_2 \notin X\} \cdots \Pr\{R_{n-1} \notin X\} \cdot \Pr\{R_n \in X\}
\end{aligned}$$

In the second step, we use the fact that the die rolls are mutually independent.

Let $p$ be the probability that a particular die roll is a winner; that is, $p = \Pr\{R_i \in X\} = |X|/6$. Then we can rewrite the probability that the player wins more succinctly:

$$\begin{aligned}
\Pr\{\text{win}\} &= p + (1-p)p + (1-p)^2 p + \ldots + (1-p)^{n-1} p \\
&= p\left[1 + (1-p) + (1-p)^2 + \ldots + (1-p)^{n-1}\right]
\end{aligned}$$

The sum in square-brackets is called a *geometric sum*. In the next section, we'll see how to replace this sum with a compact expression, and then we'll return to analyzing the game.

## 2.1 Geometric Sums

A geometric sum has the form:

$$1 + x + x^2 + \ldots + x^{n-1}$$

Later in the course, we'll devote considerable effort to finding closed-form expressions for summations like this one. (*Closed-form expression* is an informal term referring to an expression involving no summation notation, product notation, triple-dots, or other blechy stuff.) For geometric sums, we can find a closed form via a neat trick. To begin, set $y$ equal to the whole sum:

$$y = 1 + x + x^2 + \ldots + x^{n-1}$$

Then we have:

$$xy = x + x^2 + x^3 + \ldots + x^n$$

Subtracting the second equation from the first, we find:

$$\begin{aligned} y - xy &= 1 - x^n \\ \Rightarrow \quad y(1-x) &= 1 - x^n \\ \Rightarrow \quad y &= \frac{1-x^n}{1-x} \quad (\text{for } x \neq 1) \end{aligned}$$

Putting this all together, we have a simple formula for the sum of a geometric series:

$$1 + x + x^2 + \ldots + x^{n-1} = \frac{1-x^n}{1-x} \quad (\text{for } x \neq 1)$$

## 2.2 Which Game is Better?

Now we can find the player's odds of winning the dice game by plugging in the formula for a geometric sum:

$$\begin{aligned} \Pr\{\text{win}\} &= p\left[1 + (1-p) + (1-p)^2 + \ldots + (1-p)^{n-1}\right] \\ &= p \cdot \frac{1-(1-p)^n}{1-(1-p)} \\ &= 1 - (1-p)^n \end{aligned}$$

This is the probability that a player rolls a winning number within $n$ tries, if she wins on each try with probability $p$.

From another perspective, this formula is obvious! We could have reasoned as follows:

$$\begin{aligned}
\Pr\{\text{win on some roll}\} &= 1 - \Pr\{\text{lose on every roll}\} \\
&= 1 - (\Pr\{\text{lose on one roll}\})^n \\
&= 1 - (1 - \Pr\{\text{win on one roll}\})^n \\
&= 1 - (1-p)^n
\end{aligned}$$

This is a common situation; there are two different ways to compute the same probability. Doing both is wise, if you want to be confident in your answer.

Let's see what this formula says about the two games:

- Pick one number and roll four times. Here $p = \frac{1}{6}$ and $n = 4$, so the probability of winning is $1 - (1 - 1/6)^4 = 0.517\ldots$.
- Pick two different numbers and roll two times. Now $p = \frac{1}{3}$ and $n = 2$, so the probability is $1 - (1 - 1/3)^2 = 0.555\ldots$.

The second game is just a little better!

# 3 Streaks

Was the table of $H$'s and $T$'s below generated by flipping a fair coin 100 times, or by someone tapping the H and T keys in a what felt like a random way?

$$\begin{array}{c}
HTTTHTHTTHTTHTHTHTHT \\
TTTHHTHHTHTHTTHHHTHT \\
HHTHHTTTHHHTHTTHHHHT \\
THTTHHTHTHTHTHTTHTHH \\
HTTHHHTHTHHHTHTHHHTH
\end{array}$$

There is no way to be sure. However, this sequence has a distinctive feature that is common in "random" human-generated sequences and unusual in truly random sequences: namely, there is no long streak of H's or T's. In fact, no symbol appears above more than four times in a row. How likely is that? If we flip a fair coin 100 times, what is the probability that we never get five heads in a row?

## 3.1 From a Probability Problem to a Counting Problem

The sample space for this experiment is $\{H,T\}^{100}$; that is, the set of all length-100 sequences of H's and T's. If the coin tosses are fair and independent, then all $2^{100}$ such sequences are equally likely. Therefore, we need only count the number of sequences with no streak of five heads; given that, the probability that a random length-100 sequence contains no such streak is:

$$\Pr\{\text{sequence has no } HHHHH\} = \frac{\#\text{ sequences with no } HHHHH}{2^{100}}$$

This is a common situation. We have reduced a probability problem to a counting problem. Unfortunatley, we have no hope of solving the counting problem by direct computation. No computer can consider all $2^{100}$ sequences of $H$'s and $T$'s, keeping track of how many lack a streak of five heads. But, on the bright side, there is a big bag of mathematical tricks for solving counting problems. In this case, we'll use a *recurrence equation.* The recurrence equation approach involves two steps:

1. Solve some small problems.
2. Solve the $n$-th problem using preceding solutions.

Let's see how this approach plays out in the analysis of streaks.

## 3.2 Step 1: Solve Small Instances

Let $S_n$ be the set of length-$n$ sequences of $H$'s and $T$'s that do not contain a streak of five heads. Our eventual goal is to compute $|S_{100}|$. But for now, let's just compute $|S_n|$ for some very small values of $n$:

$$\begin{aligned}
|S_1| &= 2 \quad (H \text{ and } T) \\
|S_2| &= 4 \quad (HH,\ HT,\ TH, \text{ and } TT) \\
|S_3| &= 8 \\
|S_4| &= 16 \\
|S_5| &= 31 \quad (HHHHH \text{ is excluded!})
\end{aligned}$$

These are called *base cases.*

## 3.3 Step 2: Solve the $n$-th Problem Using Preceding Solutions

We can classify the sequences in $S_n$ into five groups:

1. Sequences that end with a $T$.
2. Sequences that end with $TH$.
3. Sequences that end with $THH$.
4. Sequences that end with $THHH$.
5. Sequences that end with $THHHH$.

Every sequence in $S_n$ falls into exactly one of these groups. Thus, the size of $S_n$ is the sum of the sizes of these five groups.

How many sequences are there in the first group? That is, how many sequences in $S_n$ end with $T$? The preceding $n-1$ symbols in such a sequence can not contain a streak of five heads. Therefore, those $n-1$ symbols form a sequence in $S_{n-1}$. On the other hand, putting a $T$ at then end of any sequence in $S_{n-1}$ gives a sequence in $S_n$. Therefore, the number of sequences in the first group is exactly equal to $|S_{n-1}|$.

How many sequences are there in the second group? Arguing as before, the preceding $n-2$ symbols in each such sequence must form a sequence in $S_{n-2}$. On the other hand, appending $TH$ to any sequence in $S_{n-2}$ gives a sequence in the second group. Therefore, there are exactly $|S_{n-2}|$ sequences in the second group.

By similar reasoning, the number of sequences in the third group is $|S_{n-3}|$, the number in the fourth is $|S_{n-4}|$, and the number in the fifth is $|S_{n-5}|$. Therefore, we have:

$$|S_n| \quad = \quad |S_{n-1}| + |S_{n-2}| + |S_{n-3}| + |S_{n-4}| + |S_{n-5}| \quad (\text{for } n > 5)$$

This *recurrence equation* expresses the solution to a large problem ($|S_n|$) in terms of the solutions to smaller problems ($S_{n-1}$, $S_{n-2}$, $\ldots$).

By combining the base cases and the recurrence equation, we can compute $|S_6|$, $|S_7|$, $|S_8|$, and so forth until we reach $|S_{100}|$.

$$\begin{aligned}
|S_6| &= |S_5| + |S_4| + |S_3| + |S_2| + |S_1| \\
&= 31 + 16 + 8 + 4 + 2 + 1 \\
&= 63 \\
|S_7| &= |S_6| + |S_5| + |S_4| + |S_3| + |S_2| \\
&= 63 + 31 + 16 + 8 + 4 + 2 \\
&= 124 \\
|S_8| &= \ldots
\end{aligned}$$

Computing $|S_{100}|$ still requires about 500 additions, so a computer helps. Plugging the value of $|S_{100}|$ into our earlier probability formula, we find:

$$\Pr\{\text{sequence has no } HHHHH\} \quad = \quad 0.193\ldots$$

Thus, four out of five sequences of 100 coin tosses contain a streak of five heads. By symmetry, we also know that four out of five sequences contain a streak of five tails. If we suppose that these two events are nearly independent, then only about one random sequence in twenty-five contains no streak of five heads or five tails. This is the situation for the sequence given at the start of this section. Sure enough, I made up that sequence up by "randomly" tapping keys!