Lecture Notes

1 Random Variables

This lecture introduces the idea of a *random variable*. This name is a misnomer, since a random variable is actually a function.

Definition A *random variable* is a function that maps every outcome in the sample space of an experiment to a real number.

For example, consider the experiment of tossing three independent, unbiased coins. The sample space S for this experiment consists of eight outcomes: HHH, HTH, TTT, etc. Let $C : S \mapsto \mathcal{R}$ be the function defined by:

$$C(w) = \text{number of heads appearing in outcome } w \in S$$

Then C is a random variable, because it maps every outcome in the sample space S to a real number. For example, $C(HHH) = 3$, $C(HTH) = 2$, $C(TTT) = 0$, etc.

Similarly, for the same experiment, we can define a random variable $M : S \mapsto \mathcal{R}$ as follows.

$$M(w) = \begin{cases} 1 & \text{if all 3 coins match in outcome } w \in S \\ 0 & \text{otherwise} \end{cases}$$

For example, $M(HHH) = 1$, $M(HTH) = 0$, $M(TTT) = 1$, and so on.

The remainder of this section is divided into short segments that describe basic features of random variables. Throughout, we will use these two random variables, C and M, as examples; keep in mind that C **c**ounts heads and M indicates that all coins **m**atch.

1.1 Indicator Random Variables

The random variable M is an example of an *indicator random variable*.

Definition An *indicator random variable* is a random variable that maps every outcome to either 0 or 1.

Indicator random variables are also called *Bernoulli* or *characteristic* random variables. Typically, indicator random variables identify all outcomes that share some characteristic; outcomes with the characteristic are mapped to 1, and outcomes without the characteristic are mapped to 0. For example, the random variable M indicates outcomes with the characteristic that all three coins match.

1.2 Events Defined by a Random Variable

There is a natural relationship between random variables and events. Recall that an event is just a subset of the outcomes in the sample space of an experiment.

The relationship is simplest for an indicator random variable. An indicator random variable partitions the sample space into two blocks: outcomes mapped to 1 and outcomes mapped to 0. These two sets of outcomes are events. For example, the random variable M partitions the sample space as follows:

$$\underbrace{HHH \quad TTT}_{\text{mapped to 1}} \quad \underbrace{HHT \quad HTH \quad HTT \quad THH \quad THT \quad TTH}_{\text{mapped to 0}}$$

Thus, the random variable M defines two events, the event that all coins match (denoted $M = 1$) and the event that not all coins match (denoted $M = 0$).

A general random variable may partition the sample space into many blocks. A block contains all outcomes mapped to the same value by the random variable. Each block is a set of outcomes and therefore an event. For example, the random variable C partitions the sample space into four blocks:

$$\underbrace{TTT}_{\text{mapped to 0}} \quad \underbrace{TTH \quad THT \quad HTT}_{\text{mapped to 1}} \quad \underbrace{THH \quad HTH \quad HHT}_{\text{mapped to 2}} \quad \underbrace{HHH}_{\text{mapped to 3}}$$

Thus, the random variable C defines four events, the event that no coin is heads (denoted $C = 0$), the event that one coin is heads ($C = 1$), the event that two coins are heads ($C = 2$), and the event that three coins are heads ($C = 3$).

We can define other events in terms of a random variable as well. For example, the event $C \geq 2$, consists of all outcomes mapped to 2 or more. In general, if A is a set of real numbers, the event $C \in A$ consists of all outcomes mapped to an element of A. For example, $C \in \{1, 3\}$ is the event that there are an odd number of heads.

1.3 Probability of Events Defined by a Random Variable

Recall that the probability of an event is the sum of the probabilities of the outcomes it contains. From this rule, we can compute the probability of various events associated with a random variable. For example, if $R : S \mapsto \mathcal{R}$ is a random variable and x is a real number, then

$$\Pr(R = x) \quad = \quad \sum_{w \in S : R(w) = x} \Pr(w)$$

For example, we can compute $\Pr(C = 2)$ as follows:

$$\begin{aligned}
\Pr(C=2) &= \sum_{w \in S : C(w)=2} \Pr(w) \\
&= \Pr(THH) + \Pr(HTH) + \Pr(HHT) \\
&= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\
&= \frac{3}{8}
\end{aligned}$$

Here S is the sample space $\{HHH, HHT, HTH, \ldots\}$. The first equation uses the definition of the probability of an event. In the second step, we identify the three outcomes corresponding to terms in the summation. In the third step, we observe that every outcome has probability $\frac{1}{8}$, since the three coins are fair and independent.

Similarly, we can compute $\Pr(M = 1)$ as follows

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

The justification for each step is the same as before. We can find the probability of the event that $C \geq 2$ in the same way:

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

If the range of R is $\mathcal{N}$, an expression of the form $\Pr(R \geq x)$ can also be evaluated with the summation $\sum_{i \geq x} \Pr(R = i)$. That is, instead of summing over outcomes mapped to at least x as above, we can sum probabilities over all values of the random variable that are greater than or equal to x. The result is the same, since both summations cover the same outcomes.

Finally, we find the probability of the event that $C \in \{1, 3\}$.

$$\begin{aligned}
\Pr(C \in \{1,3\}) &= \sum_{w \in S : C(w) \in \{1,3\}} \Pr(w) \\
&= \Pr(TTH) + \Pr(THT) + \Pr(HTT) + \Pr(HHH) \\
&= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\
&= \frac{1}{2}
\end{aligned}$$

As in the preceding example, an expression of the form $\Pr(R \in A)$ can also be evaluated by summing probabilities over values in A, instead of by summing over outcomes. That is, $\Pr(R \in A) = \sum_{a \in A} \Pr(R = a)$.

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 all three coins are not the same ($M = 0$). We can compute this probability using the familiar Product Rule:

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

1.5 Independence

The notion of independence does not carry over to random variables so easily. In analogy to the last lecture, we will define independence for a pair of random variables and then define mutual independence for two or more random variables.

Definition Two random variables R_1 and R_2 are *independent* if for all $x_1, x_2 \in \mathcal{R}$, we have:

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

There is an equivalent definition for the independence of two random variables in terms of conditional probability. We will make use of both definitions.

Definition Two random variables R_1 and R_2 are *independent* if for all $x_1, x_2 \in \mathcal{R}$ such that $\Pr(R_2 = x_2) \neq 0$, we have:

$$\Pr(R_1 = x_1 \mid R_2 = x_2) = \Pr(R_1 = x_1)$$

The second definition may be more intuitive; it says that the probability that R_1 has a particular value is unaffected by the value of R_2.

1.5.1 Proving that Two Random Variables are Not Independent

Are C and M independent? Intuitively, no; the number of heads (C) not only affects, but completely determines whether all three coins match (M). To prove this, let's use the first definition of independence. We must find some $x_1, x_2 \in \mathcal{R}$ such that the condition in the first definition is false. For example, the condition does not hold for $x_1 = 2$ and $x_2 = 1$.

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

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.

1.5.2 A Dice Example

Suppose that we roll two fair, independent dice. We can regard the numbers that turn up as random variables, D_1 and D_2. For example, if the outcome is $w = (3, 5)$, then $D_1(w) = 3$ and $D_2(w) = 5$.

Let $T = D_1 + D_2$. Then T is also a random variable, since it is a function mapping each outcome to a real number, namely the sum of the numbers shown on the two dice. For outcome $w = (3, 5)$, we have $T(w) = 3 + 5 = 8$.

Define S as follows:

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

That is, $S = 1$ if the sum of the dice is 7, and $S = 0$ if the sum of the dice is not 7. For example, for outcome $w = (3, 5)$, we have $S(w) = 0$, since the sum of the dice is 8. Since S is a function mapping each outcome to a real number, S is also a random variable. In particular, S is an indicator random variable, since every outcome is mapped to 0 or 1.

The definitions of random variables T and S illustrate a general rule: *any function of random variables is also random variable.*

Are D_1 and T independent? That is, is the sum of the two dice (T) independent of the outcome of the first die (D_1)? Intuitively, the answer appears to be no! To prove this let's use the second definition of independence. We must find $x_1, x_2 \in \mathcal{R}$ such that $\Pr(x_2) \neq 0$ and the condition in the second definition does not hold. For example, we can choose $x_1 = 2$ and $x_2 = 3$.

$$\Pr((T = 2) \mid (D_1 = 3)) = 0 \quad \text{but} \quad \Pr(T = 2) = \frac{1}{36} \neq 0$$

The first probability is zero, since if we roll a three on the first die ($D_1 = 3$), then there is no way that the sum of both dice is two ($T = 2$). On the other hand, if we throw both dice, the probability that the sum is two is $\frac{1}{36}$, since we could roll two ones.

Are S and D_1 independent? That is, is the probability that the sum of both dice is seven (S) independent of the outcome of the first die (D_1)? Once again, intuition suggests that the answer is "no". Surprisingly, however, these two random variables *are* actually independent!

Proving that two random variables are independent requires some work. Let's use the second definition of independence. We must show that for all x_1, x_2 in $\mathcal{R}$ such that $\Pr(D_1 = x_2) \neq 0$, we have:

$$\Pr(S = x_1 \mid D_1 = x_2) = \Pr(S = x_1)$$

If x_1 is neither 0 nor 1, then the condition holds trivially because both sides are zero. Furthermore, the condition $\Pr(D_1 = x_2) \neq 0$ holds only when x_2 is 1, 2, 3, 4, 5, or 6. Still, we must check the above condition in $2 \cdot 6 = 12$ cases. Two observations make this easier.

First, there are $6 \cdot 6 = 36$ outcomes in the sample space for this experiment. The outcomes are equiprobable, so each outcome has probability $\frac{1}{36}$. The two dice sum to seven in six outcomes: $1 + 6$, $2 + 5$, $3 + 4$, $4 + 3$, $5 + 2$, and $6 + 1$. Therefore, the probability of rolling a seven, $\Pr(S = 1)$, is $\frac{6}{36} = \frac{1}{6}$.

Second, after we know the result of the first die, there is always exactly one value for the second die that makes the sum seven. For example, if the first die is 2, then the sum is seven only if the second die is a 5. Therefore, $\Pr(S = 1 \mid D_1 = x_2) = \frac{1}{6}$ for $x_2 = 1, 2, 3, 4, 5$, or 6.

These two observations establish the independence condition in six cases:

$$\begin{aligned}
\Pr(S = 1 \mid D_1 = 1) = {} & \quad \tfrac{1}{6} \quad = \Pr(S = 1) \\
\Pr(S = 1 \mid D_1 = 2) = {} & \quad \tfrac{1}{6} \quad = \Pr(S = 1) \\
\cdots \quad & \quad \cdots \quad \cdots \\
\Pr(S = 1 \mid D_1 = 6) = {} & \quad \tfrac{1}{6} \quad = \Pr(S = 1)
\end{aligned}$$

The remaining cases are complementary to the the first six. For example, we know that $\Pr(S = 0) = \frac{5}{6}$, since the complementary event, $S = 1$, has probability $\frac{1}{6}$.

$$\begin{aligned}
\Pr(S = 0 \mid D_1 = 1) = {} & \quad \tfrac{5}{6} \quad = \Pr(S = 0) \\
\Pr(S = 0 \mid D_1 = 2) = {} & \quad \tfrac{5}{6} \quad = \Pr(S = 0) \\
\cdots \quad & \quad \cdots \quad \cdots \\
\Pr(S = 0 \mid D_1 = 6) = {} & \quad \tfrac{5}{6} \quad = \Pr(S = 0)
\end{aligned}$$

We have established that the independence condition holds for all necessary $x_1, x_2 \in \mathcal{R}$. This proves that S and D_1 are independent after all!

1.6 Mutual Independence

The definition of mutual independence for random variables is similar to the definition for events.

Definition Random variables $R_1, R_2, \ldots R_n$ are *mutually independent* if for all $x_1, x_2, \ldots x_n$, we have:

$$\Pr(R_1 = x_1 \cap R_2 = x_2 \cap \ldots \cap R_n = x_n) = \prod_{i=1}^{n} \Pr(R_i = x_i)$$

1.7 Probability Distributions

A random variable is a function from the sample space of an experiment to the real numbers. As a result, every random variable is bound up in some particular experiment. Often, however, we want to describe a random variable independent of any experiment. This consideration motivates the following two definitions.

Definition The *probability distribution function (pdf)* for a random variable $R : S \mapsto \mathcal{R}$ is the function $f : \mathcal{R} \mapsto [0,1]$ defined by:

$$f(x) = \Pr(R = x)$$

The probability distribution function is also sometimes called the point distribution function. A consequence of this definition is that $\sum_x f(x) = 1$, since we are summing the probabilities of all outcomes in the sample space.

Definition The *cumulative distribution function* for a random variable $R : S \mapsto \mathcal{R}$ is the function $F : \mathcal{R} \mapsto [0,1]$ defined by:

$$\begin{aligned} F(x) &= \Pr(R \leq x) \\ &= \sum_{y \leq x} f(y) \end{aligned}$$

Note that neither the probability distribution function nor the cumulative distribution function involves the sample space of an experiment; both are functions from $\mathcal{R}$ to $[0,1]$. This allows us to study random variables without reference to a particular experiment. In particular, we will look at three distributions today and will see more in upcoming lectures.

Bernoulli Distribution

For our first example, let R be a Bernoulli random variable that is 0 with probability p and 1 with probability $1-p$. We can compute the probability distribution function f at 0 and 1 as follows:

$$\begin{aligned} f(0) &= \Pr(R = 0) = p \\ f(1) &= \Pr(R = 1) = 1 - p \end{aligned}$$

Similarly, we can compute the cumulative distribution function F:

$$\begin{aligned} F(0) &= \Pr(R \leq 0) = p \\ F(1) &= \Pr(R \leq 1) = 1 \end{aligned}$$

Uniform Distribution

Now let R be a random variable that is uniform on $[1, N]$. That is, R takes on value k with probability $\frac{1}{N}$ for all $1 \leq k \leq N$. The probability distribution function and the cumulative distribution function are given below.

$$\left.\begin{aligned} f(k) &= \Pr(R = k) = \tfrac{1}{N} \\ F(k) &= \Pr(R \leq k) = \tfrac{k}{N} \end{aligned}\right. \qquad \text{where } 1 \leq k \leq N$$

Uniform distributions are very common. For example, the outcome of a fair die is uniform on $[1, 6]$.

2 The Numbers Game

Here is a game that turns out to involve random variables with the uniform distribution.

There are two envelopes. Each contains a number in the range $0, 1, \ldots 100$, and the two numbers are distinct. To win the game, we must determine which envelope contains the larger number. Our only advantage is that we are allowed to peek at the number in one envelope; we can choose which one. Can we devise a strategy that gives us a better than 50% chance of winning?

For example, suppose we are playing the game and are shown the two envelopes. Now we could guess randomly which envelope contains the larger number, and not even bother to peek in one envelope. With this strategy, we have a 50% chance of winning.

Suppose we try to do better. We peek in the left envelope and see the number 12. Since 12 is a small number, we guess that the right envelope probably contains the larger number. Now, we might be correct. On the other hand, maybe the the person who wrote the numbers decided to be tricky, and made *both* numbers small! Then our guess is not so good!

An important point to remember is that the numbers in the envelope may *not* be random. We should assume that the person who writes the numbers is trying to defeat us; he may use randomness or he may not— we don't know!

2.1 Intuition Behind the Winning Strategy

Amazingly, there is a strategy that wins more than 50% of the time, regardless of the numbers in the envelopes.

Suppose we somehow knew a number x between the larger and smaller number. Now we peek in an envelope and see some number. If this number is larger than x, then it must be the larger number. If the number we see is smaller than x, then the larger number must be in the other envelope. In other words, if we know x, then we are guaranteed to win.

Of course, we do not know the number x, so what can we do? Guess!

There is some probability that we guess x correctly. If so, then we are guaranteed to win. If we guess incorrectly, then we are no worse off than before; our chance of winning is still 50%. Combining these two cases, our overall chance of winning is better than 50%!

This argument sounds almost comically implausible, but we will rigorously justify it in the next section!

The key is *how* we guess the number x. That is, what is the probability distribution function of x? The best answer turns out to be a uniform distribution.

2.2 Analysis of the Winning Strategy

Let's formally state the strategy described in the preceding section and then compute our chance of winning.

Call the numbers in the envelopes y and z and suppose $y < z$. For generality, suppose that each number is in the range $0, 1, \ldots, n$. Above, we considered the case $n = 100$. The number we see by peeking is denoted r. Here is the winning strategy:

1. Guess a number x from $\{\frac{1}{2}, 1\frac{1}{2}, 2\frac{1}{2}, \ldots, n - \frac{1}{2}\}$ with the uniform distribution. That is, each value is selected with probability $\frac{1}{n}$. (We constrain x to be something-and-a-half to avoid ties.)
2. Peek into a random envelope. We see a value r that is either y or z. Each envelope is chosen with probability $\frac{1}{2}$, and the choice is independent of the number x.
3. Hope that $y < x < z$. If $r > x$, then guess that the larger number is the envelope we peeked into. If $r < x$, then guess that the larger number is in the other envelope.

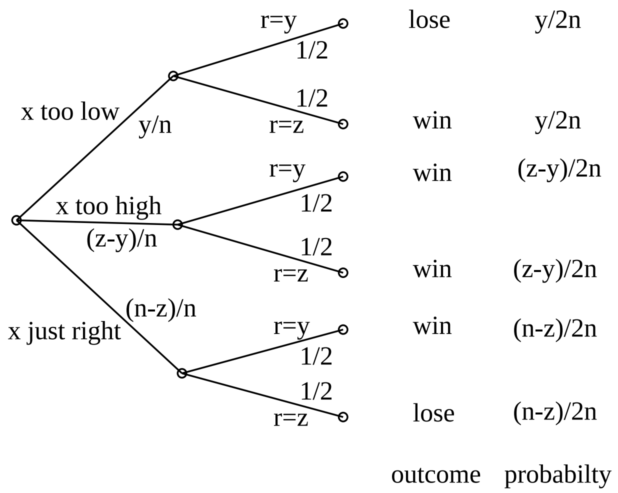

Figure 1: *This is the tree diagram for the Numbers Game.*

We can compute the probability of winning by using the tree diagram in Figure 1 and the usual four-step method.

Step 1: Find the sample space. We either choose x too low, too high, or just right. Then we either choose $r = y$ or $r = z$. As indicated in the figure, this gives a total of six outcomes.

Step 2: Define events of interest. We are interested in the event that we correctly pick the larger number. This event consists of four outcomes, which are marked "win" in the figure.

Step 3: Compute outcome probabilities. As usual, we first assign probabilities to edges. First, we guess x. The probability that our guess of x is too low is $\frac{y}{n}$, the probability that our guess is too high is $\frac{n-z}{n}$, and the probability of a correct guess is $\frac{z-y}{n}$. We then select an envelope; $r = y$ and $r = z$ occur with equal probability, independent of the choice of x. The probability of an outcome is the product of the probabilities on the corresponding root-to-leaf path, as shown in the figure.

Step 4: Compute event probabilities. The probability of winning is the sum of the probabilities of the four winning outcomes. This gives:

$$\begin{aligned}
\Pr(\text{winning}) &= \frac{y}{2n} + \frac{z-y}{2n} + \frac{z-y}{2n} + \frac{n-z}{2n} \\
&= \frac{n+z-y}{2n} \\
&= \frac{1}{2} + \frac{z-y}{2n} \\
&\geq \frac{1}{2} + \frac{1}{2n}
\end{aligned}$$

In the final equality, we use the fact that the larger number z is at least 1 greater than the smaller number y, since they must be distinct.

We conclude that the probability of winning with this strategy is at least $\frac{1}{2} + \frac{1}{2n}$, regardless of the numbers in the envelopes!

For example, if the numbers in the envelopes are in the range $0, \ldots 100$, then the probability of winning is at least $\frac{1}{2} + \frac{1}{200} = 50.5\%$. Even better, if the numbers are constrained to be in the range $0, \ldots, 10$, then the probability of winning rises to 55%! By Las Vegas standards, these are great odds!

2.3 Optimality of the Winning Strategy

What strategy should our opponent use in putting the numbers into the envelopes? That is, how can he ensure that we do not get, say, a 60% chance of winning?

Of course, our opponent could try to be clever, putting in two low numbers and then two high numbers, etc. But then there is no guarantee that we will not catch on and start winning every time!

It turns out that our opponent should also use a randomized strategy involving the uniform distribution. In particular, he should choose y from $[0, n-1]$ uniformly, and then let $z = y + 1$. That is, he should randomly choose a pair of consecutive numbers like $(6, 7)$ or $(73, 74)$ with the uniform distribution. In homework, we will prove that this strategy is optimal:

Claim 2.1 If the opponent uses the strategy above, then $\Pr(\text{we win}) \leq \frac{1}{2} + \frac{1}{2n}$ for every strategy we can adopt.

In summary, we can win with probability *at least* $\frac{1}{2} + \frac{1}{2n}$ regardless of what our opponent does, and our opponent can ensure that we win with probability *at most* $\frac{1}{2} + \frac{1}{2n}$ regardless of what we do.

The homework also considers the case where our opponent is allowed to put any non-negative integers into the envelopes; he can use numbers as large as he likes. This case is more complicated, because the uniform distribution on $\mathcal{N}$ is not even well-defined. Nevertheless, there is a strategy that guarantees a better than 50% chance of winning regardless of what numbers are in the envelopes!

3 The Binomial Distribution

We now introduce a third distribution, called the *binomial distribution*. This is the most important and commonly occurring distribution in computer science. We will actually define one important special case of the binomial distribution and then define the general case.

Definition The *unbiased binomial distribution* is the function $f_n : \mathcal{R} \mapsto [0, 1]$ defined by

$$f_n(k) = \binom{n}{k} \frac{1}{2^n}$$

where n is a parameter that is at least 1.

Definition The *general binomial distribution* is the function $f_{n,p} : \mathcal{R} \mapsto [0,1]$ defined by

$$f_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where n and p are parameters such that $n \geq 1$ and $0 < p < 1$.

(In both definitions, if k is not an integer $0, 1, \ldots, n$, then f is zero.)

The unbiased binomial distribution is a special case of the general binomial distribution where the parameter p is equal to $\frac{1}{2}$.

3.1 The Troubled Space Station Mir

Suppose that the Mir space station has n parts, each of which is faulty with probability p. Furthermore, assume that faults occur independently. Let the random variable R be the number of faulty parts. What is the probability distribution of R?

Since the probability distribution function is defined by $f(k) = \Pr(R = k)$, our problem is to find $\Pr(R = k)$. We can do this with the usual four-step method, though we will not draw a tree diagram.

Step 1: Find the sample space. We can characterize Mir with a string of W's and F's of length n. A W in the i-th position indicates that the i-th part is working, and an F indicates that the i-th part is faulty. Each such string is an outcome, and the sample space S is the set of all 2^n such strings.

Step 2: Define events of interest. We want to find the probability that there are exactly k faulty parts; that is, we are interested in the event that $R = k$.

Step 3: Compute outcome probabilities. Since faults occur independently, the probability of an outcome such as $FWFWW$ is simply a product such as $p(1-p)p(1-p)(1-p) = p^2(1-p)^3$. Each F contributes a p term and each W contributes a $(1-p)$ term. In general, the probability of an outcome with k faulty parts and $n-k$ working parts is $p^k(1-p)^{n-k}$.

Step 4: Compute event probabilities.

We can compute the probability that k parts are faulty as follows:

$$\begin{aligned}
\Pr(R = k) &= \sum_{\substack{w \in S: \\ w \text{ has } k\ F\text{'s}}} p^k (1-p)^{n-k} \\
&= (\#\text{ of length } n \text{ strings with } k\ F\text{'s}) \cdot p^k (1-p)^{n-k} \\
&= \binom{n}{k} p^k (1-p)^{n-k}
\end{aligned}$$

The first equation uses the definition of the probability of an event. The second line follows because all terms in the summation are equal. In the final line we use the fact that there are $\binom{n}{k}$ strings of length n with k F's.

We can now see that the probability distribution function for the number of faulty parts is precisely the general binomial distribution:

$$f(k) = \Pr(R = k) = \binom{n}{k} p^k (1-p)^{n-k} = f_{n,p}(k)$$

In general, the binomial distribution arises whenever we have n independent Bernoulli variables with the same distribution. In this case, the Bernoulli variables indicated whether a part was faulty or not. As another example, if we flip n fair coins, then the number of heads is an unbiased binomial distribution.

3.2 The Shape of the Binomial Distribution

The binomial distribution is somewhat complicated. For example, it is not even immediately clear that $\sum_k f_{n,p}(k) = 1$. This fact follows from the Binomial Theorem:

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

More generally, we would like to know the shape of the binomial distribution. That is, what is the value of $f_{n,p}(k)$ for particular n, p, and k? Of course, we could simply plug in the values of these variables, but a simpler, approximate expression would be nice.

For example, if we flip 100 fair coins, what is the probability that we get exactly 50 heads? What is the probability that we get 25 heads or fewer?

Approximating $f_{n,p}(k)$

We can approximate the value of $f_{n,p}(k)$ using Stirling's formula. For convenience, we set $k = \alpha n$ where α is a number between 0 and 1. The result is:

$$f_{n,p}(\alpha n) = \frac{2^{\left(\alpha \log_2\left(\frac{p}{\alpha}\right) + (1-\alpha)\log_2\left(\frac{1-p}{1-\alpha}\right)\right)\cdot n} \overbrace{e^{a_n - a_{\alpha n} - a_{n-\alpha n}}}^{\sim 1}}{\sqrt{2\pi\alpha(1-\alpha)n}}$$

This expression looks nasty, but is really very useful. As usual, the a_i symbols arise from the error in Stirling's approximation; a_i denotes a value between $\frac{1}{12i+1}$ and $\frac{1}{12i}$.

The Maximum Value of $f_{n,p}(k)$

The maximum value of $f_{n,p}(\alpha n)$ occurs when $\alpha = p$. This matches intuition. For example, in the Mir problem, each part is faulty with probability p, so we would expect exactly pn faulty parts to be the likeliest case. Substituting, $\alpha = p$ into our approximation for $f_{n,p}(\alpha n)$ gives:

$$f_{n,p}(pn) \leq \frac{1}{\sqrt{2\pi p(1-p)n}}$$

The two sides of this inequality are actually asymptotically equal. We can use this formula to find the probability that exactly 50 heads come up in 100 tosses of a fair coin by substituting $n = 100$ and $p = \frac{1}{2}$:

$$\begin{aligned}
\Pr(50 \text{ heads}) &= f_{100,\frac{1}{2}}\left(\frac{1}{2}n\right) \\
&\leq \frac{1}{\sqrt{50\pi}} \approx 0.079788
\end{aligned}$$

The probability of throwing exactly 50 heads in 100 tosses is at most 8%! In fact, the bound given above is very close to the true value; in this case, the exact answer is $0.079589\ldots$

We can also compute the probability of throwing exactly 25 heads in 100 tosses. In this case, we substitute $n = 100$, $p = \frac{1}{2}$, and $\alpha = \frac{1}{4}$ into the formula for $f_{n,p}(\alpha n)$:

$$\begin{aligned}
f_{n,p}(\alpha n) &= \frac{2^{\left(\alpha \log_2\left(\frac{p}{\alpha}\right)+(1-\alpha)\log_2\left(\frac{1-p}{1-\alpha}\right)\right)\cdot n} e^{a_n - a_{\alpha n} - a_{n-\alpha n}}}{\sqrt{2\pi\alpha(1-\alpha)n}} \\
&= \frac{2^{(-0.1887)\cdot 100} \cdot 0.99\ldots}{10.85\ldots} \\
&\approx 1.913 \cdot 10^{-7}
\end{aligned}$$

The odds are less than 1 in 5 million for throwing exactly 25 heads in 100 tosses!

The main term in our approximation of $f_{n,p}(\alpha n)$ is the power of 2. If $p = \alpha$, then this term is 1. However, if $p \neq \alpha$, then this term is of the form 2^{-cn} for some $c > 0$. As a consequence, when n grows large, $f_{n,p}(\alpha n)$ shrinks exponentially. This suggests that the binomial distribution has the shape shown in Figure 2.

The Cumulative Distribution Function

What is the probability of tossing 25 or fewer heads? Of course, we could sum the probability of zero heads, one head, two heads, ..., and 24 heads. But there is also a simple formula in terms of the probability distribution function.

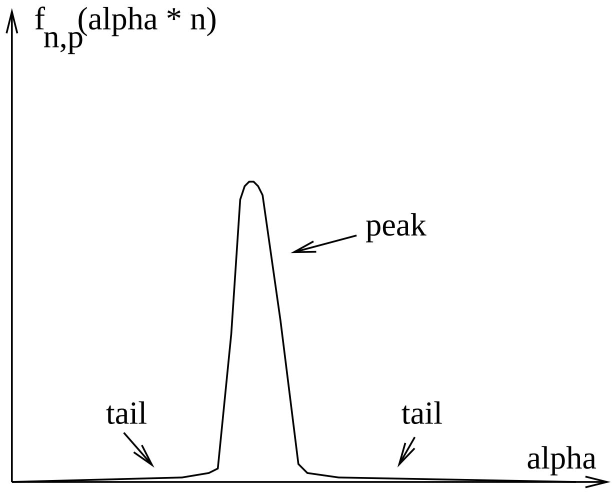

Figure 2: *This diagram shows the approximate shape of the binomial distribution function, $f_{n,p}(\alpha n)$. The central peak is centered at $\alpha = pn$ and has height $\Theta(\frac{1}{\sqrt{n}})$ and width $\Theta(\sqrt{n})$. The "tails" on either side fall off very quickly.*

Theorem 3.1

$$F_{n,p}(\alpha n) \leq \frac{1-\alpha}{1-\frac{\alpha}{p}} f_{n,p}(\alpha n) \qquad \text{for } \alpha < p$$

We can compute the probability of throwing 25 or fewer heads by plugging in the values $n = 100$, $\alpha = \frac{1}{4}$, and $p = \frac{1}{2}$. This gives:

$$\begin{aligned} \Pr(\text{at most 25 heads}) &= F_{100,\frac{1}{2}}(\frac{1}{4} \cdot 100) \\ &\leq \frac{3/4}{1/2} f_{100,\frac{1}{2}}(25) \\ &= \frac{3}{2} \cdot 1.913 \ldots \cdot 10^{-7} \end{aligned}$$

In other words, the probability of throwing 25 or fewer heads is at most 1.5 times the probability of throwing exactly 25 heads. Therefore, we are at least twice as likely to throw exactly 25 heads as to throw 24 or fewer! This is somewhat surprising; the cases of 0 heads, 1 head, 2 heads, ..., 24 heads are all together less likely than the single case of 25 heads. This shows how quickly the tails of the binomial distribution fall off!

3.3 Transmission Across a Noisy Channel

Suppose that we are transmitting bits across a noisy channel. (For example, say your modem uses a phone line that faintly picks up a local radio station.) Suppose we transmit 10,000 bits, and each arriving bit is incorrect with probability 0.01. Assume that these errors occur independently. What is the probability that more than 2% of the bits are erroneous?

We can solve this problem using Theorem 3.1. However, one trick is required. The theorem only holds if $\alpha < p$; therefore, we have to work in terms of correct bits instead of erroneous bits.

$$\begin{aligned}
\Pr(> \text{ than } 2\% \text{ errors}) &= \Pr(\leq \ 98\% \text{ correct}) \\
&= F_{n,0.99}(0.98n) \\
&\leq 1.98\frac{2^{-0.005646 \cdot 10{,}000}}{0.3509\sqrt{10,000}} \\
&\leq 2^{-60}
\end{aligned}$$

The probability that more than 2% of the bits are erroneous is incredibly small! This again demonstrates the extreme improbability of outcomes on the tails of the binomial distribution.

3.4 Polling

Another good example of binomial distributions comes up in polling. The Gallup polling service reported that on November 10, only 37% of American adults thought Louise Woodward was guilty. Furthermore, Gallup asserted that since they asked the opinions of 623 adults, "one can say with 95 percent confidence that the error attributable to sampling and other random effects is within plus or minus 4 percentage points". Can we confirm this claim?

We want to determine p, the fraction of Americans that think Louise is guilty. Our plan is to sample the opinion of n people chosen uniformly and at random, *with replacement.* That is, we might poll the same person twice! This may seem inefficient, but it simplifies the analysis. If G is the number of people in our sample that say Louise is guilty, then we will claim that p is about $\frac{G}{n}$.

The problem is that our sample might not be representative. For example, maybe everyone in the country outside of North Dakota thinks Louise is guilty, but by bad luck the sample contained mostly North Dakotans. Then our poll would give the wrong answer.

Let ϵ be the margin of error we can tolerate, and let δ be the probability that our result lies outside this margin. How many people must we poll so that our result is within ϵ percent of national opinion with probability at least $1 - \delta$? For example, Gallup claims that for $\epsilon = 0.04$ and $\delta = 0.05$, polling 623 people is sufficient.

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

$$\begin{aligned}
\delta &= \underbrace{\Pr\left(\frac{G}{n} < p - \epsilon\right)}_{\substack{\text{too many in sample} \\ \text{say ``not guilty''}}} + \underbrace{\Pr\left(\frac{G}{n} > p + \epsilon\right)}_{\substack{\text{too many in sample} \\ \text{say ``guilty''}}} \\
&= \Pr(G < (p - \epsilon)n) + \Pr(G > (p + \epsilon)n)
\end{aligned}$$

Each term in the definition of δ can be evaluated using Theorem 3.1. In the second term, we must use the same trick as in the Noisy Channel problem to ensure that $\alpha < p$. We observe that $\Pr(\frac{G}{n} > p + \epsilon) = \Pr(\frac{n-G}{n} < 1 - p - \epsilon)$, where $\frac{n-G}{n}$ is the fraction of people polled who say that Louise is not guilty, and $1 - p$ is the fraction of all Americans who say that she is not guilty. This gives:

$$\delta \leq F_{n,p}((p-\epsilon)n) + F_{n,1-p}((1-p-\epsilon)n)$$

This is an expression for the probability that our poll is off by more than the margin of error. The problem is that the expression contains p, the fraction of Americans that think Louise is guilty. This is the number we are trying to determine by polling!

Fortunately, we can upper bound δ by using the following fact:

Fact 1 For all ϵ, the maximum value of δ occurs when $p = \frac{1}{2}$.

The fact implies that to get an upper bound on δ, we can pretend that half of the people think Louise is guilty. This gives:

$$\begin{aligned}
\delta &\leq F_{n,\frac{1}{2}}((\frac{1}{2}-\epsilon)n) + F_{n,1-\frac{1}{2}}((1-\frac{1}{2}-\epsilon)n) \\
&= 2F_n((\frac{1}{2}-\epsilon)n)
\end{aligned}$$

Now suppose that we want a margin of error of 4% as Gallup claimed. Plugging in $\epsilon = 0.04$ gives:

$$\begin{aligned}
\delta &\leq 2F_n(0.46n) \\
&\leq 2 \cdot 6.75 \cdot \frac{2^{-0.004622\ldots \cdot n}}{1.2492\ldots\sqrt{n}}
\end{aligned}$$

We want to poll enough people so that δ is less than 0.05. The easiest way is to plug in values for n, the number of people polled:

n = people polled	upper bound on probability poll is wrong		
500	9.7%		
600	6.4%		
623	5.9%	$\leftarrow$	Gallup's poll size
650	5.3%		
662	5.0%	$\leftarrow$	our poll size
700	4.3%		

Gallup's poll size is just about right. By our calculation, polling 662 people is sufficient to determine public opinion to within 4% with confidence of 95%. We can be certain that Gallup's poll of 623 people gives 95% confidence with at most a 4.13% margin of error. But the real margin of error may be less, since we made several approximations. Still, our approximations must be quite good, since 4.13% is quite close to Gallup's claim of 4%.

The 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!