Random Variables, Distributions and Expectation

1 Random Variables

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 mathematical tool.

1.1 Definition

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:

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

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

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

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

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

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!

1.2 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.

1.3 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:

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

1.4 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*, namely, *sets* of outcomes.

1.5 Independence

The notion of independence carries over from events to random variables as well. Random variables R_1 and R_2 are ***independent*** if for all x_1 in the codomain of R_1, and x_2 in the codomain of R_2, we have:

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

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 for all x_1 and x_2 in the codomains of R_1 and R_2 respectively, 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:

$$\Pr\{[C = x_1] \cap [M = x_2]\} \quad \neq \quad \Pr\{C = x_1\} \cdot \Pr\{M = x_2\}.$$

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

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

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. Random variables $R_1, R_2, \ldots, R_n$ are ***mutually independent*** if for all $x_1, x_2, \ldots, x_n$, in the codomains of $R_1, R_2, \ldots, R_n$ respectively, we have:

$$\begin{aligned}\Pr\{[R_1 = x_1] \cap [R_2 = x_2] \cap \cdots \cap [R_n = x_n]\} \\ = \quad \Pr\{R_1 = x_1\} \cdot \Pr\{R_2 = x_2\} \cdots \Pr\{R_n = x_n\}.\end{aligned}$$

A consequence of this definition of mutual independence is that the probability that any *subset* of the variables takes a particular set of values is equal to the product of the probabilites that the individual variables take their values. Thus, for example, if $R_1, R_2, \ldots, R_{100}$ are mutually independent random variables, then it follows that:

$$\Pr\{[R_1 = 7] \cap [R_7 = 9.1] \cap [R_{23} = \pi]\} \quad = \quad \Pr\{R_1 = 7\} \cdot \Pr\{R_7 = 9.1\} \cdot \Pr\{R_{23} = \pi\}.$$

2 The Birthday Principle

There are 100 students in a lecture hall. What is the probability that some two people share a birthday? Maybe about $1/3$? Let's check! We'll use the following two variables throughout our analysis:

- Let n be the number of people in the group.
- Let d be the number of days in the year.

Furthermore, we'll make the assumption that birthdays are uniformly-distributed, independent random variables. This assumption is not really valid in the real world, since more babies are born at certain times of year and the birthdays of twins are clearly not independent. However, our analysis of this problem applies to many situations in computer science that are unaffected by twins, leap days, and romantic holidays anyway, so we won't dwell on those complications.

The sample space for this experiment consists of all ways of assigning birthdays to the people of the group. There are d^n such assignments, since the first person can have d different birthdays, the second person can have d different birthdays, and so forth. Furthermore, every such assignment is equally probable by our assumption that birthdays are uniformly-distributed and mutually independent, so the sample space is uniform.

Let D be the event that everyone has a distinct birthday. This is the complement of the event that we're interested in, but the probability of D is easier to evaluate. Later we can use the fact that $\Pr\left\{\overline{D}\right\} = 1 - \Pr\{D\}$ to compute the probability we really want. Anyway, event D consists of $d(d-1)(d-2)\cdots(d-n+1)$ outcomes, since we can select the birthday of the first person in d days, the birthday of the second person in $d-1$ ways, and so forth. Therefore, the probability that everyone has a different birthday is:

$$\Pr\{D\} = \frac{d(d-1)(d-2)\cdots(d-n+1)}{d^n}.$$

For $n = 100$, this probability is actually fantastically small —less than one in a million! If there are 100 people in a room, two are almost certain to share a birthday.

Let's use an approximation to rewrite the right side of the preceding equation in a more insightful form:

$$\begin{aligned}\Pr\{D\} &= \left(1 - \frac{0}{d}\right)\left(1 - \frac{1}{d}\right)\left(1 - \frac{2}{d}\right)\cdots\left(1 - \frac{n-1}{d}\right)\\ &\approx e^0 \cdot e^{-1/d} \cdot e^{-2/d} \cdots e^{-(n-1)/d}\\ &= e^{\frac{-n(n-1)}{2d}}.\end{aligned}$$

In the first step, we pair each term in the numerator with a d term in the denominator. Next, we use the approximation $e^{-x} \approx 1 - x$, which is pretty accurate when x is small.[1] In the last step, we combine exponents using the familiar formula $0 + 1 + 2 + \cdots + (n-1) = n(n-1)/2$.

The exponent in the final expression above is -1 when $n \approx \sqrt{2d}$. This is roughly the break-even point, where the probability that two people share a birthday is in the ballpark of $1/2$. This leads to a rule called the ***birthday principle***, which is useful in many contexts in computer science:

> If there are d days in a year and $\sqrt{2d}$ people in a room, then the probability that two share a birthday is about $1 - 1/e \approx 0.632$.

[1] This approximation is obtained by truncating the Taylor series $e^{-x} = 1 - x + x^2/2! - x^3/3! + \cdots$

For example, this principle says that if you have $\sqrt{2 \cdot 365} \approx 27$ people in a room, then the probability that two share a birthday is about 0.632. The actual probability is about 0.626, so the approximation is quite good.

The Birthday Principle is a great rule of thumb with surprisingly many applications. For example, cryptographic systems and digital signature schemes must be hardened against "birthday attacks". The principle also tells us how many items can be inserted into a hash table before one starts to experience collisions.

3 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 $\text{PDF}_R : V \to [0, 1]$ defined by:

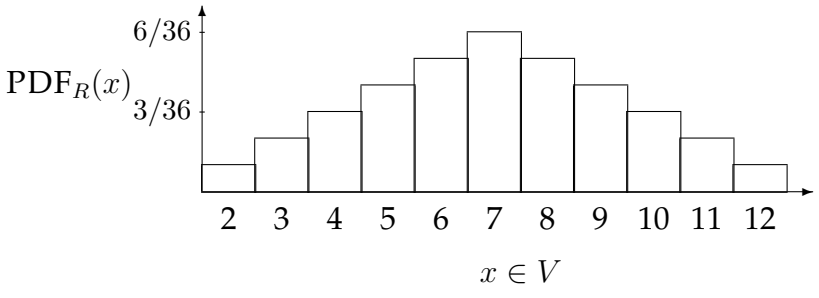
$$\text{PDF}_R(x) = \Pr\{R = x\}$$

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 $V = \{2, 3, \ldots, 12\}$. A plot of the probability density function is shown below:

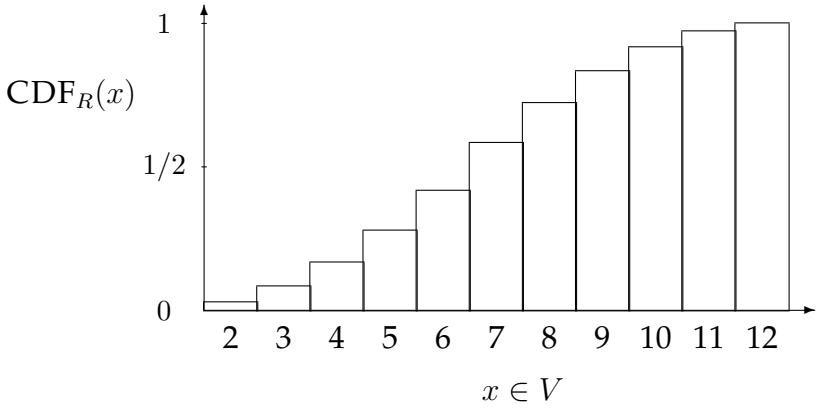


The lump in the middle indicates that sums close to 7 are the most likely. The total area of all the rectangles is 1 since the dice must take on exactly one of the sums in $V = \{2, 3, \ldots, 12\}$.

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

We'll now look at three important distributions and some applications.

3.1 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 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:

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

3.2 Uniform Distribution

A random variable that takes on each possible value with the same probability is called ***uniform***. For example, the probability density function of a random variable U that is uniform on the set $\{1, 2, \ldots, N\}$ 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 $\{1, 2, \ldots, 6\}$.

3.3 The Numbers Game

Let's play a game! I have two envelopes. Each contains an integer in the range $0, 1, \ldots, 100$, 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 evelope 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!

3.3.1 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!

3.3.2 Analysis of the Winning Strategy

For generality, suppose that I can choose numbers from the set $\{0, 1, \ldots, n\}$. 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:

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

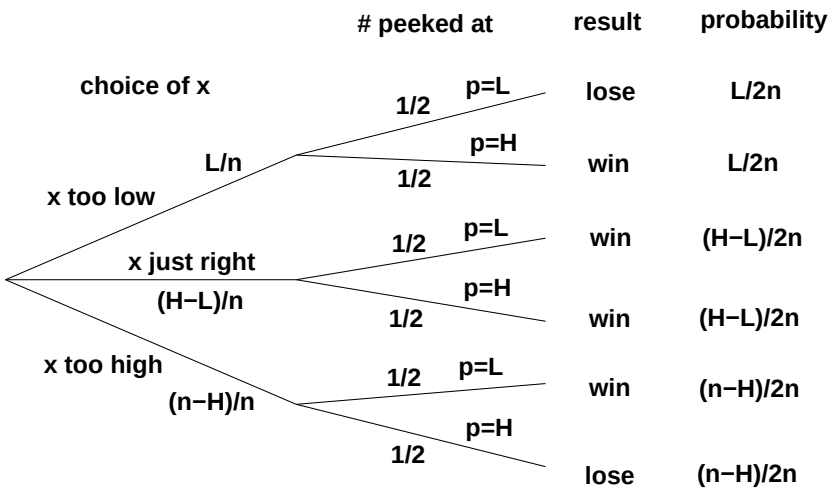
But what probability distribution should you use?

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 evelopes. 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 $0, 1, \ldots, 100$, 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 $0, \ldots, 10$, then your probability of winning rises to 55%! By Las Vegas standards, those are great odds!

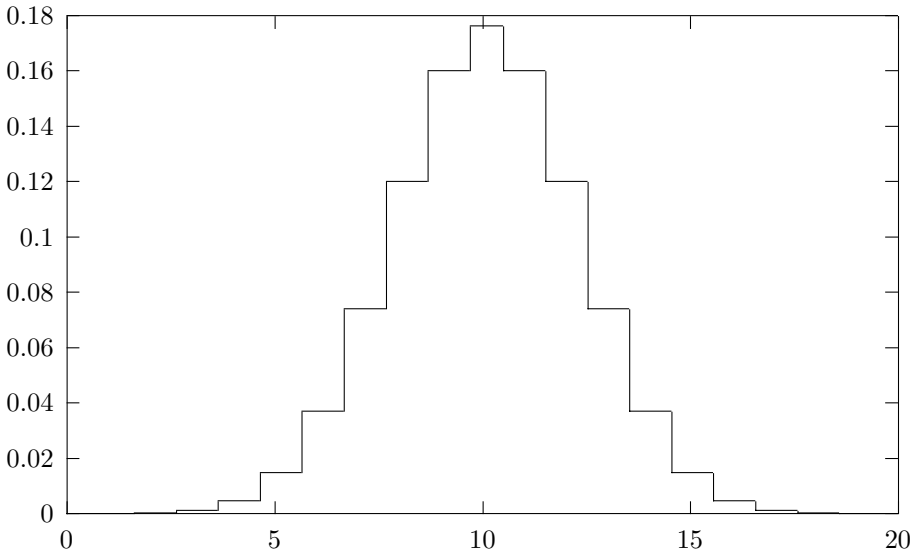
3.4 Binomial Distribution

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.

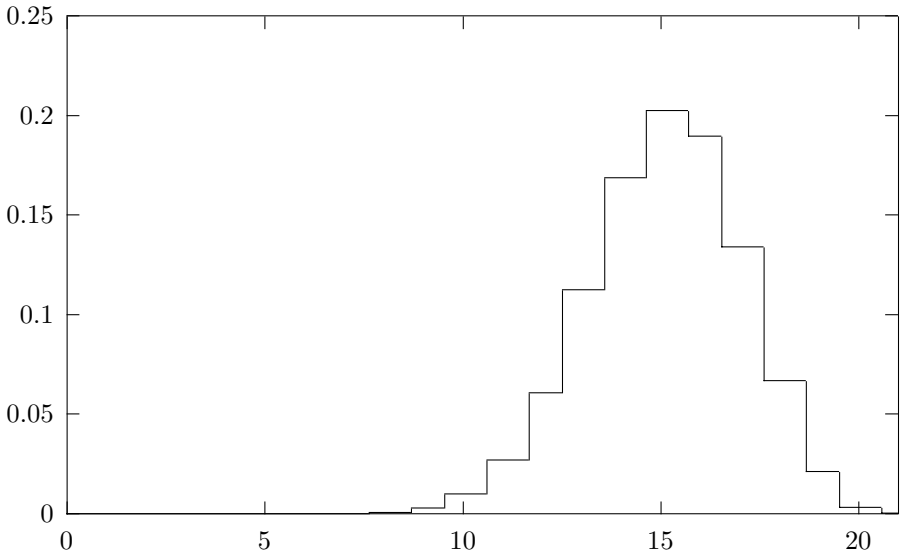
3.4.1 The General Binomial Distribution

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 probabilty $p = 0.75$. 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.



3.4.2 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 αn where α 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 αn heads in n tosses of a coin that comes up heads with probability p is:

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

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%.

3.5 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 (α). 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 summmation 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 (1) 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 α, 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 $n = 100$:

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

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. In our example, the probability of flipping 75 or more heads is the same as the probability of flipping 25 or fewer tails. By the above analysis, this is also extremely small.

3.6 Polling

Suppose we want to estimate the fraction of the U.S. voting population who would favor Hillary Clinton over Rudy Giuliani in the year 2008 presidential election.[2] Let p be this unknown fraction. Let's suppose we have some random process —say throwing darts at voter registration lists— which will select each voter with equal probability. We can define a Bernoulli variable, K, by the rule that $K = 1$ if the random voter most prefers Clinton, and $K = 0$ otherwise.

Now to estimate p, we take a large number, n, of random choices of voters[3] and count the fraction who favor Clinton. That is, we define variables $K_1, K_2, \ldots$, where K_i is interpreted to be the indicator variable for the event that the ith chosen voter prefers Clinton. Since our choices are made independently, the K_i's are independent. So formally,

[2] We can only keep our fingers crossed for this race to happen – when they ran against each other for the U.S. Senate in 2000, they generated some of the best entertainment in TV history.

[3] We're choosing a random voter n times *with replacement*. That is, we don't remove a chosen voter from the set of voters eligible to be chosen later; so we might choose the same voter more than once in n tries! We would get a slightly better estimate if we required n *different* people to be chosen, but doing so complicates both the selection process and its analysis with little gain in accuracy.

we model our estimation process by simply assuming we have mutually independent Bernoulli variables $K_1, K_2, \ldots,$ each with the same probability, p, of being equal to 1. Now let S_n be their sum, that is,

$$S_n ::= \sum_{i=1}^{n} K_i. \tag{2}$$

So S_n has the binomial distribution with parameter, n, which we can choose, and unknown parameter, p.

The variable S_n/n describes the fraction of voters *in our sample* who favor Clinton. We would expect that S_n/n should be something like p. We will use the sample value, S_n/n, as our *statistical estimate* of p.

In particular, suppose we want our estimate of p to be within 0.04 of p at least 95% of the time. Namely, we want

$$\Pr\left\{\left|\frac{S_n}{n} - p\right| \leq 0.04\right\} \geq 0.95\ .$$

Let ϵ be the margin of error we can tolerate, and let δ be the probability that our result lies outside this margin. We're interested in having $\epsilon = 0.04$ and $\delta \leq 0.05$, but the derivation will be clearer if we postpone plugging these values in until the end.

We want to determine the number, n, of times we must poll voters so that the value, S_n/n, of our estimate will, with probability at least $1-\delta$, be within ϵ of the actual fraction in the nation favoring Clinton.

We can define δ, the probability that our poll is off by more than the margin of error ϵ, as follows:

$$\begin{aligned}
\delta &= \underbrace{\Pr\left\{\frac{S_n}{n} \leq p - \epsilon\right\}}_{\substack{\text{too many in sample} \\ \text{prefer ``Giuliani''}}} + \underbrace{\Pr\left\{\frac{S_n}{n} \geq p + \epsilon\right\}}_{\substack{\text{too many in sample} \\ \text{prefer ``Clinton''}}} \\
&= \Pr\{S_n \leq (p-\epsilon)n\} + \Pr\{S_n \geq (p+\epsilon)n\}\ .
\end{aligned}$$

Now

$$\text{CDF}_{S_n}((p-\epsilon)n) ::= \Pr\{S_n \leq (p-\epsilon)n\}$$

Also,

$$\Pr\{S_n \geq (p+\epsilon)n\} = \Pr\{n - S_n \leq ((1-p)-\epsilon)n\}\ .$$

But $T_n ::= n - S_n$ is simply the number of voters in the sample who prefer Giuliani, which is a sum of Bernoulli random variables with parameter $1-p$, and therefore

$$\Pr\{T_n \leq ((1-p)-\epsilon)n\} = \text{CDF}_{T_n}((1-p)-\epsilon)n.$$

Hence

$$\delta = \text{CDF}_{S_n}((p-\epsilon)n) + \text{CDF}_{T_n}(((1-p)-\epsilon)n). \tag{3}$$

So we have reduced getting a good estimate of the required sample size to finding good bounds on two cumulative binomial distributions with parameters p and $1-p$ respectively.

Using the bound on the cumulative binomial distribution function allows us to calculate an expression bounding (3) in terms of n, ϵ and p. The problem is that this bound would contain p, the fraction of Americans that prefer Clinton. This is the unknown number we are trying to determine by polling! Fortunately, there is a simple way out of this circularity. Since (3) is symmetric in p, it has an inflection point when $p = 1/2$, and this inflection point is, in fact, its maximum:

Fact. For all ϵ, n, the maximum value of δ in equation (3) occurs when $p = 1/2$.

In other words, the binomial tails fall off most slowly when $p = 1/2$. Using this fact, and plugging into the equations for $\text{CDF}_{S_n}((p-\epsilon)n)$ and $\text{CDF}_{T_n}(((1-p)-\epsilon)n)$, we get the following theorem:

Theorem 3.1 (Binomial Sampling). *Let $K_1, K_2, \ldots,$ be a sequence of mutually independent 0-1-valued random variables with the same expectation, p, and let*

$$S_n ::= \sum_{i=1}^{n} K_i.$$

Then, for $1/2 > \epsilon > 0$,

$$\Pr\left\{\left|\frac{S_n}{n} - p\right| \geq \epsilon\right\} \leq \frac{1+2\epsilon}{2\epsilon} \cdot \frac{2^{-n(1-H((1/2)-\epsilon))}}{\sqrt{2\pi(1/4-\epsilon^2)n}}. \tag{4}$$

We want $\epsilon = 0.04$, so plugging into (4) gives

$$\delta \leq 13.5 \cdot \frac{2^{-n(0.00462)}}{1.2492\sqrt{n}} \tag{5}$$

where δ is the probability that our estimate is not within ϵ of p. We want to poll enough people so that $\delta \leq 0.05$. The easiest way to find the necessary sample size n is to plug in values for n to find the smallest one where in the righthand side of (5) is ≤ 0.05:

n = people polled	upper bound on probability poll is wrong	
500	9.7%	
600	6.4%	
623	5.9%	
650	5.3%	
664	5.0%	← our poll size
700	4.3%	

So 95% of the time, polling 664[4] people will yield a fraction that is within 0.04 of the actual fraction of voters preferring Clinton.

A remarkable point is that the population of the country has no effect on the poll size! Whether there are a thousand people or a billion in the country, polling only a few hundred is sufficient!

This method of estimation by sampling a quantity —voting preference in this example— is a technique that can obviously be used to estimate many other unknown quantities.

Problem 1. Explaining Sampling to a Jury

We just showed that merely sampling 662 voters will yield a fraction that, 95% of the time, is within 0.04 of the of the actual fraction of voters who prefer Clinton. The actual size of the voting population (10's of millions) was never considered because *it did not matter*.

Suppose you were going to serve as an expert witness in a trial. How would you explain why the number of people necessary to poll *does not depend on the population size*?

4 Confidence Levels

Suppose a pollster uses a sample of 662 random voters to estimate the fraction of voters who prefer Clinton, and the pollster finds that 364 of them prefer Clinton. It's tempting, **but sloppy**, to say that this means:

False Claim. *With probability 0.95, the fraction, p, of voters who prefer Clinton is* $364/662 \pm 0.04$. *Since* $364/662 - 0.04 > 0.50$, *there is a 95% chance that more than half the voters prefer Clinton.*

What's objectionable about this statement is that it talks about the probability or "chance" that a real world fact is true, namely that the actual fraction, p, of voters favoring Clinton is more than 0.50. But p is what it is, and it simply makes no sense to talk about the probability that it is something else. For example, suppose p is actually 0.49; then it's nonsense to ask about the probability that it is within 0.04 of 364/662 —it simply isn't.

A more careful summary of what we have accomplished goes this way:

> We have described a probabilistic procedure for estimating the value of the actual fraction, p. The probability that *our estimation procedure* will yield a value within 0.04 of p is 0.95.

This is a bit of a mouthful, so special phrasing closer to the sloppy language is commonly used. The pollster would describe his conclusion by saying that

> At the 95% *confidence level*, the fraction of voters who prefer Clinton is $364/662 \pm 0.04$.

[4] An exact calculation of the binomial CDF shows that a somewhat smaller poll size of 612 would be sufficient.

It's important to remember that confidence levels refer to the results of estimation procedures for real-world quantities. The real-world quantity being estimated is typically unknown, but fixed; it is not a random variable, so it makes no sense to talk about the probability that it has some property.

5 Expected Value

The ***expectation*** or ***expected value*** of a random variable is a single number that tells you a lot about the behavior of the variable. Roughly, the expectation is the average value, where each value is weighted according to the probability that it comes up. Formally, the expected value of a random variable R defined on a sample space S is:

$$\mathrm{E}[R] = \sum_{w \in S} R(w) \Pr\{w\}$$

To appreciate its signficance, suppose S is the set of students in a class, and we select a student uniformly at random. Let R be the selected student's exam score. Then $\mathrm{E}[R]$ is just the class average— the first thing everyone want to know after getting their test back! In the same way, expectation is usually the first thing one wants to determine about any random variable.

Let's work through an example. Let R be the number that comes up on a fair, six-sided die. Then the expected value of R is:

$$\begin{aligned}
\mathrm{E}[R] &= \sum_{k=1}^{6} k \cdot \frac{1}{6} \\
&= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\
&= \frac{7}{2}
\end{aligned}$$

This calculation shows that the name "expected value" is a little misleading; the random variable might *never* actually take on that value. You can't roll a $3\frac{1}{2}$ on an ordinary die!

5.1 Equivalent Definitions of Expectation

There are some other ways of writing the definition of expectation. Sometimes using one of these other formulations can make computing an expectation a lot easier. One option is to group together all outcomes on which the random variable takes on the same value.

Theorem 5.1.

$$\mathrm{E}[R] = \sum_{x \in \mathrm{range}(R)} x \cdot \Pr\{R = x\}$$

Proof. We'll transform the left side into the right. Let $[R = x]$ be the event that $R = x$.

$$\begin{aligned}
\mathrm{E}[R] &= \sum_{w \in S} R(w) \Pr\{w\} \\
&= \sum_{x \in \text{range}(R)} \sum_{w \in [R=x]} R(w) \Pr\{w\} \\
&= \sum_{x \in \text{range}(R)} \sum_{w \in [R=x]} x \Pr\{w\} \\
&= \sum_{x \in \text{range}(R)} \left(x \cdot \sum_{w \in [R=x]} \Pr\{w\} \right) \\
&= \sum_{x \in \text{range}(R)} x \cdot \Pr\{R = x\}
\end{aligned}$$

On the second line, we break the single sum into two. The outer sum runs over all possible values x that the random variable takes on, and the inner sum runs over all outcomes taking on that value. Thus, we're still summing over every outcome in the sample space exactly once. On the last line, we use the definition of the probability of the event $[R = x]$. □

Corollary 5.2. *If R is a natural-valued random variable, then:*

$$\mathrm{E}[R] = \sum_{i=0}^{\infty} i \cdot \Pr\{R = i\}$$

When you are considering a random variable *that takes on values only in the natural numbers*, $\mathbb{N} ::= \{0, 1, 2, \dots\}$, there is yet another way to write the expected value:

Theorem 5.3. *If R is a natural-valued random variable, then:*

$$\mathrm{E}[R] = \sum_{i=0}^{\infty} \Pr\{R > i\}$$

Proof. Consider the sum:

$$\begin{array}{ccccccc}
\Pr\{R=1\} & + & \Pr\{R=2\} & + & \Pr\{R=3\} & + & \cdots \\
 & + & \Pr\{R=2\} & + & \Pr\{R=3\} & + & \cdots \\
 & & & + & \Pr\{R=3\} & + & \cdots \\
 & & & & & + & \cdots
\end{array}$$

The columns sum to $1 \cdot \Pr\{R = 1\}$, $2 \cdot \Pr\{R = 2\}$, $3 \cdot \Pr\{R = 3\}$, etc. Thus, the whole sum is equal to:

$$\sum_{i=0}^{\infty} i \cdot \Pr\{R = i\} = \mathrm{E}[R]$$

Here, we're using Corollary 5.2. On the other hand, the rows sum to $\Pr\{R > 0\}$, $\Pr\{R > 1\}$, $\Pr\{R > 2\}$, etc. Thus, the whole sum is also equal to:

$$\sum_{i=0}^{\infty} \Pr\{R > i\}$$

These two expressions for the whole sum must be equal, which proves the theorem. □

5.2 Expected Value of an Indicator Variable

The expected value of an indicator random variable for an event is just the probability of that event. (Remember that a random variable I_A is the indicator random variable for event A, if $I_A = 1$ when A occurs and $I_A = 0$ otherwise.)

Lemma 5.4. *If I_A is the indicator random variable for event A, then*

$$\mathrm{E}[I_A] = \Pr\{A\}.$$

Proof.

$$\begin{aligned}\mathrm{E}[I_A] &= 1 \cdot \Pr\{I_A = 1\} + 0 \cdot \Pr\{I_A = 0\} \\ &= \Pr\{I_A = 1\} \\ &= \Pr\{A\}. && \text{(Def. of } I_A\text{)}\end{aligned}$$

□

For example, if A is the event that a coin with bias p comes up heads, $\mathrm{E}[I_A] = \Pr\{I_A = 1\} = p$.

5.3 Mean Time to Failure

Let's look at a problem where one of these alternative definitions of expected value is particularly helpful. A computer program crashes at the end of each hour of use with probability p, if it has not crashed already. What is the expected time until the program crashes?

If we let R be the number of hours until the crash, then the answer to our problem is $\mathrm{E}[R]$. This is a natural-valued variable, so we can use the formula:

$$\mathrm{E}[R] = \sum_{i=0}^{\infty} \Pr\{R > i\}$$

We have $R > i$ only if the system remains stable after i opportunities to crash, which happens with probability $(1-p)^i$. Plugging this into the formula above gives:

$$\begin{aligned} \mathrm{E}[R] &= \sum_{i=0}^{\infty} (1-p)^i \\ &= \frac{1}{1-(1-p)} \\ &= \frac{1}{p} \end{aligned}$$

The closed form on the second line comes from the formula for the sum of an infinite geometric series where the ratio of consecutive terms is $1-p$.

So, for example, if there is a 1% chance that the program crashes at the end of each hour, then the expected time until the program crashes is $1/0.01 = 100$ hours. The general principle here is well-worth remembering: if a system fails at each time step with probability p, then the expected number of steps up to the first failure is $1/p$.

5.3.1 Making a Baby Girl

A couple really wants to have a baby girl. There is a 50% chance that each child they have is a girl, and the genders of their children are mutually independent. If the couple insists on having children until they get a girl, then how many baby boys should they expect first?

This is really a variant of the previous problem. The question, "How many hours until the program crashes?" is mathematically the same as the question, "How many children must the couple have until they get a girl?" In this case, a crash corresponds to having a girl, so we should set $p = 1/2$. By the preceding analysis, the couple should expect a baby girl after having $1/p = 2$ children. Since the last of these will be the girl, they should expect just one boy.

Something to think about: If every couple follows the strategy of having children until they get a girl, what will eventually happen to the fraction of girls born in this world?