

## Random Variables

We've used probability 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 new questions are about matters of degree: how much, how hard, how long? To approach these questions, we need a new mathematical tool.

### 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 is:

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

Now  $C$  is a function that maps each outcome in the sample space to a number as follows:

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

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

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

The functions  $C$  and  $M$  are examples of **random variables**. In general, a random variable is a function whose domain is the sample space. (The codomain can be anything, but we'll usually use a subset of the real numbers.) Notice that the name "random variable" is a misnomer; random variables are actually functions!

## 1.1 Indicator Random Variables

An **indicator random variable** (or simply an **indicator** or a **Bernoulli random variable**) 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: 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 and inequalities involving random variables. For example:

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

As another example:

$$\begin{aligned}\Pr(C \geq 2) &= \Pr(THH) + \Pr(HTH) + \Pr(HHT) + \Pr(HHH) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{2}\end{aligned}$$

This is pretty wild; one normally thinks of equations and inequalities as either true or false. But when variables are replaced by random variables, there is a *probability* that the relationship holds!

### 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 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 line 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, which *sets* of outcomes. So taking their intersection is perfectly valid!

### 1.4 Independence

The notion of independence carries over from events to random variables as well. Random variables  $R_1$  and  $R_2$  are *independent* if

$$\Pr(R_1 = x_1 \cap R_2 = x_2) = \Pr(R_1 = x_1) \cdot \Pr(R_2 = x_2)$$

for all  $x_1$  in the codomain of  $R_1$  and  $x_2$  in the codomain of  $R_2$ .

As with events, we can formulate independence for random variables in an equivalent and perhaps more intuitive way: random variables  $R_1$  and  $R_2$  are independent if and only if

$$\Pr(R_1 = x_1 \mid R_2 = x_2) = \Pr(R_1 = x_1) \text{ or } \Pr(R_2 = x_2) = 0$$

for all  $x_1$  in the codomain of  $R_1$  and  $x_2$  in the codomain of  $R_2$ . 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 that case, we have:

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

The notion of independence generalizes to a set of random variables as follows. Random variables  $R_1, R_2, \dots, R_n$  are **mutually independent** if

$$\begin{aligned} \Pr(R_1 = x_1 \cap R_2 = x_2 \cap \dots \cap R_n = x_n) \\ = \Pr(R_1 = x_1) \cdot \Pr(R_2 = x_2) \cdot \dots \cdot \Pr(R_n = x_n) \end{aligned}$$

for all  $x_1, \dots, x_n$  in the codomains of  $R_1, \dots, R_n$ .

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 probabilities of the individual assignments. Thus, for example, if  $R_1, R_2, \dots, R_{100}$  are mutually independent random variables with codomain  $\mathbb{N}$ , then it follows that:

$$\Pr(R_1 = 9 \cap R_7 = 84 \cap R_{23} = 13) = \Pr(R_1 = 9) \cdot \Pr(R_7 = 84) \cdot \Pr(R_{23} = 13)$$

(This follows by summing over all possible values of the other random variables; we omit the details.)

## 1.5 An Example with Dice

Suppose that we roll two fair, independent dice. The sample space for this experiment consists of all pairs  $(r_1, r_2)$  where  $r_1, r_2 \in \{1, 2, 3, 4, 5, 6\}$ . Thus, for example, the outcome  $(3, 5)$  corresponds to rolling a 3 on the first die and a 5 on the second. The probability of each outcome in the sample space is  $1/6 \cdot 1/6 = 1/36$  since the dice are fair and independent.

We can regard the numbers that come up on the individual dice as random variables  $D_1$  and  $D_2$ . So  $D_1(3, 5) = 3$  and  $D_2(3, 5) = 5$ . Then the expression  $D_1 + D_2$  is another random variable; let's call it  $T$  for “total”. More precisely, we've defined:

$$T(w) = D_1(w) + D_2(w) \quad \text{for every outcome } w$$

Thus,  $T(3, 5) = D_1(3, 5) + D_2(3, 5) = 3 + 5 = 8$ . In general, any function of random variables is itself a random variable. For example,  $\sqrt{D_1} + \cos(D_2)$  is a strange, but well-defined random variable.

Let's also define an indicator random variable  $S$  for the event that the total of the two dice is seven:

$$S(w) = \begin{cases} 1 & \text{if } T(w) = 7 \\ 0 & \text{if } T(w) \neq 7 \end{cases}$$

So  $S$  is equal to 1 when the sum is seven and is equal to 0 otherwise. For example,  $S(4, 3) = 1$ , but  $S(5, 3) = 0$ .

Now let's consider a couple questions about independence. First, are  $D_1$  and  $T$  independent? Intuitively, the answer would seem to be "no" since the number that comes up on the first die strongly affects the total of the two dice. But to prove this, we must find integers  $x_1$  and  $x_2$  such that:

$$\Pr(D_1 = x_1 \cap T = x_2) \neq \Pr(D_1 = x_1) \cdot \Pr(T = x_2)$$

For example, we might choose  $x_1 = 2$  and  $x_2 = 3$ . In this case, we have

$$\Pr(T = 2 \mid D_1 = 3) = 0$$

since the total can not be only 2 when one die alone is 3. On the other hand, we have:

$$\begin{aligned} \Pr(T = 2) \cdot \Pr(D_1 = 3) &= \Pr(\{1, 1\}) \cdot \Pr(\{(3, 1), (3, 2), \dots, (3, 6)\}) \\ &= \frac{1}{36} \cdot \frac{6}{36} \neq 0 \end{aligned}$$

So, as we suspected, these random variables are not independent.

Are  $S$  and  $D_1$  independent? Once again, intuition suggests that the answer is "no". The number on the first die ought to affect whether or not the sum is equal to seven. But this time intuition turns out to be wrong! These two random variables actually are independent.

Proving that two random variables are independent takes some work. (Fortunately, this is an uncommon task; usually independence is a modeling assumption. Only rarely do random variables unexpectedly turn out to be independent.) In this case, we must show that

$$\Pr(S = x_1 \cap D_1 = x_2) = \Pr(S = x_1) \cdot \Pr(D_1 = x_2) \tag{1}$$

for all  $x_1 \in \{0, 1\}$  and all  $x_2 \in \{1, 2, 3, 4, 5, 6\}$ . We can work through all these possibilities in two batches:

- Suppose that  $x_1 = 1$ . Then for every value of  $x_2$  we have:

$$\begin{aligned} \Pr(S = 1) &= \Pr((1, 6), (2, 5), \dots, (6, 1)) = \frac{1}{6} \\ \Pr(D_1 = x_2) &= \Pr((x_2, 1), (x_2, 2), \dots, (x_2, 6)) = \frac{1}{6} \\ \Pr(S = 1 \cap D_1 = x_2) &= \Pr((x_2, 7 - x_2)) = \frac{1}{36} \end{aligned}$$

Since  $1/6 \cdot 1/6 = 1/36$ , the independence condition is satisfied.

- Otherwise, suppose that  $x_1 = 0$ . Then we have  $\Pr(S = 0) = 1 - \Pr(S = 1) = 5/6$  and  $\Pr(D_1 = x_2) = 1/6$  as before. Now the event

$$S = 0 \cap D_1 = x_2$$

consists of 5 outcomes: all of  $(x_2, 1), (x_2, 2), \dots, (x_2, 6)$  except for  $(x_2, 7 - x_2)$ . Therefore, the probability of this event is  $5/36$ . Since  $5/6 \cdot 1/6 = 5/36$ , the independence condition is again satisfied.

Thus, the outcome of the first die roll is independent of the fact that the sum is 7. This is a strange, isolated result; for example, the first roll is *not* independent of the fact that the sum is 6 or 8 or any number other than 7. But this example shows that the mathematical notion of independent random variables— while closely related to the intuitive notion of “unrelated quantities”— is not exactly the same thing.

## 2 Probability Distributions

A random variable is defined to be a function whose domain is the sample space of an experiment. Often, however, random variables with essentially the same properties show up in completely different experiments. For example, some random variable that come up in polling, in primality testing, and in coin flipping all share some common properties. If we could study such random variables in the abstract, divorced from the details any particular experiment, then our conclusions would apply to *all* the experiments where that sort of random variable turned up. Such general conclusions could be very useful. There are a couple tools that capture the essential properties of a random variable, but leave other details of the associated experiment behind.

The **probability density function (pdf)** for a random variable  $R$  with codomain  $V$  is a function  $\text{PDF}_R : V \rightarrow [0, 1]$  defined by:

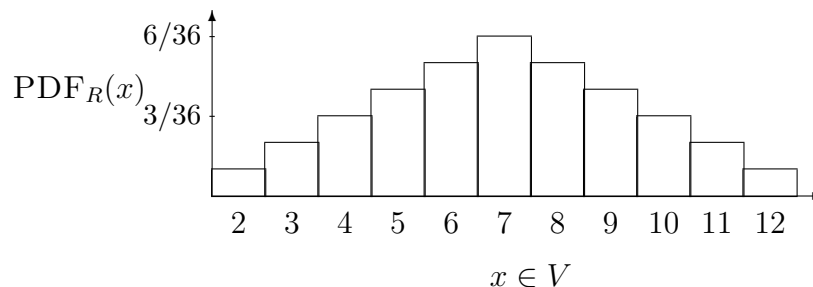
$$\text{PDF}_R(x) = \Pr(R = x)$$

A consequence of this definition is that

$$\sum_{x \in V} \text{PDF}_R(x) = 1$$

since the random variable always takes on exactly one value in the set  $V$ .

As an example, let’s return to the experiment of rolling two fair, independent dice. As before, let  $T$  be the total of the two rolls. This random variable takes on values in the set  $V = \{2, 3, \dots, 12\}$ . A plot of the probability density function is shown below:

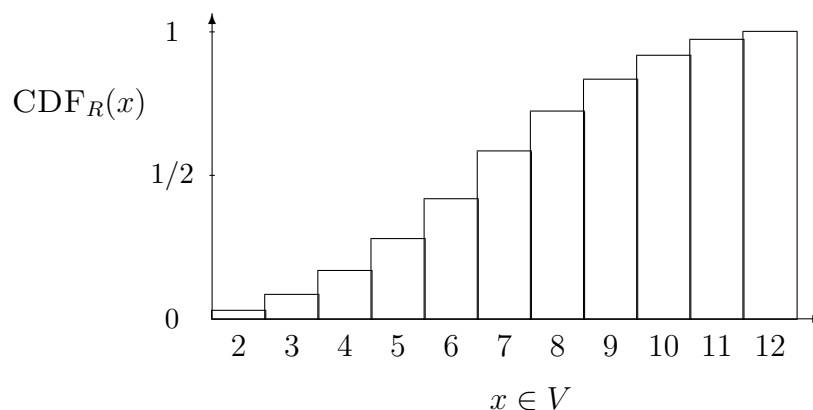


The lump in the middle indicates that sums close to 7 are the most likely. The total area of all the rectangles is 1 since the dice must take on exactly one of the sums in  $V = \{2, 3, \dots, 12\}$ .

A closely-related idea is the **cumulative distribution function (cdf)** for a random variable  $R$ . This is a function  $\text{CDF}_R : V \rightarrow [0, 1]$  defined by:

$$\text{CDF}_R(x) = \Pr(R \leq x)$$

As an example, the cumulative distribution function for the random variable  $T$  is shown below:



The height of the  $i$ -th bar in the cumulative distribution function is equal to the *sum* of the heights of the leftmost  $i$  bars in the probability density function. This follows from the definitions of pdf and cdf:

$$\begin{aligned} \text{CDF}_R(x) &= \Pr(R \leq x) \\ &= \sum_{y \leq x} \Pr(R = y) \\ &= \sum_{y \leq x} \text{PDF}_R(y) \end{aligned}$$

In summary,  $\text{PDF}_R(x)$  measures the probability that  $R = x$  and  $\text{CDF}_R(x)$  measures the probability that  $R \leq x$ . Both the  $\text{PDF}_R$  and  $\text{CDF}_R$  capture the same information about the random variable  $R$ — you can derive one from the other— but sometimes one is

more convenient. The key point here is that neither the probability density function nor the cumulative distribution function involves the sample space of an experiment. Thus, through these functions, we can study random variables without reference to a particular experiment.

For the remainder of today, we'll look at three important distributions and some applications.

## 2.1 Bernoulli Distribution

Indicator random variables are perhaps the most common type because of their close association with events. The probability density function of an indicator random variable  $B$ , that is the pdf for the number of heads (0 or 1) when you flip a (possibly biased) coin once, is always

$$\begin{aligned}\text{PDF}_B(0) &= p \\ \text{PDF}_B(1) &= 1 - p\end{aligned}$$

where  $0 \leq p \leq 1$ . The corresponding cumulative distribution function is:

$$\begin{aligned}\text{CDF}_B(0) &= p \\ \text{CDF}_B(1) &= 1\end{aligned}$$

This is called the *Bernoulli distribution*. The number of heads flipped on a (possibly biased) coin has a Bernoulli distribution.

## 2.2 Uniform Distribution

A random variable that takes on each possible values with the same probability is called *uniform*. For example, the probability density function of a random variable  $U$  that is uniform on the set  $\{1, 2, \dots, N\}$  is:

$$\text{PDF}_U(k) = \frac{1}{N}$$

And the cumulative distribution function is:

$$\text{CDF}_U(k) = \frac{k}{N}$$

Uniform distributions come up all the time. For example, the number rolled on a fair die is uniform on the set  $\{1, 2, \dots, 6\}$ .

## 2.3 The Numbers Game

Let's play a game! I have two envelopes. Each contains an integer in the range  $0, 1, \dots, 100$ , and the numbers are distinct. To win the game, you must determine which envelope contains the larger number. To give you a fighting chance, I'll let you peek at the number in one envelope selected at random. Can you devise a strategy that gives you a better than 50% chance of winning?

For example, you could just pick an envelope at random and guess that it contains the larger number. But this strategy wins only 50% of the time. Your challenge is to do better.

So you might try to be more clever. Suppose you peek in the left envelope and see the number 12. Since 12 is a small number, you might guess that that other number is larger. But perhaps I'm sort of tricky and put small numbers in *both* envelopes. Then your guess might not be so good!

An important point here is that the numbers in the envelopes may *not* be random. I'm picking the numbers and I'm choosing them in a way that I think will defeat your guessing strategy. I'll only use randomization to choose the numbers if that serves *my* end: making you lose!

### 2.3.1 Intuition Behind the Winning Strategy

Amazingly, there is a strategy that wins more than 50% of the time, regardless of what numbers I put in the envelopes!

Suppose that you somehow knew a number  $x$  *between* my lower number and higher numbers. Now you peek in an envelope and see one or the other. If it is bigger than  $x$ , then you know you're peeking at the higher number. If it is smaller than  $x$ , then you're peeking at the lower number. In other words, if you know an number  $x$  between my lower and higher numbers, then you are certain to win the game.

The only flaw with this brilliant strategy is that you do *not* know  $x$ . Oh well.

But what if you try to *guess*  $x$ ? There is some probability that you guess correctly. In this case, you win 100% of the time. On the other hand, if you guess incorrectly, then you're no worse off than before; your chance of winning is still 50%. Combining these two cases, your overall chance of winning is better than 50%!

Informal arguments about probability, like this one, often sound plausible, but do not hold up under close scrutiny. In contrast, this argument sounds completely implausible—but is actually correct!

### 2.3.2 Analysis of the Winning Strategy

For generality, suppose that I can choose numbers from the set  $\{0, 1, \dots, n\}$ . Call the lower number  $L$  and the higher number  $H$ .

Your goal is to guess a number  $x$  between  $L$  and  $H$ . To avoid confusing equality cases, you select  $x$  at random from among the half-integers:

$$\left\{ \frac{1}{2}, 1\frac{1}{2}, 2\frac{1}{2}, \dots, n - \frac{1}{2} \right\}$$

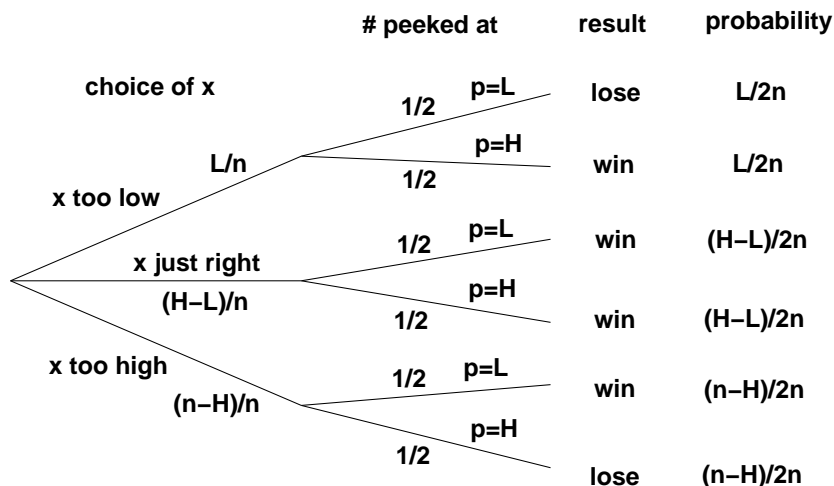
But what probability distribution should you use?

The uniform distribution turns out to be your best bet. An informal justification is that if I figured out that you were unlikely to pick some number—say  $50\frac{1}{2}$ —then I'd always put 50 and 51 in the envelopes. Then you'd be unlikely to pick an  $x$  between  $L$  and  $H$  and would have less chance of winning.

After you've selected the number  $x$ , you peek into an envelope and see some number  $p$ . If  $p > x$ , then you guess that you're looking at the larger number. If  $p < x$ , then you guess that the other number is larger.

All that remains is to determine the probability that this strategy succeeds. We can do this with the usual four-step method and a tree diagram.

**Step 1: Find the sample space.** You either choose  $x$  too low ( $< L$ ), too high ( $> H$ ), or just right ( $L < x < H$ ). Then you either peek at the lower number ( $p = L$ ) or the higher number ( $p = H$ ). This gives a total of six possible outcomes.



**Step 2: Define events of interest.** The four outcomes in the event that you win are marked in the tree diagram.

**Step 3: Assign outcome probabilities.** First, we assign edge probabilities. Your guess  $x$  is too low with probability  $L/n$ , too high with probability  $(n - H)/n$ , and just right with probability  $(H - L)/n$ . Next, you peek at either the lower or higher number with equal probability. Multiplying along root-to-leaf paths gives the outcome probabilities.

**Step 4: Compute event probabilities.** The probability of the event that you win is

the sum of the probabilities of the four outcomes in that event:

$$\begin{aligned} \Pr(\text{win}) &= \frac{L}{2n} + \frac{H-L}{2n} + \frac{H-L}{2n} + \frac{n-H}{2n} \\ &= \frac{1}{2} + \frac{H-L}{2n} \\ &\geq \frac{1}{2} + \frac{1}{2n} \end{aligned}$$

The final inequality relies on the fact that the higher number  $H$  is at least 1 greater than the lower number  $L$  since they are required to be distinct.

Sure enough, you win with this strategy more than half the time, regardless of the numbers in the envelopes! For example, if I choose numbers in the range  $0, 1, \dots, 100$ , then you win with probability at least  $\frac{1}{2} + \frac{1}{200} = 50.5\%$ . Even better, if I'm allowed only numbers in the range  $0, \dots, 10$ , then your probability of winning rises to  $55\%$ ! By Las Vegas standards, those are great odds!

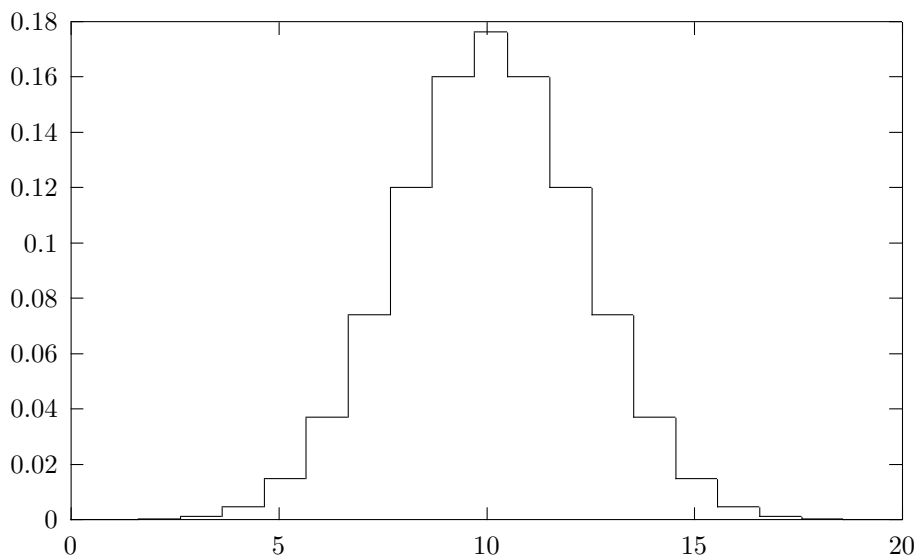
## 2.4 Binomial Distribution

Of the more complex distributions, the ***binomial distribution*** is surely the most important in computer science. The standard example of a random variable with a binomial distribution is the number of heads that come up in  $n$  independent flips of a coin; call this random variable  $H$ . If the coin is fair, then  $H$  has an *unbiased binomial density function*:

$$\text{PDF}_H(k) = \binom{n}{k} 2^{-n}$$

This follows because there are  $\binom{n}{k}$  sequences of  $n$  coin tosses with exactly  $k$  heads, and each such sequence has probability  $2^{-n}$ .

Here is a plot of the unbiased probability density function  $\text{PDF}_H(k)$  corresponding to  $n = 20$  coins flips. The most likely outcome is  $k = 10$  heads, and the probability falls off rapidly for larger and smaller values of  $k$ . These falloff regions to the left and right of the main hump are usually called the ***tails of the distribution***.



An enormous number of analyses in computer science come down to proving that the tails of the binomial and similar distributions are very small. In the context of a problem, this typically means that there is very small probability that something *bad* happens, which could be a server or communication link overloading or a randomized algorithm running for an exceptionally long time or producing the wrong result.

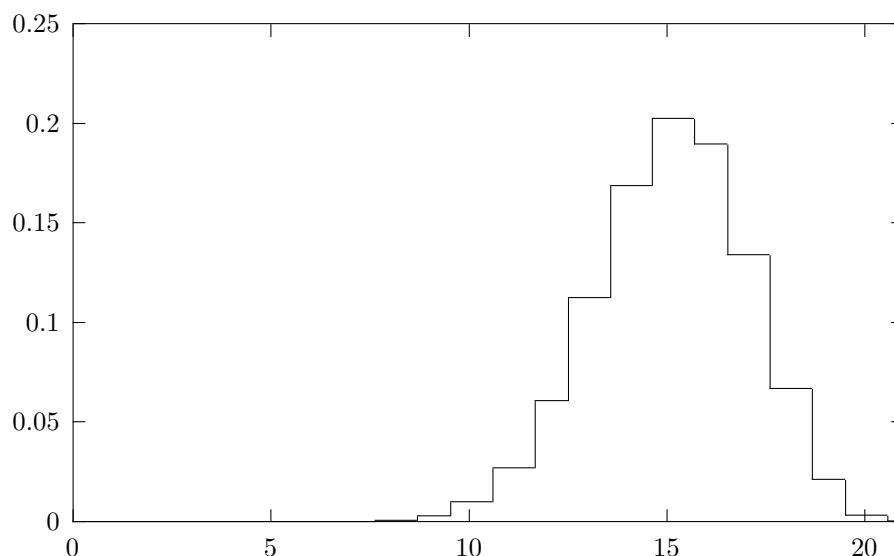
### 2.4.1 The General Binomial Distribution

Now let  $J$  be the number of heads that come up on  $n$  independent coins, each of which is heads with probability  $p$ . Then  $J$  has a *general binomial density function*:

$$\text{PDF}_J(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

As before, there are  $\binom{n}{k}$  sequences with  $k$  heads and  $n-k$  tails, but now the probability of each such sequence is  $p^k(1-p)^{n-k}$ .

As an example, the plot below shows the probability density function  $\text{PDF}_J(k)$  corresponding to flipping  $n = 20$  independent coins that are heads with probability  $p = 0.75$ . The graph shows that we are most likely to get around  $k = 15$  heads, as you might expect. Once again, the probability falls off quickly for larger and smaller values of  $k$ .



### 2.4.2 Approximating the Binomial Density Function

There is an approximate closed-form formula for the general binomial density function, though it is a bit unwieldy. First, we need an approximation for a key term in the exact formula,  $\binom{n}{k}$ . For convenience, let's replace  $k$  by  $\alpha n$  where  $\alpha$  is a number between 0 and 1. Then, from Stirling's formula, we find that:

$$\binom{n}{\alpha n} \leq \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}}$$

where  $H(\alpha)$  is the famous *entropy function*:

$$H(\alpha) = \alpha \log_2 \frac{1}{\alpha} + (1 - \alpha) \log_2 \frac{1}{1 - \alpha}$$

This upper bound on  $\binom{n}{\alpha n}$  is very tight and serves as an excellent approximation.

Now let's plug this formula into the general binomial density function. The probability of flipping  $\alpha n$  heads in  $n$  tosses of a coin that comes up heads with probability  $p$  is:

$$\text{PDF}_J(\alpha n) \leq \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot p^{\alpha n} (1-p)^{(1-\alpha)n} \quad (2)$$

This formula is ugly as a bowling shoe, but quite useful. For example, suppose we flip a fair coin  $n$  times. What is the probability of getting *exactly*  $\frac{1}{2}n$  heads? Plugging  $\alpha = 1/2$  and  $p = 1/2$  into this formula gives:

$$\begin{aligned} \text{PDF}_J(\alpha n) &\leq \frac{2^{nH(1/2)}}{\sqrt{2\pi(1/2)(1-(1/2))n}} \cdot 2^{-n} \\ &= \sqrt{\frac{2}{\pi n}} \end{aligned}$$

Thus, for example, if we flip a fair coin 100 times, the probability of getting exactly 50 heads is about  $1/\sqrt{50\pi} \approx 0.079$  or around 8%.

## 2.5 Approximating the Cumulative Binomial Distribution Function

Suppose a coin comes up heads with probability  $p$ . As before, let the random variable  $J$  be the number of heads that come up on  $n$  independent flips. Then the probability of getting *at most*  $k$  heads is given by the cumulative binomial distribution function:

$$\begin{aligned} \text{CDF}_J(k) &= \Pr(J \leq k) \\ &= \sum_{i=0}^k \text{PDF}_J(i) \\ &= \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

Evaluating this expression directly would be a lot of work for large  $k$  and  $n$ , so now an approximation would be really helpful. Once again, we can let  $k = \alpha n$ ; that is, instead of thinking of the absolute number of heads ( $k$ ), we consider the fraction of flips that are heads ( $\alpha$ ). The following approximation holds provided  $\alpha < p$ :

$$\begin{aligned} \text{CDF}_J(\alpha n) &\leq \frac{1 - \alpha}{1 - \alpha/p} \cdot \text{PDF}_J(\alpha n) \\ &\leq \frac{1 - \alpha}{1 - \alpha/p} \cdot \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot p^{\alpha n} (1-p)^{(1-\alpha)n} \end{aligned}$$

In the first step, we upper bound the summation with a geometric sum and apply the formula for the sum of a geometric series. (The details are dull and omitted.) Then we insert the approximate formula (2) for  $\text{PDF}_J(\alpha n)$  from the preceding section.

You have to press a lot of buttons on a calculator to evaluate this formula for a specific choice of  $\alpha$ ,  $p$ , and  $n$ . (Even computing  $H(\alpha)$  is a fair amount of work!) But for large  $n$ , evaluating the cumulative distribution function exactly requires vastly *more* work! So don't look gift blessings in the mouth before they hatch. Or something.

As an example, the probability of flipping at most 25 heads in 100 tosses of a fair coin is obtained by setting  $\alpha = 1/4$ ,  $p = 1/2$  and  $n = 100$ :

$$\text{CDF}_J(n/4) \leq \frac{1 - (1/4)}{1 - (1/4)/(1/2)} \cdot \text{PDF}_J(n/4) \leq \frac{3}{2} \cdot 1.913 \cdot 10^{-7}$$

This says that flipping 25 or fewer heads is extremely unlikely, which is consistent with our earlier claim that the tails of the binomial distribution are very small. In fact, notice that the

probability of flipping *25 or fewer* heads is only 50% more than the probability of flipping *exactly 25* heads. Thus, flipping exactly 25 heads is twice as likely as flipping any number between 0 and 24!

**Caveat:** The upper bound on  $\text{CDF}_J(\alpha n)$  holds only if  $\alpha < p$ . If this is not the case in your problem, then try thinking in complementary terms; that is, look at the number of tails flipped instead of the number of heads.