16 Random Variables and Distributions

Thus far, we have For example, we computed the probability

, or

~~So far we~~ focused on probabilities of events, that you win the Monty Hall game, that

But, in many cases, we would

you have a rare medical condition, given that you tested positive. ~~Now we focus on~~

like to know more. For example,

~~quantitative questions:~~ *How many* contestants must play the Monty Hall game until one

gambling with

of them finally wins? *How long* will this condition last? *How much* will I lose ~~playing~~

~~with~~ strange dice all night? To answer such questions, we

si~~lly Math games all day?~~ ~~Random variables are the mathematical tool for addressing~~

need to work with ~~a new tool:~~ random variables.

~~such questions, and in this chapter we work out their basic properties, especially prop-~~

~~erties of their *mean or expected value*.~~

~~like to know more. For example, how long will the illness last or how much~~

16.1 ~~Random Variable Examples~~ Definitions and Examples

Definition 16.1.1. A *random variable*, R, on a probability space is a total function whose domain is the sample space.

The codomain of R can be anything, but will usually be a subset of the real numbers. Notice that the name "random variable" is a misnomer; random variables are actually functions!

For example, suppose we toss three ~~mutually~~ independent,[1] 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

---

1 Going forward, when we talk about flipping independent coins, we will assume that they are mutually independent.

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

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

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

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

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

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

So C and M are random variables.

16.1.1 Indicator Random Variables

An *indicator random variable* is a random variable that maps every outcome to either 0 or 1. ~~These~~ Indicator random variables are also called *Bernoulli variables*. 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, E, partitions the sample space into those outcomes in E and those not in E. So E is naturally associated with an indicator random variable, I_E, where $I_E(w) = 1$ for outcomes $w \in E$ and $I_E(w) = 0$ for outcomes $w \notin E$. Thus, $M = I_E$

where E is the event that all three coins match.

16.1.2 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 properties of random variables. For example,

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

~~EDITING NOTE:~~

As another example:

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

16.1.4 Conditional Probability

INSERT T goes here—

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

INSERT T

16.1.3 Functions of Random Variables

Random variables can be combined to form other random variables. For example, suppose that you roll two unbiased, independent 6-sided dice. Let D_i ~~d~~ be the random variable denoting the ~~ith~~ outcome of the ith die for $i = 1, 2$. ~~The~~ For example,

$$\Pr[D_1 = 3] = 1/6.$$

Then let $T = D_1 + D_2$. T ~~is then~~ is also a random variable and it denotes the sum of the two dice. For example,

$$\Pr[T = 2] = 1/36$$

and

$$\Pr[T = 7] = 1/6.$$

Random variables can be combined in complicated ways as we will see in Chapter 18. For example,

$$Y = e^T$$

is also a random variable. In this case,

$$\Pr[Y = e^2] = 1/36$$

and

$$\Pr[Y = e^7] = 1/6.$$

this probability using the definition of conditional probability:

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

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.

~~16.1.3~~ 16.1.5 Independence

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

$$\Pr\{R_1 = x_1 \text{ AND } R_2 = x_2\} = \Pr\{R_1 = x_1\} \cdot \Pr\{R_2 = x_2\}.$$

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

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

whenever the lefthand conditional probability is defined, that is, whenever $\Pr\{R_2 = x_2\} > 0$.

For ~~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:

$$\Pr\{C = x_1 \text{ AND } M = x_2\} \neq \Pr\{C = x_1\} \cdot \Pr\{M = x_2\}.$$

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

$$\Pr\{C = 2 \text{ AND } M = 1\} = 0 \cancel{\neq \frac{1}{4} \cdot \frac{3}{8} = \Pr\{M = 1\} \cdot \Pr\{C = 2\}},$$

and

$$\Pr\{M=1\} \cdot \Pr\{C=2\} = \frac{1}{4} \cdot \frac{3}{8} \neq 0.$$

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.

On the other hand, let H_1 be the indicator variable for the event that the first flip is a Head, so

$$[H_1 = 1] = \{HHH, HTH, HHT, HTT\}.$$

Then H_1 is independent of M, since

$$\Pr\{M = 1\} = 1/4 = \Pr\{M = 1 \mid H_1 = 1\} = \Pr\{M = 1 \mid H_1 = 0\}$$

$$\Pr\{M = 0\} = 3/4 = \Pr\{M = 0 \mid H_1 = 1\} = \Pr\{M = 0 \mid H_1 = 0\}$$

This example is an instance of a simple lemma:

Lemma 16.1.2. *Two events are independent iff their indicator variables are independent.*

As with events, the notion of independence generalizes to more than two random variables.

Definition 16.1.3. Random variables $R_1, R_2, \ldots, R_n$ are *mutually independent* iff

$$\begin{aligned} &\Pr\{R_1 = x_1 \text{ AND } R_2 = x_2 \text{ AND} \cdots \text{ AND } R_n = x_n\} \\ &= \Pr\{R_1 = x_1\} \cdot \Pr\{R_2 = x_2\} \cdots \Pr\{R_n = x_n\}. \end{aligned}$$

for all $x_1, x_2, \ldots, x_n$.

A consequence of ~~th~~ Definition 16.1.3 is

~~It is a simple exercise to show~~ that the probability that any *subset* of the variables takes a particular set of values is equal to the product of the probabilities 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:

$$\Pr\{R_1 = 7 \text{ AND } R_7 = 9.1 \text{ AND } R_{23} = \pi\} = \Pr\{R_1 = 7\} \cdot \Pr\{R_7 = 9.1\} \cdot \Pr\{R_{23} = \pi\}.$$

The proof is based on summing over all possible values for all of the other random variables.

~~We have~~

~~INSERT A goes here~~

16.1.6 Distribution Functions ← subsection

~~16.2 Probability Distributions~~

~~16.2.1 Probability Density Functions~~

A random variable maps outcomes to values, ~~but~~ ~~Som~~ Often, random variables that show up for different spaces of outcomes wind up behaving in much the same way because they have the same probability of ~~taking~~ having any given value. ~~Namely~~ Hence, random variables on different probability spaces may wind up having the same probability density function.

Definition 16.2.1. Let R be a random variable with codomain V. The *probability density function (pdf)* of R is a function $\text{PDF}_R : V \to [0, 1]$ defined by:

$$\text{PDF}_R(x) ::= \begin{cases} \Pr\{R = x\} & \text{if } x \in \text{range}(R), \\ 0 & \text{if } x \notin \text{range}(R). \end{cases}$$

A consequence of this definition is that

$$\sum_{x \in \text{range}(R)} \text{PDF}_R(x) = 1.$$

This ~~follows~~ is because R has a value for each outcome, so summing the probabilities over all outcomes is the same as summing over the probabilities of each value in the range of R.

As an example, ~~let's return to the experiment of rolling two fair~~ ~~consider~~ suppose that you roll two unbiased, independent 6-sided dice. ~~As before,~~ let T be the ~~total~~ random variable that equals the sum 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:~~ in Figure F2.

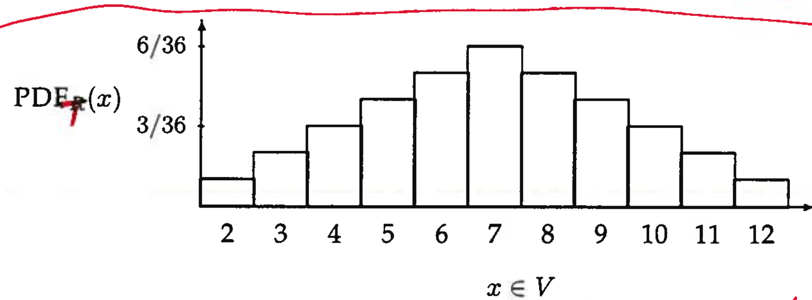

Figure F2: The ~~pdf for T~~ probability density function for the sum of two 6-sided dice.

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 $V = \{2, 3, \ldots, 12\}$.

concept to a pdf

A closely-related ~~idea~~ is the *cumulative distribution function (cdf)* for a random variable ~~R~~ whose codomain is real numbers. This is a function $\mathrm{CDF}_R : \mathbb{R} \to [0,1]$ defined by:

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

As an example, the cumulative distribution function for the random variable T is shown ~~below:~~ in Figure F3,

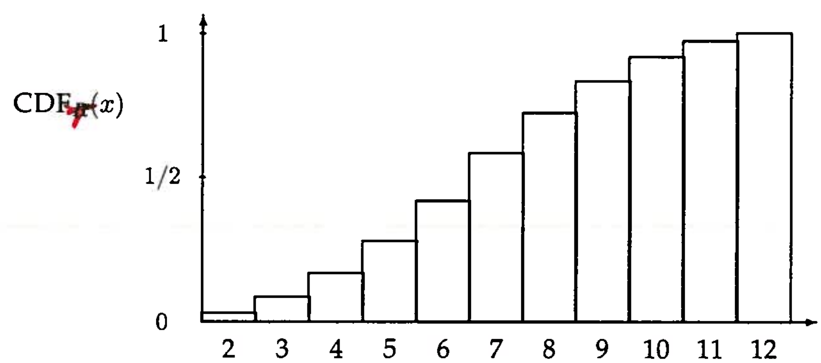

Figure F3: The ~~cdf~~ cumulative distribution function for the sum of ~~two~~ two 6-sided dice.

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 distribution, and ~~as~~ the binomial distribution. We we look more closely at these ~~distribu~~ common distributions in the next several sections.

the definitions of pdf and cdf:

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

$$= \sum_{y \leq x} \Pr\{R = y\}$$

$$= \sum_{y \leq x} \text{PDF}_R(y)$$

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.~~

~~**EDITING NOTE**: Thus, through these functions, we can study random variables without reference to a particular experiment.~~

One of the really interesting things about density functions and ~~a~~ distribution functions is that many random variables turn out to have the same ~~pdf~~ pdf and cdf. In other words, even though R and S ~~are~~ are different random variables on different probability spaces, it is often the case that ~~PDF~~

$$\text{PDF}_R = \text{PDF}_S.$$

In fact, some pdf's are so common that they are given a special name. ~~such as~~ ~~The three most important distributions in computer sci~~ For example, the three most important distributions in computer science are the Bernoulli distribution, the uniform

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

16.2 ~~16.2.1~~ Bernoulli Distributions ← section

The Bernoulli distribution is the simplest and most common distribution function. That's because it is the distribution function for an indicator random variable.

~~Indicator random variables are perhaps the most common type because of their close as-~~ ~~In particular, the~~ ~~sociation with events. The probability density function of an indicator random variable~~, ~~with parameter p is (...) is~~ ~~B, is always~~

Specifically, the Bernoulli distribution ~~In fact~~ ~~more specific~~ ~~is specified by~~ has a probability density function $f_p : \{0, 1\} \to [0, 1]$ of the form where

$$\cancel{\mathrm{PDF}_B(0) = p} \quad f_p(0) = p \text{ and}$$

$$\cancel{\mathrm{PDF}_B(1) = 1 - p} \quad f_p(1) = 1 - p,$$

for some $p \in [0, 1]$. ~~where $0 \le p \le 1$~~. The corresponding cumulative distribution function is:

$$\cancel{\mathrm{CDF}_B(0) = p} \quad F_p(0) = p$$

$$\cancel{\mathrm{CDF}_B(1) = 1} \quad F_p(1) = 1.$$

16.3 ← section

~~16.2.2~~ Uniform Distributions

16.3.1 Definition

A random variable that takes on each possible value with the same probability is ~~called~~ said to be *uniform*. ~~For example, the probability density function of a random variable U that is uniform on the set $\{1,2,\ldots,N\}$ is:~~ The uniform distribution has a pdf of the form $f_n : \{1, 2, \ldots, n\} \to [0,1]$ where

$$\cancel{\mathrm{PDF}_U(k) = \frac{1}{N}} \quad f_n(k) = \frac{1}{n} \quad \&$$

for some $n \in \mathbb{N}^+$. The ~~for $k = 1, 2, \ldots, N$~~

~~And the~~ cumulative distribution function is:

$$\cancel{\mathrm{CDF}_U(k) = \frac{k}{N}} \quad F_n(k) = \frac{k}{n}.$$

Uniform distributions ~~come up all the time~~ arise frequently in practice. For example, the number rolled on a fair die is uniform on the set $\{1, 2, \ldots, 6\}$.

~~16.2.3~~ 16.3.2 The Numbers Game

Enough definitions — let's ~~Let's~~ play a game! ~~I~~ we 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~~ we'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~~ one envelope and see the number 12. Since 12 is a small number, you might guess that that ~~other number~~ the number in the other envelope is larger. But perhaps ~~I'm sort of~~ we're been 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~~ we're picking the numbers and ~~I'm~~ we're choosing them in a way that ~~I~~ we think will defeat your guessing strategy. ~~I'll~~ we'll only use randomization to choose the numbers if that serves ~~my~~ our purposes, ~~end: making~~ which is to make you lose!

Intuition Behind the Winning Strategy

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

Suppose that you somehow knew a number x ~~*between* my lower number and higher numbers~~ that was in between the numbers in the envelopes. Now you peek in ~~an~~ one envelope and see ~~one or the other~~ a number. 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 a number x between ~~my~~ the numbers in the envelopes,

~~lower and higher numbers~~, then you are certain to win the game.

such an

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!

Analysis of the Winning Strategy

we

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, as shown in Figure F4.

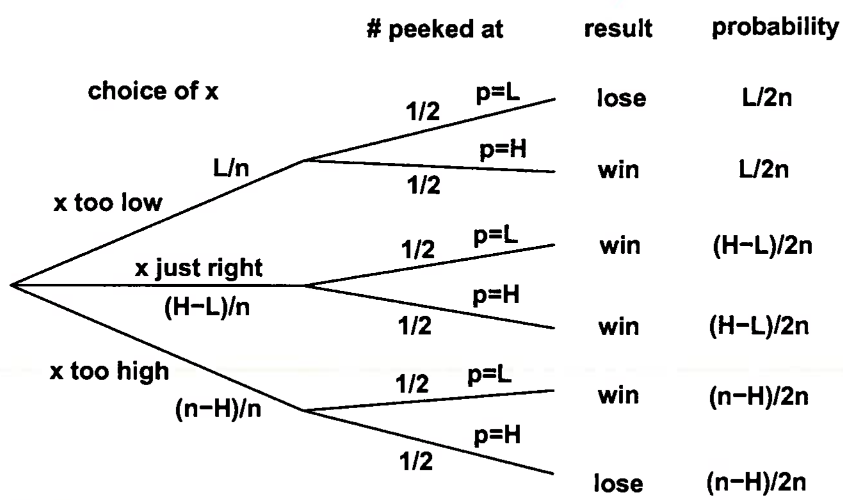

Figure F4: The tree diagram for the numbers game.

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!

~~Now, it's back to m~~ — INSERT D goes here —

16.4 ~~16.2.4~~ Binomial Distributions ⟸ section

16.4.1 Definitions

The third commonly-used distribution in computer science is the ~~The~~ *binomial distribution* ~~plays an important role in Computer Science as it does in most other sciences~~. 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_n. If the coin is fair, then H_n has~~. If the coin is fair, then the number of heads has an unbiased Binomial distribution, specified by the pdf $f_n : \{1, 2, \ldots, k\} \to [0,1]$ where ~~an *unbiased binomial density function*. In particular~~

$$f_n(k) \; \cancel{\mathrm{PDF}_{H_n}(k)} = \binom{n}{k} 2^{-n}.$$

~~for $k = 1, 2, \ldots, n$.~~ for some $n \in \mathbb{N}^+$. This is ~~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}.

The cumulative distribution ~~function for H_n~~

~~is~~

$$\cancel{F_n(k) = \sum_{i=0}^{k} \binom{n}{k} 2^{-n}}.$$

INSERT D

It turns out that this strategy is optimal. We won't prove it here but if I also use randomness in the right way to select the numbers that go in the envelopes, then the strategy that we described for the player is optimal. In particular,

Claim: **If ~~[illegible]~~ y is uniformly chosen from [0, n-1] and z = y +1, then for any player strategy,**

$$\mathbf{Pr(win)} \leq \frac{1}{2} + \frac{1}{2n} .$$

Randomized Algorithms

The ~~best~~ strategy to win the numbers game is an example of a randomized algorithm — it uses randomly numbers to influence decisions.

Questions?

~~This is our first example of a randomized protocol.~~ Protocols and algorithms that make use of random numbers are very important in computer science. There are many problems for which the best known solutions are based on a random number generator.

For example, the most commonly used protocol for deciding when to send a broadcast on a shared bus or Ethernet is a randomized algorithm known as exponential backoff. ~~Explain~~ ← ~~purple~~ One of the most commonly used sorting algorithms used in practice (called quicksort) uses random numbers. You'll see many more examples ~~in 6.046~~ if ~~when~~ you take an algorithms course. In each case, randomness is used to improve the ~~performance of the algorithm or protocol~~ ~~chance~~ probability that the algorithm runs quickly or otherwise performs well.

$f_n(k)$

~~Here is a plot of the unbiased probability density function PDF$_{B_n}(k)$ corresponding~~

A plot of $f_{20}(k)$ is shown in Figure F5.

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*. We'll talk a lot more about these tails in ~~a moment~~ Section 16.4.

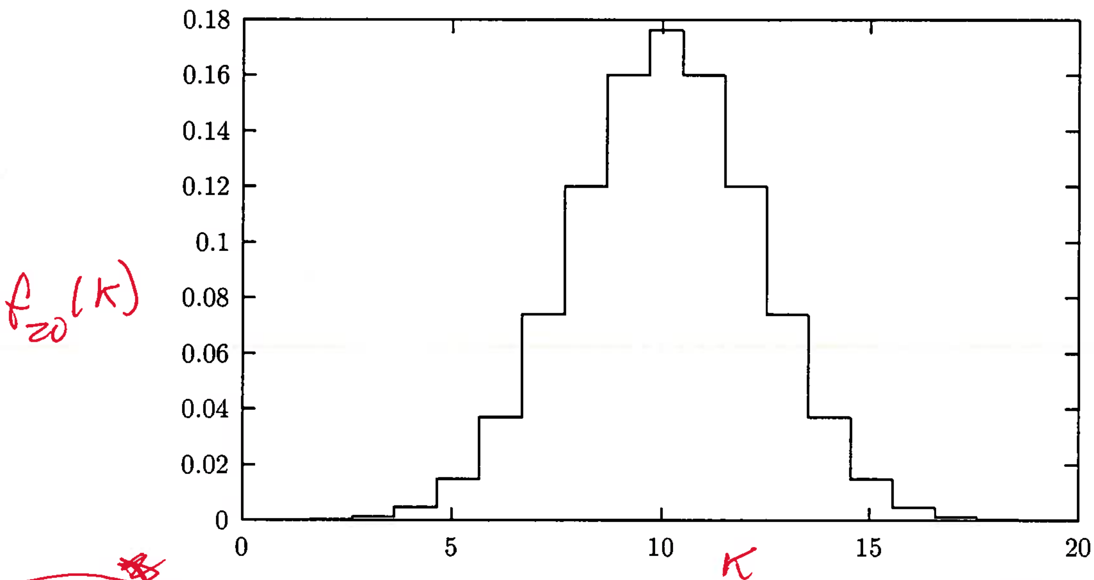

Figure F5: The pdf for the unbiased binomial distribution for $n=20$, $f_{20}(k)$.

This is insert H and goes to p 1052

In many fields, including Computer Science, probability analyses come down to getting small bounds on the tails of the binomial distribution. In the context of a problem,

The cumulative distribution function for the unbiased binomial distribution is

$$F_n(k) = \sum_{i=0}^{k} \binom{n}{i} 2^{-n}$$

for $k \in \{1, 2, \ldots, n\}$.

unlikely that too many bits are corrupted in a message, or that too many servers or communication links become overloaded, or that a randomized ~~algorit~~ algorithm runs for too long.

this typically means that there is very small probability that ~~something *bad* happens,~~ too many *bad* things happen. For example, we would like to know that is very ~~which could be a server or communication link overloading or a randomized algorithm~~ ~~unlikely that too~~ ~~running for an exceptionally long time or producing the wrong result.~~

As an example, ~~we can calculate~~ lets compute the probability of flipping ~~at most 25 heads~~ 75 or more tails in 100 tosses of a fair coin ~~and see that it is very small, namely, less than 1 in 3,000,000.~~ This is the same as the probability of flipping at most 25 heads. ~~which is~~ ~~Plugging $n = 100$, $p = 1/2$, and $\alpha = 1/4$ into Equation 16.1, we find that this probability is~~

~~**EDITING NOTE**: Add calculation that the ratio of the $k-1$st and kth terms for $k \leq 25$ is less than 1/4(?), so the probability of $< k$ heads is less than 1/2 the prob of exactly k heads.~~

~~In fact, the tail of the distribution falls off so rapidly that the probability of flipping exactly 25 heads is nearly twice the probability of flipping fewer than 25 heads! That is, the probability of flipping exactly 25 heads —small as it is —is still nearly twice as large as the probability of flipping exactly 24 heads *plus* the probability of flipping exactly 23~~

End of Insert H

heads *plus* ... ~~the~~ probability of flipping ~~no heads~~.

The General Binomial Distribution

If the coins are biased so that each coin

~~Now~~ ~~let J be the number of heads that come up on n independent coins, each of which~~

the number of heads has a general binomial distribution

is heads with probability p. Then ~~J has a general binomial density function~~:

specified by the pdf $f_{n,p} : \{1, 2, \ldots, n\} \to [0, 1]$ where

$$f_{n,p}(k) \; \cancel{\mathrm{PDF}_J(k)} = \binom{n}{k} p^k (1-p)^{n-k}.$$

~~for $k = 1, 2, \ldots, n$~~

~~This is because~~ for some $n \in \mathbb{N}^+$ and $p \in [0, 1]$. This is because

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

For ~~As an~~ example, the plot ~~below~~ in Figure F7 shows the probability density function ~~$\mathrm{PDF}_J(k)$~~ $f_{n,p}(k)$ 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.

The cumulative distribution function for the general binomial distribution is

$$F_{n,p}(k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \quad \text{(eqn F5)}$$

for ~~$k \in \{1, 2, \dots, n\}$~~ $k \in \{1, 2, \dots, n\}$.

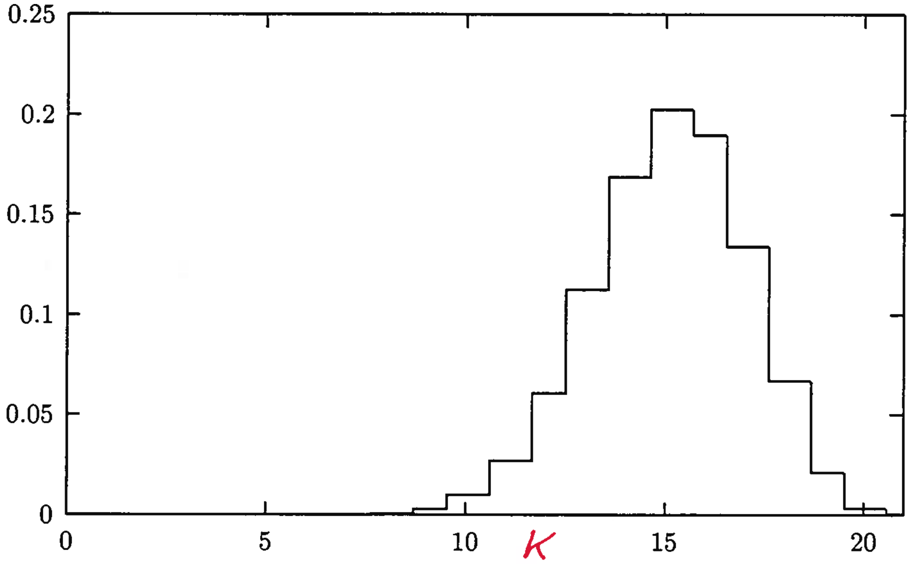

Figure F7: The pdf for the general binomial distribution $f_{n,p}(k)$ for ~~n = 20 and~~ $n = 20$ and $p = .75$.

~~EDITING NOTE:~~

~~Approximating the Binomial Density Function~~

~~16.4.2 Approximations ing the~~

16.4.2 Approximating the Binomial Density Function

Computing the general binomial density function is daunting if not impossible when k and n are large. ~~n is up in the thousands~~s. Fortunately, there is an approximate closed-form formula for this function based on an approximation for the binomial coefficient. In the formula, k

~~using theorem 9.3.1 6.1~~

is replaced by αn where α is a number between 0 and 1.

Lemma 16.2.2.

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

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

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

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

and $\frac{1}{12m+1} \le \epsilon(m) \le \frac{1}{12m}$ for all $m \in \mathbb{N}^+$.

The graph of H is shown in Figure 16.1.

Lemma 16.2.2 provides ~~The upper bound(16.2.2) on the binomial coefficient is tight enough to serve as~~ an excellent approximation for binomial coefficients. We'll skip its derivation, which consists of plugging in Theorem 9.6.1 ~~Stirling's formula~~ for the factorials in the binomial coefficient and then simplifying.

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

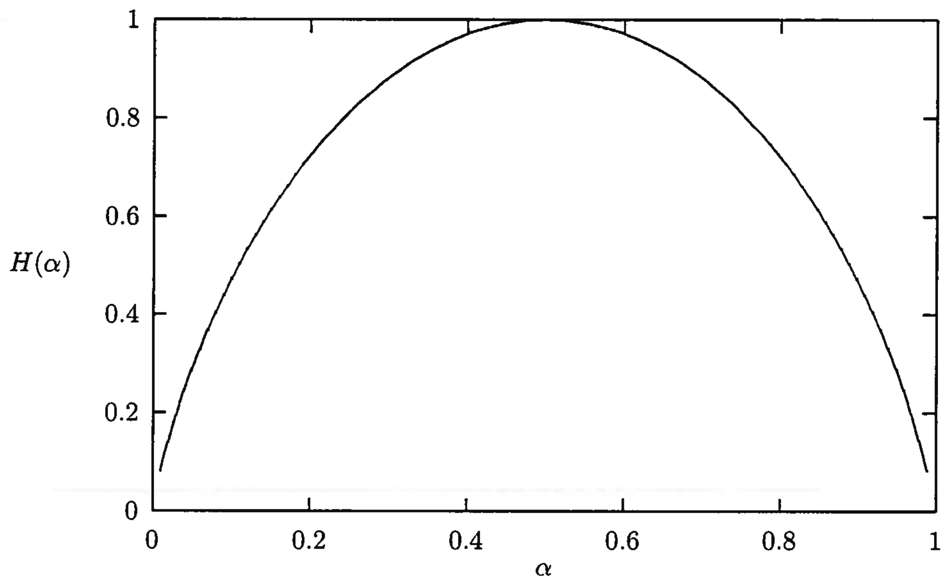

Figure 16.1: The Entropy Function

is: INSERT G goes here

$$\cancel{\mathrm{PDF}_J(\alpha n) \leq \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot p^{\alpha n}(1-p)^{(1-\alpha)n}} \qquad \cancel{(16.1)}$$

This formula is as ugly as a bowling shoe, but is useful because it's easy to evaluate. For example, suppose we flip a fair coin n times. What is the probability of getting *exactly* pn ~~$\frac{1}{2}n$~~ heads? Plugging $\alpha =$ p ~~$1/2$ and $p = 1/2$ and $H(1/2) = 1$~~ into Equation 16.1 ~~(16.1)~~ gives:

$$\cancel{\mathrm{PDF}_J(\alpha n) \leq \frac{2^{nH(1/2)}}{\sqrt{2\pi(1/2)(1-(1/2))n}} \cdot 2^{-n}}$$

$$\cancel{= \sqrt{\frac{2}{\pi n}}}$$

Insert R goes here

Thus, for example, if we flip a fair coin (where $p = 1/2$) $n =$ 100 times, the probability of getting exactly 50 heads is ~~about $1/\sqrt{50\pi} \approx 0.079$ or around 8%.~~ 0.079... or around 8%.[1]

[1] The ~~errors~~ contribution of the $e^{\epsilon(n) - 2\epsilon(n/2)}$ ~~$e^{\epsilon_n - 2\epsilon_{n/2}}$~~ term is so small for $n = 100$ that it disappears in the ... .

INSERT G

$$f_{n,p}(\alpha n) = \frac{2^{nH(\alpha)}\, p^{\alpha n} (1-p)^{(1-\alpha)n}\, e^{\substack{\varepsilon(n)-\varepsilon((1-\alpha)n)-\varepsilon(\alpha n) \\ \cancel{a_n - a_{(1-\alpha)n} - a_{\alpha n}}}}}{\sqrt{2\pi\alpha(1-\alpha)n}}$$

Eqn 6.1 $\longrightarrow$

$$= \frac{2^{n\left(\alpha \log\frac{p}{\alpha} + (1-\alpha)\log\frac{1-p}{1-\alpha}\right)}\, e^{\substack{\varepsilon(n)-\varepsilon((1-\alpha)n)- \\ \cancel{a_n - a_{(1-\alpha)n} - a}}}}{\sqrt{2\pi\alpha(1-\alpha)n}}$$

$$\sim \frac{2^{n\left(\alpha \log\frac{p}{\alpha} + (1-\alpha)\log\frac{1-p}{1-\alpha}\right)}}{\sqrt{2\pi\alpha(1-\alpha)n}}.$$

INSERT R

$$f_{n,p}(pn) = \frac{\cancel{\frac{n(0)}{2}}\; e^{\varepsilon(n) - 2\varepsilon(n/2)}\;\cancel{{}^{a_n - 2a_{n/2}}}}{\sqrt{2\pi\, p(1-p)\, n}}$$

$$\cancel{\sim \frac{2}{\sqrt{\pi n}}}$$

$$\cancel{\sim \sqrt{\frac{2}{\pi n}}}$$

$$\cancel{\sim \frac{1}{\sqrt{p(1-p)\, 2\pi n}}}$$

$$\sim \frac{1}{\sqrt{2\pi\, p(1-p)\, n}}$$

16.4.3 Bounding the Tails

— INSERT H goes here —
(it is text on pp 1045-6)

~~— INSERT I goes here —~~

~~**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* αn heads is given by the cumulative binomial distribution function: ~~Equation 15.~~

$$\cancel{\mathrm{CDF}_J(\alpha n) = \Pr\{J \leq \alpha n\} = \sum_{i=0}^{\alpha n} \mathrm{PDF}_J(i)} \tag{16.2}$$

$$F_{n,p}(\alpha n) = \sum_{i=0}^{\alpha n} \binom{n}{i} p^i (1-p)^{n-i}$$

We can bound this sum by bounding the ratio of successive terms. INSERT I goes here ~~This yields a geometric sum from 0 to $\mathrm{PDF}_J(\alpha n)$ that bounds (16.2). Then applying the formula for a geometric sum gives~~

$$\cancel{\mathrm{CDF}_J(\alpha n) \leq \frac{1-\alpha}{1-\alpha/p} \cdot \mathrm{PDF}_J(\alpha n),} \tag{16.3}$$

~~which holds providing $\alpha < p$. This is all we need, since we already have the bound (16.1) for $\mathrm{PDF}_J(\alpha n)$.~~

~~It would be awkward to evaluate (16.3) with a calculator, but it's easy to write a pro-~~

INSERT I

In particular, for $i \le \alpha n$,

$$\frac{\binom{n}{i-1} p^{i-1} (1-p)^{n-(i-1)}}{\binom{n}{i} p^{i} (1-p)^{n-i}} = \frac{\dfrac{n!\; p^{i-1} (1-p)^{n-i+1}}{(i-1)!\,(n-i+1)!}}{\dfrac{n!\; p^{i} (1-p)^{n-i}}{i!\,(n-i)!}}$$

$$= \frac{i\,(1-p)}{(n-i+1)\,p}$$

$$\le \frac{\alpha n\,(1-p)}{(n-\alpha n+1)\,p}$$

$$\le \frac{\alpha(1-p)}{(1-\alpha)\,p}.$$

This means that for $\alpha < p$,

$$F_{n,p}(\alpha n) \le f_{n,p}(\alpha n) \sum_{i=0}^{\infty} \left[\frac{\alpha(1-p)}{(1-\alpha)p}\right]^i$$

$$= \frac{f_{n,p}(\alpha n)}{1 - \dfrac{\alpha(1-p)}{(1-\alpha)p}}$$

$$= \left(\frac{1-\alpha}{1-\alpha/p}\right) f_{n,p}(\alpha n). \qquad \text{(Eqn F7)}$$

In other words, the probability of at most αn heads is at most

$$\frac{1-\alpha}{1-\alpha/p}$$

times the probability of exactly αn heads.

For ~~the~~ our scenario, where $p = 1/2$ and $\alpha = 1/4$,

$$\frac{1-\alpha}{1-\alpha/p} = \frac{3/4}{1/2} = 3/2.$$

Plugging $n = 100$, $\alpha = 1/4$, and $p = 1/2$ into Equation 16.1, we find that the probability of at most 25 heads in 100 coin flips is

$$F_{100,1/2}(25) < \frac{3}{2} \cdot \frac{2^{100\left(\frac{1}{4}\log 2 + \frac{3}{4}\log\frac{2}{3}\right)}}{\sqrt{75\pi/2}}\, e^{\epsilon(100)-\epsilon(75)-\epsilon(25)}$$

$$\leq 2.9 \cdot 10^{-7}.$$

~~gram to do it. 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 $n = 100$:~~

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

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 ~~$\mathrm{CDF}_J(\alpha n)$~~ $F_{n,p}(\alpha n)$ in Equation F7 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.~~

— INSERT L goes here —

16.3 Average & Expected Value

This goes to CH 17

The *expectation* of a random variable is its average value, where each value is weighted according to the probability that it comes up. The expectation is also called the *expected value* or the *mean* of the random variable.

For example, suppose we select a student uniformly at random from the class, and let R be the student's quiz score. Then $\mathrm{E}[R]$ is just the class average —the first thing everyone wants to know after getting their test back! For similar reasons, the first thing you usually want to know about a random variable is its expected value.

INSERT L

16.5 Continuous Distributions

You may have noticed that all of the distributions we have discussed thus far are for finite sample spaces. That's because ~~in computer sci~~ finite distributions are the most common in computer science. They are also the easiest to work with.

~~In~~ In other ~~fields,~~ ~~More generally,~~ there are ~~distributio~~ important distributions on infinite sample spaces. ~~A~~ ~~The~~ A good example is the normal distribution. The standard normal distribution is defined by the pdf

$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

~~f~~ for $x \in \mathbb{R}$. A graph of $f(x)$ is shown in Figure L1. ~~The cumulative distribution function is given by the for~~

$$F(x) = \int_{-\infty}^{x} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz.$$

More generally, there are important distributions on infinite sample spaces. We will breifly mention some of the most important in the following subsections. For the most part in this text, however, our focus will continue to be on finite probability spaces

— INSERT M goes here —

16.5.2 ~~The~~ Normal Distributions

The standard normal distribution is defined by the pdf $f : \mathbb{R} \to [0,1]$ where

$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

A graph of $f(x)$ is shown in Figure L1.

16.5.1 Continuous Uniform Distributions

~~The un~~ We have already talked about the uniform distribution on a finite sample space $\{1, 2, \ldots, n\}$. The uniform distribution can also be defined on the infinite sample space $[0, n]$. In this case, the pdf is $f_n : [0, n] \Rightarrow [0, 1]$ where

$$f_n(x) = \frac{1}{n}$$

and the ~~cumulative d~~ cdf is

$$F_n(x) = \frac{x}{n} .$$

~~For~~ The difference between the continuous and discrete uniform distributions is that ~~for~~ the pdf for the continuous ~~can~~ uniform distribution is nonzero for any real $x \in [0, n]$ whereas the pdf ~~for~~ the discrete uniform distribution is nonzero only for integer $x \in [1, n]$.

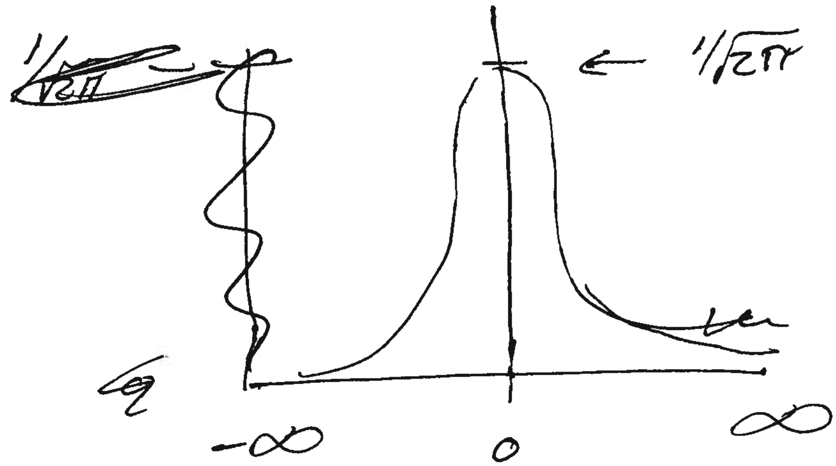

Figure L1: The plot of the pdf for the standard normal distribution.

The cumulative distribution function ~~is given~~ for the standard normal is

$$F(x) = \int_{-\infty}^{x} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz .$$

~~we will Henceforth,~~ ~~For the most parts~~ ~~we will restrict our attention to discrete distributions in this~~ going ~~text~~ ~~forward~~

The general normal distribution is defined based on two parameters: μ (the mean) and σ^2 (the variance). ~~and~~ It's pdf is the function $f_{\mu,\sigma^2}: \mathbb{R} \to [0,1]$ where

$$f_{\mu,\sigma^2}(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} . \qquad \text{(eqn N1)}$$

~~In some~~

The normal distribution is similar to the binomial distribution in some respects. ~~In fact, the binomial distribution is sometimes thought of referred to as the discrete version of the normal distribution.~~ ~~For example, if you set you cons are looking at the problem of flipping flipping αn heads in n unbiased, independent coin tosses~~ For example, if we set $\mu = \frac{n}{2}$, $\sigma^2 = \frac{n}{4}$ and $x = \alpha n$ in Equation N1, the resulting pdf is exponentially small in n, which is similar to the behavior of Equation 16.1 when $p = 1/2$. We'll talk further about the relationship in chapter 18, where the reasons for the choices of μ, σ^2 and x above will become apparent.