18 Deviation from the Mean

18.1 Why the Mean?

In the previous chapter we took it for granted that expectation is important, and we developed a bunch of techniques for calculating expected values. But why should we care about this value? After all, a random variable may never take a value anywhere near its expected value.

The most important reason to care about the mean value comes from its connection to estimation by sampling. For example, suppose we want to estimate the average age, income, family size, or other measure of a population. To do this, we determine a random process for selecting people —say throwing darts at census lists. This process makes the selected person's age, income, and so on into a random variable whose *mean* equals the *actual average* age or income of the population. So we can select a random sample of people and calculate the average of people in the sample to estimate the true average in the whole population. Many fundamental results of probability theory explain exactly how the reliability of such estimates improves as the sample size increases, and in this chapter we'll examine a few such results.

In particular, when we make an estimate by repeated sampling, we need to know how much confidence we should have that our estimate is OK. Technically, this reduces to finding the probability that an estimate *deviates* a lot from its expected value. This topic of *deviation from the mean* is the focus of this final chapter.

18.2 Markov's Theorem

Markov's theorem is an easy result that gives a generally rough estimate of the probability that a random variable takes a value *much larger* than its mean.

The idea behind Markov's Theorem can be explained with a simple example of *intelligence quotient*, IQ. This quantity was devised so that the average IQ measurement would be 100. Now from this fact alone we can conclude that at most 1/3 the population can have an IQ of 300 or more, because if more than a third had an IQ of 300, then the average would have to be *more* than $(1/3)300 = 100$, contradicting the fact that the average is 100. So the probability that a randomly chosen person has an IQ of 300 or more is at most 1/3. Of course this is not a very

strong conclusion; in fact no IQ of over 300 has ever been recorded. But by the same logic, we can also conclude that at most 2/3 of the population can have an IQ of 150 or more. IQ's of over 150 have certainly been recorded, though again, a much smaller fraction than 2/3 of the population actually has an IQ that high.

But although these conclusions about IQ are weak, they are actually the strongest general conclusions that can be reached about a random variable using *only* the fact that it is nonnegative and its mean is 100. For example, if we choose a random variable equal to 300 with probability 1/3, and 0 with probability 2/3, then its mean is 100, and the probability of a value of 300 or more really is 1/3. So we can't hope to get a better upper bound based solely on this limited amount of information.

Theorem 18.2.1 (Markov's Theorem). *If R is a nonnegative random variable, then for all $x > 0$*

$$\Pr[R \geq x] \leq \frac{\mathrm{Ex}[R]}{x}.$$

Proof. For any $x > 0$

$$\begin{aligned}
\mathrm{Ex}[R] ::= & \sum_{y \in \mathrm{range}(R)} y \Pr[R = y] \\
\geq & \sum_{\substack{y \geq x, \\ y \in \mathrm{range}(R)}} y \Pr[R = y] \qquad \text{(because } R \geq 0\text{)} \\
\geq & \sum_{\substack{y \geq x, \\ y \in \mathrm{range}(R)}} x \Pr[R = y] \\
= & \; x \sum_{\substack{y \geq x, \\ y \in \mathrm{range}(R)}} \Pr[R = y] \\
= & \; x \Pr[R \geq x]. \qquad (18.1)
\end{aligned}$$

Dividing the first and last expression (18.1) by x gives the desired result. ∎

Our focus is deviation from the mean, so it's useful to rephrase Markov's Theorem this way:

Corollary 18.2.2. *If R is a nonnegative random variable, then for all $c \geq 1$*

$$\Pr[R \geq c \cdot \mathrm{Ex}[R]\,] \leq \frac{1}{c}. \qquad (18.2)$$

This Corollary follows immediately from Markov's Theorem(18.2.1) by letting x be $c \cdot \mathrm{Ex}[R]$.

18.2.1 Applying Markov's Theorem

Let's consider the Hat-Check problem again. Now we ask what the probability is that x or more men get the right hat, this is, what the value of $\Pr[G \geq x]$ is.

We can compute an upper bound with Markov's Theorem. Since we know $\mathrm{Ex}[G] = 1$, Markov's Theorem implies

$$\Pr[G \geq x] \leq \frac{\mathrm{Ex}[G]}{x} = \frac{1}{x}.$$

For example, there is no better than a 20% chance that 5 men get the right hat, regardless of the number of people at the dinner party.

The Chinese Appetizer problem is similar to the Hat-Check problem. In this case, n people are eating appetizers arranged on a circular, rotating Chinese banquet tray. Someone then spins the tray so that each person receives a random appetizer. What is the probability that everyone gets the same appetizer as before?

There are n equally likely orientations for the tray after it stops spinning. Everyone gets the right appetizer in just one of these n orientations. Therefore, the correct answer is $1/n$.

But what probability do we get from Markov's Theorem? Let the random variable, R, be the number of people that get the right appetizer. Then of course $\mathrm{Ex}[R] = 1$ (right?), so applying Markov's Theorem, we find:

$$\Pr[R \geq n] \leq \frac{\mathrm{Ex}[R]}{n} = \frac{1}{n}.$$

So for the Chinese appetizer problem, Markov's Theorem is tight!

On the other hand, Markov's Theorem gives the same $1/n$ bound for the probability everyone gets their hat in the Hat-Check problem in the case that all permutations are equally likely. But the probability of this event is $1/(n!)$. So for this case, Markov's Theorem gives a probability bound that is way off.

18.2.2 Markov's Theorem for Bounded Variables

Suppose we learn that the average IQ among MIT students is 150 (which is not true, by the way). What can we say about the probability that an MIT student has an IQ of more than 200? Markov's theorem immediately tells us that no more than 150/200 or 3/4 of the students can have such a high IQ. Here we simply applied Markov's Theorem to the random variable, R, equal to the IQ of a random MIT student to conclude:

$$\Pr[R > 200] \leq \frac{\mathrm{Ex}[R]}{200} = \frac{150}{200} = \frac{3}{4}.$$

"where get this?
Oh 150 = E[R]

But let's observe an additional fact (which may be true): no MIT student has an IQ less than 100. This means that if we let $T ::= R - 100$, then T is nonnegative and $\text{Ex}[T] = 50$, so we can apply Markov's Theorem to T and conclude:

$$\Pr[R > 200] = \Pr[T > 100] \leq \frac{\text{Ex}[T]}{100} = \frac{50}{100} = \frac{1}{2}.$$

So only half, not 3/4, of the students can be as amazing as they think they are. A bit of a relief!

In fact, we can get better bounds applying Markov's Theorem to $R - b$ instead of R for any lower bound $b > 0$ on R (see Problem 18.2). Similarly, if we have any upper bound, u, on a random variable, S, then $u - S$ will be a nonnegative random variable, and applying Markov's Theorem to $u - S$ will allow us to bound the probability that S is much *less* than its expectation.

18.3 Chebyshev's Theorem

We got more mileage out of Markov's Theorem by applying it to $R - b$ rather than R. More generally, a really good trick for getting stronger bounds on a random variable R out of Markov's Theorem is to apply some cleverly chosen function of R.

Choosing functions that are powers of $|R|$ turns out to be specially useful. In particular, since $|R|^\alpha$ is nonnegative, Markov's inequality also applies to the event $[|R|^\alpha \geq x^\alpha]$. But this event is equivalent to the event $[|R| \geq x]$, so we have:

Lemma 18.3.1. *For any random variable R, $\alpha \in \mathbb{R}^+$, and $x > 0$,*

$$\Pr[|R| \geq x] \leq \frac{\text{Ex}[|R|^\alpha]}{x^\alpha}.$$

Rephrasing (18.3.1) in terms of the random variable, $|R - \text{Ex}[R]|$, that measures R's deviation from its mean, we get

$$\Pr[|R - \text{Ex}[R]| \geq x] \leq \frac{\text{Ex}[(R - \text{Ex}[R])^\alpha]}{x^\alpha}. \tag{18.3}$$

The case when $\alpha = 2$ is turns out to be so important that numerator of the right hand side of (18.3) has been given a name:

Definition 18.3.2. The *variance*, $\text{Var}[R]$, of a random variable, R, is:

$$\text{Var}[R] ::= \text{Ex}[(R - \text{Ex}[R])^2].$$

For all the places I saw this, I should be an expert!

The restatement of (18.3) for $\alpha = 2$ is known as *Chebyshev's Theorem.*

Theorem 18.3.3 (Chebyshev). *Let R be a random variable and $x \in \mathbb{R}^+$. Then*

$$\Pr[|R - \mathrm{Ex}[R]| \geq x] \leq \frac{\mathrm{Var}[R]}{x^2}.$$

The expression $\mathrm{Ex}[(R - \mathrm{Ex}[R])^2]$ for variance is a bit cryptic; the best approach is to work through it from the inside out. The innermost expression, $R - \mathrm{Ex}[R]$, is precisely the deviation of R above its mean. Squaring this, we obtain, $(R - \mathrm{Ex}[R])^2$. This is a random variable that is near 0 when R is close to the mean and is a large positive number when R deviates far above or below the mean. So if R is always close to the mean, then the variance will be small. If R is often far from the mean, then the variance will be large.

18.3.1 Variance in Two Gambling Games

The relevance of variance is apparent when we compare the following two gambling games.

Game A: We win \$2 with probability 2/3 and lose \$1 with probability 1/3.

Game B: We win \$1002 with probability 2/3 and lose \$2001 with probability 1/3.

Which game is better financially? We have the same probability, 2/3, of winning each game, but that does not tell the whole story. What about the expected return for each game? Let random variables A and B be the payoffs for the two games. For example, A is 2 with probability 2/3 and -1 with probability 1/3. We can compute the expected payoff for each game as follows:

$$\mathrm{Ex}[A] = 2 \cdot \frac{2}{3} + (-1) \cdot \frac{1}{3} = 1,$$
$$\mathrm{Ex}[B] = 1002 \cdot \frac{2}{3} + (-2001) \cdot \frac{1}{3} = 1$$

both same

The expected payoff is the same for both games, but they are obviously very different! This difference is not apparent in their expected value, but is captured by variance. We can compute the $\mathrm{Var}[A]$ by working "from the inside out" as follows:

$$A - \mathrm{Ex}[A] = \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ -2 & \text{with probability } \frac{1}{3} \end{cases}$$

$$(A - \mathrm{Ex}[A])^2 = \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ 4 & \text{with probability } \frac{1}{3} \end{cases}$$

$$\mathrm{Ex}[(A - \mathrm{Ex}[A])^2] = 1 \cdot \frac{2}{3} + 4 \cdot \frac{1}{3}$$

$$\mathrm{Var}[A] = 2.$$

but other far more risky

Similarly, we have for Var[B]:

$$\begin{aligned} B - \mathrm{Ex}[B] &= \begin{cases} 1001 & \text{with probability } \frac{2}{3} \\ -2002 & \text{with probability } \frac{1}{3} \end{cases} \\ (B - \mathrm{Ex}[B])^2 &= \begin{cases} 1,002,001 & \text{with probability } \frac{2}{3} \\ 4,008,004 & \text{with probability } \frac{1}{3} \end{cases} \\ \mathrm{Ex}[(B - \mathrm{Ex}[B])^2] &= 1,002,001 \cdot \frac{2}{3} + 4,008,004 \cdot \frac{1}{3} \\ \mathrm{Var}[B] &= 2,004,002. \end{aligned}$$

The variance of Game A is 2 and the variance of Game B is more than two million! Intuitively, this means that the payoff in Game A is usually close to the expected value of \$1, but the payoff in Game B can deviate very far from this expected value.

High variance is often associated with high risk. For example, in ten rounds of Game A, we expect to make \$10, but could conceivably lose \$10 instead. On the other hand, in ten rounds of game B, we also expect to make \$10, but could actually lose more than \$20,000!

18.3.2 Standard Deviation

Because of its definition in terms of the square of a random variable, the variance of a random variable may be very far from a typical deviation from the mean. For example, in Game B above, the deviation from the mean is 1001 in one outcome and -2002 in the other. But the variance is a whopping 2,004,002. From a dimensional analysis viewpoint, the "units" of variance are wrong: if the random variable is in dollars, then the expectation is also in dollars, but the variance is in square dollars. For this reason, people often describe random variables using standard deviation instead of variance.

Definition 18.3.4. The *standard deviation*, σ_R, of a random variable, R, is the square root of the variance:

$$\sigma_R ::= \sqrt{\mathrm{Var}[R]} = \sqrt{\mathrm{Ex}[(R - \mathrm{Ex}[R])^2]}.$$

So the standard deviation is the square root of the mean of the square of the deviation, or the *root mean square* for short. It has the same units —dollars in our example —as the original random variable and as the mean. Intuitively, it measures the average deviation from the mean, since we can think of the square root on the outside as canceling the square on the inside.

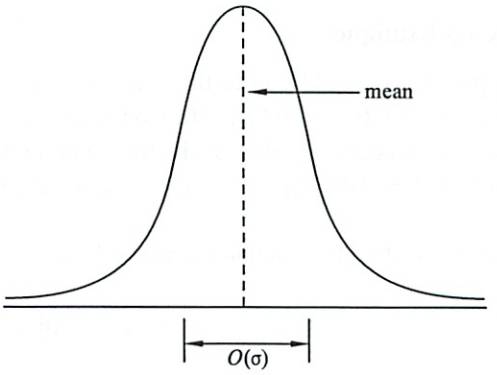


Figure 18.1 The standard deviation of a distribution indicates how wide the "main part" of it is.

Example 18.3.5. The standard deviation of the payoff in Game B is:

$$\sigma_B = \sqrt{\text{Var}[B]} = \sqrt{2,004,002} \approx 1416.$$

The random variable B actually deviates from the mean by either positive 1001 or negative 2002; therefore, the standard deviation of 1416 describes this situation reasonably well.

Intuitively, the standard deviation measures the "width" of the "main part" of the distribution graph, as illustrated in Figure 18.1.

It's useful to rephrase Chebyshev's Theorem in terms of standard deviation.

Corollary 18.3.6. *Let R be a random variable, and let c be a positive real number.*

$$\Pr[|R - \text{Ex}[R]| \geq c\sigma_R] \leq \frac{1}{c^2}.$$

Here we see explicitly how the "likely" values of R are clustered in an $O(\sigma_R)$-sized region around $\text{Ex}[R]$, confirming that the standard deviation measures how spread out the distribution of R is around its mean.

Proof. Substituting $x = c\sigma_R$ in Chebyshev's Theorem gives:

$$\Pr[|R - \text{Ex}[R]| \geq c\sigma_R] \leq \frac{\text{Var}[R]}{(c\sigma_R)^2} = \frac{\sigma_R^2}{(c\sigma_R)^2} = \frac{1}{c^2}.$$

■

The IQ Example

Suppose that, in addition to the national average IQ being 100, we also know the standard deviation of IQ's is 10. How rare is an IQ of 300 or more?

Let the random variable, R, be the IQ of a random person. So we are supposing that $\text{Ex}[R] = 100$, $\sigma_R = 10$, and R is nonnegative. We want to compute $\Pr[R \geq 300]$.

We have already seen that Markov's Theorem 18.2.1 gives a coarse bound, namely,

$$\Pr[R \geq 300] \leq \frac{1}{3}.$$

Now we apply Chebyshev's Theorem to the same problem:

$$\Pr[R \geq 300] = \Pr[|R - 100| \geq 200] \leq \frac{\text{Var}[R]}{200^2} = \frac{10^2}{200^2} = \frac{1}{400}.$$

So Chebyshev's Theorem implies that at most one person in four hundred has an IQ of 300 or more. We have gotten a much tighter bound using the additional information, namely the variance of R, than we could get knowing only the expectation.

18.4 Properties of Variance

The definition of variance of R as $\text{Ex}[(R - \text{Ex}[R])^2]$ may seem rather arbitrary. A direct measure of average deviation would be $\text{Ex}[\,|R - \text{Ex}[R]|\,]$. But the direct measure doesn't have the many useful properties that variance has, which is what this section is about.

18.4.1 A Formula for Variance

Applying linearity of expectation to the formula for variance yields a convenient alternative formula.

Lemma 18.4.1.

$$\text{Var}[R] = \text{Ex}[R^2] - \text{Ex}^2[R],$$

for any random variable, R.

Here we use the notation $\text{Ex}^2[R]$ as shorthand for $(\text{Ex}[R])^2$.

Proof. Let $\mu = \text{Ex}[R]$. Then

$$\begin{aligned}
\text{Var}[R] &= \text{Ex}[(R - \text{Ex}[R])^2] && \text{(Def 18.3.2 of variance)} \\
&= \text{Ex}[(R - \mu)^2] && \text{(def of } \mu\text{)} \\
&= \text{Ex}[R^2 - 2\mu R + \mu^2] \\
&= \text{Ex}[R^2] - 2\mu\,\text{Ex}[R] + \mu^2 && \text{(linearity of expectation)} \\
&= \text{Ex}[R^2] - 2\mu^2 + \mu^2 && \text{(def of } \mu\text{)} \\
&= \text{Ex}[R^2] - \mu^2 \\
&= \text{Ex}[R^2] - \text{Ex}^2[R]. && \text{(def of } \mu\text{)}
\end{aligned}$$

■

For example, if B is a Bernoulli variable where $p ::= \Pr[B = 1]$, then

Lemma 18.4.2.

$$\text{Var}[B] = p - p^2 = p(1-p) \tag{18.4}$$

Proof. By Lemma 17.4.2, $\text{Ex}[B] = p$. But since B only takes values 0 and 1, $B^2 = B$. So Lemma 18.4.2 follows immediately from Lemma 18.4.1. ■

18.4.2 Variance of Time to Failure

According to section 17.4.6, the mean time to failure is $1/p$ for a process that fails during any given hour with probability p. What about the variance? That is, let C be the hour of the first failure, so $\Pr[C = i] = (1-p)^{i-1}p$. We'd like to find a formula for $\text{Var}[C]$.

By Lemma 18.4.1,

$$\text{Var}[C] = \text{Ex}[C^2] - (1/p)^2 \tag{18.5}$$

so all we need is a formula for $\text{Ex}[C^2]$.

In section 17.4.6 we used conditional expectation to find the mean time to failure, and a similar approach works for the variance. Namely, the expected value of C^2 is the probability, p, of failure in the first hour times 1^2, plus $(1-p)$ times the

expected value of $(C+1)^2$. So

$$\begin{aligned}\operatorname{Ex}[C^2] &= p\cdot 1^2 + (1-p)\operatorname{Ex}[(C+1)^2]\\ &= p + (1-p)\left(\operatorname{Ex}[C^2] + \frac{2}{p} + 1\right)\\ &= p + (1-p)\operatorname{Ex}[C^2] + (1-p)\left(\frac{2}{p}+1\right) \text{ so}\end{aligned}$$

$$\begin{aligned}p\operatorname{Ex}[C^2] &= p + (1-p)\left(\frac{2}{p}+1\right)\\ &= \frac{p^2 + (1-p)(2+p)}{p} \text{ and}\end{aligned}$$

$$\operatorname{Ex}[C^2] = \frac{2-p}{p^2}$$

mean time to failure

18.4.3 Dealing with Constants

It helps to know how to calculate the variance of $aR+b$:

Theorem 18.4.3. *Let R be a random variable, and a a constant. Then*

$$\operatorname{Var}[aR] = a^2\operatorname{Var}[R]. \tag{18.6}$$

Proof. Beginning with the definition of variance and repeatedly applying linearity of expectation, we have:

$$\begin{aligned}\operatorname{Var}[aR] ::= \operatorname{Ex}[(aR - \operatorname{Ex}[aR])^2]\\ &= \operatorname{Ex}[(aR)^2 - 2aR\operatorname{Ex}[aR] + \operatorname{Ex}^2[aR]]\\ &= \operatorname{Ex}[(aR)^2] - \operatorname{Ex}[2aR\operatorname{Ex}[aR]] + \operatorname{Ex}^2[aR]\\ &= a^2\operatorname{Ex}[R^2] - 2\operatorname{Ex}[aR]\operatorname{Ex}[aR] + \operatorname{Ex}^2[aR]\\ &= a^2\operatorname{Ex}[R^2] - a^2\operatorname{Ex}^2[R]\\ &= a^2\left(\operatorname{Ex}[R^2] - \operatorname{Ex}^2[R]\right)\\ &= a^2\operatorname{Var}[R] \qquad \text{(by Lemma 18.4.1)}\end{aligned}$$

■

It's even simpler to prove that adding a constant does not change the variance, as the reader can verify:

Theorem 18.4.4. *Let R be a random variable, and b a constant. Then*

$$\operatorname{Var}[R+b] = \operatorname{Var}[R]. \tag{18.7}$$

Recalling that the standard deviation is the square root of variance, this implies that the standard deviation of $aR + b$ is simply $|a|$ times the standard deviation of R:

Corollary 18.4.5.

$$\sigma_{aR+b} = |a|\sigma_R.$$

18.4.4 Variance of a Sum

In general, the variance of a sum is not equal to the sum of the variances, but variances do add for *independent* variables. In fact, *mutual* independence is not necessary: *pairwise* independence will do. This is useful to know because there are some important situations involving variables that are pairwise independent but not mutually independent.

Theorem 18.4.6. *If R_1 and R_2 are independent random variables, then*

$$\text{Var}[R_1 + R_2] = \text{Var}[R_1] + \text{Var}[R_2]. \tag{18.8}$$

Proof. We may assume that $\text{Ex}[R_i] = 0$ for $i = 1, 2$, since we could always replace R_i by $R_i - \text{Ex}[R_i]$ in equation (18.8). This substitution preserves the independence of the variables, and by Theorem 18.4.4, does not change the variances.

Now by Lemma 18.4.1, $\text{Var}[R_i] = \text{Ex}[R_i^2]$ and $\text{Var}[R_1 + R_2] = \text{Ex}[(R_1 + R_2)^2]$, so we need only prove

$$\text{Ex}[(R_1 + R_2)^2] = \text{Ex}[R_1^2] + \text{Ex}[R_2^2]. \tag{18.9}$$

But (18.9) follows from linearity of expectation and the fact that

$$\text{Ex}[R_1 R_2] = \text{Ex}[R_1]\,\text{Ex}[R_2] \tag{18.10}$$

since R_1 and R_2 are independent:

$$\begin{aligned}
\text{Ex}[(R_1 + R_2)^2] &= \text{Ex}[R_1^2 + 2R_1R_2 + R_2^2] \\
&= \text{Ex}[R_1^2] + 2\,\text{Ex}[R_1R_2] + \text{Ex}[R_2^2] \\
&= \text{Ex}[R_1^2] + 2\,\text{Ex}[R_1]\,\text{Ex}[R_2] + \text{Ex}[R_2^2] \qquad \text{(by (18.10))} \\
&= \text{Ex}[R_1^2] + 2 \cdot 0 \cdot 0 + \text{Ex}[R_2^2] \\
&= \text{Ex}[R_1^2] + \text{Ex}[R_2^2]
\end{aligned}$$

■

An independence condition is necessary. If we ignored independence, then we would conclude that $\text{Var}[R+R] = \text{Var}[R]+\text{Var}[R]$. However, by Theorem 18.4.3, the left side is equal to $4\,\text{Var}[R]$, whereas the right side is $2\,\text{Var}[R]$. This implies that $\text{Var}[R] = 0$, which, by the Lemma above, essentially only holds if R is constant.

The proof of Theorem 18.4.6 carries over straightforwardly to the sum of any finite number of variables. So we have:

Theorem 18.4.7. *[Pairwise Independent Additivity of Variance] If $R_1, R_2, \ldots, R_n$ are* pairwise *independent random variables, then*

$$\text{Var}[R_1 + R_2 + \cdots + R_n] = \text{Var}[R_1] + \text{Var}[R_2] + \cdots + \text{Var}[R_n]. \tag{18.11}$$

Now we have a simple way of computing the variance of a variable, J, that has an (n, p)-binomial distribution. We know that $J = \sum_{k=1}^{n} I_k$ where the I_k are mutually independent indicator variables with $\Pr[I_k = 1] = p$. The variance of each I_k is $p(1-p)$ by Lemma 18.4.2, so by linearity of variance, we have

Lemma (Variance of the Binomial Distribution). *If J has the (n, p)-binomial distribution, then*

$$\text{Var}[J] = n\,\text{Var}[I_k] = np(1-p). \tag{18.12}$$

18.5 Estimation by Random Sampling

Polling again

Suppose we had wanted an advance estimate of the fraction of the Massachusetts voters who favored Scott Brown over everyone else in the recent Democratic primary election to fill ~~Senator Edward~~ Kennedy's seat.

Let p be this unknown fraction, and 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 Brown, and $K = 0$ otherwise.

Now to estimate p, we take a large number, n, of random choices of voters[1] and count the fraction who favor Brown. 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 Brown. Since our choices are made independently, the K_i's are

[1] 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.

independent. So formally, 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,

$$S_n ::= \sum_{i=1}^{n} K_i. \tag{18.13}$$

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 we will sample who favor Scott Brown. Most people intuitively expect this sample fraction to give a useful approximation to the unknown fraction, p —and they would be right. So we will use the sample value, S_n/n, as our *statistical estimate* of p and use the Pairwise Independent Sampling Theorem 18.5.1 to work out how good an estinate this is.

18.5.1 Sampling

Suppose we want our estimate to be within 0.04 of the Brown favoring fraction, p, at least 95% of the time. This means we want

$$\Pr\left[\left|\frac{S_n}{n} - p\right| \leq 0.04\right] \geq 0.95\,. \tag{18.14}$$

So we better determine the number, n, of times we must poll voters so that inequality (18.14) will hold.

Now S_n is binomially distributed, so from (18.12) we have

$$\mathrm{Var}[S_n] = n(p(1-p)) \leq n \cdot \frac{1}{4} = \frac{n}{4}$$

The bound of 1/4 follows from the fact that $p(1-p)$ is maximized when $p = 1-p$, that is, when $p = 1/2$ (check this yourself!).

Next, we bound the variance of S_n/n:

$$\begin{aligned}
\mathrm{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \mathrm{Var}[S_n] && \text{(by (18.6))} \\
&\leq \left(\frac{1}{n}\right)^2 \frac{n}{4} && \text{(by (18.5.1))} \\
&= \frac{1}{4n} && (18.15)
\end{aligned}$$

Now from Chebyshev and (18.15) we have:

$$\Pr\left[\left|\frac{S_n}{n} - p\right| \geq 0.04\right] \leq \frac{\mathrm{Var}[S_n/n]}{(0.04)^2} = \frac{1}{4n(0.04)^2} = \frac{156.25}{n} \tag{18.16}$$

To make our our estimate with 95% confidence, we want the righthand side of (18.16) to be at most 1/20. So we choose n so that

$$\frac{156.25}{n} \leq \frac{1}{20},$$

that is,

$$n \geq 3,125.$$

This est is better than Chebyshev

A more exact calculation of the tail of this binomial distribution shows that the above sample size is about four times larger than necessary, but it is still a feasible size to sample. The fact that the sample size derived using Chebyshev's Theorem was unduly pessimistic should not be surprising. After all, in applying the Chebyshev Theorem, we only used the variance of S_n. It makes sense that more detailed information about the distribution leads to better bounds. But working through this example using only the variance has the virtue of illustrating an approach to estimation that is applicable to arbitrary random variables, not just binomial variables.

18.5.2 Matching Birthdays

There are important cases where the relevant distributions are not binomial because the mutual independence properties of the voter preference example do not hold. In these cases, estimation methods based on the Chebyshev bound may be the best approach. Birthday Matching is an example. We already saw in Section 16.7 that in a class of 85 students it is virtually certain that two or more students will have the same birthday. This suggests that quite a few pairs of students are likely to have the same birthday. How many?

So as before, suppose there are n students and d days in the year, and let D be the number of pairs of students with the same birthday. Now it will be easy to calculate the expected number of pairs of students with matching birthdays. Then we can take the same approach as we did in estimating voter preferences to get an estimate of the probability of getting a number of pairs close to the expected number.

Unlike the situation with voter preferences, having matching birthdays for different pairs of students are not mutually independent events, but the matchings are *pairwise independent* —as explained in Section 16.7 (and proved in Problem 17.2). This will allow us to apply the same reasoning to Birthday Matching as we did for voter preference. Namely, let $B_1, B_2, \ldots, B_n$ be the birthdays of n independently chosen people, and let $E_{i,j}$ be the indicator variable for the event that the ith and jth people chosen have the same birthdays, that is, the event $[B_i = B_j]$. So our probability model, the B_i's are mutually independent variables, the $E_{i,j}$'s are pairwise independent. Also, the expectations of $E_{i,j}$ for $i \neq j$ equals the probability that $B_i = B_j$, namely, $1/d$.

Now, D, the number of matching pairs of birthdays among the n choices is simply the sum of the $E_{i,j}$'s:

$$D ::= \sum_{1 \le i < j \le n} E_{i,j}. \tag{18.17}$$

So by linearity of expectation

$$\mathrm{Ex}[D] = \mathrm{Ex}[\sum_{1 \le i < j \le n} E_{i,j}] = \sum_{1 \le i < j \le n} \mathrm{Ex}[E_{i,j}] = \binom{n}{2} \cdot \frac{1}{d}.$$

Similarly,

$$\begin{aligned} \mathrm{Var}[D] &= \mathrm{Var}[\sum_{1 \le i < j \le n} E_{i,j}] \\ &= \sum_{1 \le i < j \le n} \mathrm{Var}[E_{i,j}] && \text{(by Theorem 18.4.7)} \\ &= \binom{n}{2} \cdot \frac{1}{d}\left(1 - \frac{1}{d}\right). && \text{(by Lemma 18.4.2)} \end{aligned}$$

In particular, for a class of $n = 95$ students with $d = 365$ possible birthdays, we have $\mathrm{Ex}[D] < 12.23$ and $\mathrm{Var}[D] > 12.22(1-1/365) > 12.19$. So by Chebyshev's Theorem

$$\Pr[|D - 12.23| \ge x] < \frac{12.19}{x^2}.$$

Letting $x = 7$, we conclude that there is a better than %75 chance that in a class of 95 students, the number of pairs of students with the same birthday will be between 6 and 20.

18.5.3 Pairwise Independent Sampling

The reasoning we used above to analyze voter polling and matching birthdays is very similar. We summarize it in slightly more general form with a basic result we call the Pairwise Independent Sampling Theorem. In particular, we do not need to restrict ourselves to sums of zero-one valued variables, or to variables with the same distribution. For simplicity, we state the Theorem for pairwise independent variables with possibly different distributions but with the same mean and variance.

Theorem 18.5.1 (Pairwise Independent Sampling). *Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean, μ, and deviation, σ. Define*

$$S_n ::= \sum_{i=1}^{n} G_i. \tag{18.18}$$

Then

$$\Pr\left[\left|\frac{S_n}{n} - \mu\right| \geq x\right] \leq \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.$$

Proof. We observe first that the expectation of S_n/n is μ:

$$\begin{aligned}
\mathrm{Ex}\left[\frac{S_n}{n}\right] &= \mathrm{Ex}\left[\frac{\sum_{i=1}^n G_i}{n}\right] && \text{(def of } S_n\text{)} \\
&= \frac{\sum_{i=1}^n \mathrm{Ex}[G_i]}{n} && \text{(linearity of expectation)} \\
&= \frac{\sum_{i=1}^n \mu}{n} \\
&= \frac{n\mu}{n} = \mu.
\end{aligned}$$

The second important property of S_n/n is that its variance is the variance of G_i divided by n:

$$\begin{aligned}
\mathrm{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \mathrm{Var}[S_n] && \text{(by (18.6))} \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^n G_i\right] && \text{(def of } S_n\text{)} \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[G_i] && \text{(pairwise independent additivity)} \\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. && (18.19)
\end{aligned}$$

This is enough to apply Chebyshev's Theorem and conclude:

$$\begin{aligned}
\Pr\left[\left|\frac{S_n}{n} - \mu\right| \geq x\right] &\leq \frac{\mathrm{Var}[S_n/n]}{x^2}. && \text{(Chebyshev's bound)} \\
&= \frac{\sigma^2/n}{x^2} && \text{(by (18.19))} \\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

■

The Pairwise Independent Sampling Theorem provides a precise general statement about how the average of independent samples of a random variable approaches the mean. In particular, it proves what is known as the Law of Large

Numbers[2] : by choosing a large enough sample size, we can get arbitrarily accurate estimates of the mean with confidence arbitrarily close to 100%.

Corollary 18.5.2. *[Weak Law of Large Numbers] Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean, μ, and the same finite deviation, and let*

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

Then for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr[|S_n - \mu| \leq \epsilon] = 1.$$

18.6 Confidence versus Probability

So Chebyshev's Bound implies that sampling 3,125 voters will yield a fraction that, 95% of the time, is within 0.04 of the actual fraction of the voting population who prefer Brown.

Notice that the actual size of the voting population was never considered because *it did not matter*. People who have not studied probability theory often insist that the population size should matter. But our analysis shows that polling a little over 3000 people people is always sufficient, whether there are ten thousand, or million, or billion ... voters. You should think about an intuitive explanation that might persuade someone who thinks population size matters.

Now suppose a pollster actually takes a sample of 3,125 random voters to estimate the fraction of voters who prefer Brown, and the pollster finds that 1250 of them prefer Brown. It's tempting, **but sloppy**, to say that this means:

False Claim. *With probability 0.95, the fraction, p, of voters who prefer Brown is* $1250/3125 \pm 0.04$. *Since* $1250/3125 - 0.04 > 1/3$, *there is a 95% chance that more than a third of the voters prefer Brown to all other candidates.*

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 Brown is more than 1/3. 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 0.3; then it's nonsense to ask about the probability that it is within 0.04 of 1250/3125 —it simply isn't.

[2] This is the *Weak* Law of Large Numbers. As you might suppose, there is also a Strong Law, but it's outside the scope of 6.042.

This example of voter preference is typical: we want to estimate a fixed, unknown real-world quantity. But *being unknown does not make this quantity a random variable*, so it makes no sense to talk about the probability that it has some property.

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 Brown is $1250/3125 \pm 0.04$.

So confidence levels refer to the results of estimation procedures for real-world quantities. The phrase "confidence level" should be heard as a reminder that some statistical procedure was used to obtain an estimate, and in judging the credibility of the estimate, it may be important to learn just what this procedure was.

Problems for Section 18.2

Class Problems

Problem 18.1.

A herd of cows is stricken by an outbreak of *cold cow disease*. The disease lowers the normal body temperature of a cow, and a cow will die if its temperature goes below 90 degrees F. The disease epidemic is so intense that it lowered the average temperature of the herd to 85 degrees. Body temperatures as low as 70 degrees, **but no lower**, were actually found in the herd.

(a) Prove that at most 3/4 of the cows could have survived.

Hint: Let T be the temperature of a random cow. Make use of Markov's bound.

(b) Suppose there are 400 cows in the herd. Show that the bound of part (a) is best possible by giving an example set of temperatures for the cows so that the average herd temperature is 85, and with probability 3/4, a randomly chosen cow will have a high enough temperature to survive.

Homework Problems

Problem 18.2.

If R is a nonnegative random variable, then Markov's Theorem gives an upper

Now suppose a pollster actually takes a sample of 3,125 random voters to estimate the fraction of voters who prefer Brown, and the pollster finds that 1250 of them prefer Brown. It's tempting, **but sloppy**, to say that this means:

False Claim. *With probability 0.95, the fraction, p, of voters who prefer Brown is* $1250/3125 \pm 0.04$. *Since* $1250/3125 - 0.04 > 1/3$, *there is a 95% chance that more than a third of the voters prefer Brown to all other candidates.*

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 Brown is more than 1/3. 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 0.3; then it's nonsense to ask about the probability that it is within 0.04 of 1250/3125 —it simply isn't.

This example of voter preference is typical: we want to estimate a fixed, unknown real-world quantity. But *being unknown does not make this quantity a random variable*, so it makes no sense to talk about the probability that it has some property.

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 Brown is $1250/3125 \pm 0.04$.

So confidence levels refer to the results of estimation procedures for real-world quantities. The phrase "confidence level" should be heard as a reminder that some statistical procedure was used to obtain an estimate, and in judging the credibility of the estimate, it may be important to learn just what this procedure was.

18.7 Sums of Random Variables

If all you know about a random variable is its mean and variance, then Chebyshev's Theorem is the best you can do when it comes to bounding the probability that the random variable deviates from its mean. In some cases, however, we know

more —for example, that the random variable has a binomial distribution —and then it is possible to prove much stronger bounds. Instead of polynomially small bounds such as $1/c^2$, we can sometimes even obtain exponentially small bounds such as $1/e^c$. As we will soon discover, this is the case whenever the random variable T is the sum of n mutually independent random variables $T_1, T_2, \ldots, T_n$ where $0 \leq T_i \leq 1$. A random variable with a binomial distribution is just one of many examples of such a T. Here is another.

18.7.1 A Motivating Example

Fussbook is a new social networking site oriented toward unpleasant people.

Like all major web services, Fussbook has a load balancing problem. Specifically, Fussbook receives 24,000 forum posts every 10 minutes. Each post is assigned to one of m computers for processing, and each computer works sequentially through its assigned tasks. Processing an average post takes a computer 1/4 second. Some posts, such as pointless grammar critiques and snide witticisms, are easier. But the most protracted harangues require 1 full second.

Balancing the work load across the m computers is vital; if any computer is assigned more than 10 minutes of work in a 10-minute interval, then that computer is overloaded and system performance suffers. That would be bad, because Fussbook users are *not* a tolerant bunch.

An early idea was to assign each computer an alphabetic range of forum topics. ("That oughta work!", one programmer said.) But after the computer handling the "*pr*ivacy" and "*pr*eferred text editor" threads melted, the drawback of an ad hoc approach was clear: there are no guarantees.

If the length of every task were known in advance, then finding a balanced distribution would be a kind of "bin packing" problem. Such problems are hard to solve exactly, though approximation algorithms can come close. But in this case, task lengths are not known in advance, which is typical for workload problems in the real world.

So the load balancing problem seems sort of hopeless, because there is no data available to guide decisions. Heck, we might as well assign tasks to computers at random!

As it turns out, random assignment not only balances load reasonably well, but also permits provable performance guarantees in place of "That oughta work!" assertions. In general, a randomized approach to a problem is worth considering when a deterministic solution is hard to compute or requires unavailable information.

Some arithmetic shows that Fussbook's traffic is sufficient to keep $m = 10$ computers running at 100% capacity with perfect load balancing. Surely, more than 10 servers are needed to cope with random fluctuations in task length and imperfect

load balance. But how many is enough? 11? 15? 20? 100? We'll answer that question with a new mathematical tool.

18.7.2 The Chernoff Bound

The Chernoff[4] bound is a hammer that you can use to nail a great many problems. Roughly, the Chernoff bound says that certain random variables are very unlikely to significantly exceed their expectation. For example, if the expected load on a computer is just a bit below its capacity, then that computer is unlikely to be overloaded, provided the conditions of the Chernoff bound are satisfied.

More precisely, the Chernoff Bound says that *the sum of lots of little, independent random variables is unlikely to significantly exceed the mean of the sum.* The Markov and Chebyshev bounds lead to the same kind of conclusion but typically provide much weaker bounds. In particular, the Markov and Chebyshev bounds are polynomial, while the Chernoff bound is exponential.

Here is the theorem. The proof will come later in Section 18.7.5.

Theorem 18.7.1 (Chernoff Bound). *Let $T_1, \ldots T_n$ be mutually independent random variables such that $0 \le T_i \le 1$ for all i. Let $T = T_1 + \cdots + T_n$. Then for all $c \ge 1$,*

$$\Pr[T \ge c \operatorname{Ex}[T]] \le e^{-k \operatorname{Ex}[T]} \tag{18.23}$$

where $k = c \ln(c) - c + 1$.

The Chernoff bound applies only to distributions of sums of independent random variables that take on values in the interval $[0, 1]$. The binomial distribution is of course such a distribution, but there are lots of other distributions because the Chernoff bound allows the variables in the sum to have differing, arbitrary, and even unknown distributions over the range $[0, 1]$. Furthermore, there is no direct dependence on the number of random variables in the sum or their expectations. In short, the Chernoff bound gives strong results for lots of problems based on little information —no wonder it is widely used!

18.7.3 Chernoff Bound for Binomial Tails

The Chernoff bound is pretty easy to apply, though the details can be daunting at first. Let's walk through a simple example to get the hang of it: getting bounds on the tail of a binomial distribution, for example, bounding the probability that the number of heads that come up in 1000 independent tosses of a coin exceeds the

[4] Yes, this is the same Chernoff who figured out how to beat the state lottery —this guy knows a thing or two.

expectation by 20% or more? Let T_i be an indicator variable for the event that the ith coin is heads. Then the total number of heads is

$$T = T_1 + \cdots + T_{1000}.$$

The Chernoff bound requires that the random variables T_i be mutually independent and take on values in the range $[0, 1]$. Both conditions hold here. In this example the T_i's only take the two values 0 and 1, since they're indicators.

The goal is to bound the probability that the number of heads exceeds its expectation by 20% or more; that is, to bound $\Pr[T \geq c\,\mathrm{Ex}[T]]$ where c = 1.2. To that end, we compute k as defined in the theorem:

$$k = c\ln(c) - c + 1 = 0.0187\ldots.$$

If we assume the coin is fair, then $\mathrm{Ex}[T] = 500$. Plugging these values into the Chernoff bound gives:

$$\begin{aligned}\Pr\left[T \geq 1.2\,\mathrm{Ex}[T]\right] &\leq e^{-k\,\mathrm{Ex}[T]} \\ &= e^{-(0.0187\ldots)\cdot 500} < 0.0000834.\end{aligned}$$

So the probability of getting 20% or more extra heads on 1000 coins is less than 1 in 10,000.

The bound becomes much stronger as the number of coins increases, because the expected number of heads appears in the exponent of the upper bound. For example, the probability of getting at least 20% extra heads on a million coins is at most

$$e^{-(0.0187\ldots)\cdot 500000} < e^{-9392},$$

which is an inconceivably small number.

Alternatively, the bound also becomes stronger for larger deviations. For example, suppose we're interested in the odds of getting 30% or more extra heads in 1000 tosses, rather than 20%. In that case, $c = 1.3$ instead of 1.2. Consequently, the parameter k rises from 0.0187 to about 0.0410, which may not seem significant, but because k appears in the exponent of the upper bound, the final probability decreases from around 1 in 10,000 to about 1 in a billion!

18.7.4 Chernoff Bound for a Lottery Game

Pick-4 is a lottery game where you pay $1 to pick a 4-digit number between 0000 and 9999. If your number comes up in a random drawing, then you win $5,000. Your chance of winning is 1 in 10,000. If 10 million people play, then the expected number of winners is 1000. Whn there are exactly 1000 winners, the lottery keeps

$5 million of the $10 million paid for tickets. The lottery operator's nightmare is that the number of winners is much greater —say at the 2000 or greater point where the lottery has to pay out more than it received.. What is the probability that will happen?

Let T_i be an indicator for the event that the ith player wins. Then $T = T_1 + \cdots + T_n$ is the total number of winners. If we assume[5] that the players' picks and the winning number are random, independent and uniform, then the indicators T_i are independent, as required by the Chernoff bound.

Since 2000 winners would be twice the expected number, we choose $c = 2$, compute $k = c\ln(c) - c + 1 = 0.386\ldots$, and plug these values into the Chernoff bound:

$$\begin{aligned}\Pr[T \geq 2000] &= \Pr\left[T \geq 2\,\mathrm{Ex}[T]\right] \\ &\leq e^{-k\,\mathrm{Ex}[T]} = e^{-(0.386\ldots)\cdot 1000} \\ &< e^{-386}.\end{aligned}$$

So there is almost no chance that the lottery operator pays out double. In fact, the number of winners won't even be 10% higher than expected very often. To prove that, let $c = 1.1$, compute $k = c\ln(c) - c + 1 = 0.00484\ldots$, and plug in again:

$$\begin{aligned}\Pr\left[T \geq 1.1\,\mathrm{Ex}[T]\right] &\leq e^{-k\,\mathrm{Ex}[T]} \\ &= e^{-(0.00484)\cdot 1000} < 0.01.\end{aligned}$$

So the Pick-4 lottery may be exciting for the players, but the lottery operator has little doubt about the outcome!

Randomized Load Balancing

Now let's return to Fussbook and its load balancing problem. Specifically, we need to determine how many machines suffice to ensure that no server is overloaded; that is, assigned to do more than 10 minutes of work in a 10-minute interval. So a server is overloaded if it gets assigned more than 600 seconds of work.

To begin, let's find the probability that the first server is overloaded. Letting T be the number of seconds of work assigned to the first server, this means we want an upper bound on $\Pr[T \geq 600]$. Let T_i be the number of seconds that the first server spends on the ith task: then T_i is zero if the task is assigned to another machine,

[5] As we noted in Chapter 17, human choices are often not uniform and they can be highly dependent. For example, lots of people will pick an important date. So the lottery folks should not get too much comfort from the analysis that follows, unless they assign random 4-digit numbers to each player.

and otherwise T_i is the length of the task. So $T = \sum_{i=1}^{n} T_i$ is the total length of tasks assigned to the first server, where $n = 24{,}000$.

The Chernoff bound is applicable only if the T_i are mutually independent and take on values in the range $[0, 1]$. The first condition is satisfied if we assume that task lengths and assignments are independent. And the second condition is satisfied because processing even the most interminable harangue takes at most 1 second.

In all, there are 24,000 tasks, each with an expected length of 1/4 second. Since tasks are assigned to computers at random, the expected load on the first server is:

$$\begin{aligned} \mathrm{Ex}[T] &= \frac{24{,}000 \text{ tasks} \cdot 1/4 \text{ second per task}}{m \text{ machines}} \\ &= 6000/m \text{ seconds.} \end{aligned} \tag{18.24}$$

For example, if there are fewer than 10 machines, then the expected load on the first server is greater than its capacity, and we can expect it to be overloded. If there are exactly 10 machines, then the server is expected to run for $6000/10 = 600$ seconds, which is 100% of its capacity.

Now we can use the Chernoff bound to upper bound the probability that the first server is overloaded. We have from (18.24)

$$600 = c\,\mathrm{Ex}[T] \qquad \text{where } c ::= m/10,$$

so by the Chernoff bound

$$\Pr[T \geq 600] = \Pr[T \geq c\,\mathrm{Ex}[T]] \leq e^{-(c\ln(c)-c+1)\cdot 6000/m},$$

The probability that *some* server is overloaded is at most m times the probability that the first server is overloaded, by the Union Bound in Section 16.4.2. So

$$\begin{aligned} \Pr[\text{some server is overloaded}] &\leq \sum_{i=1}^{m} \Pr[\text{server } i \text{ is overloaded}] \\ &= m\,\Pr[\text{the first server is overloaded}] \\ &\leq m e^{-(c\ln(c)-c+1)\cdot 6000/m}, \end{aligned}$$

where $c = m/10$. Some values of this upper bound are tabulated below:

$$\begin{aligned} m &= 11: \quad 0.784\ldots \\ m &= 12: \quad 0.000999\ldots \\ m &= 13: \quad 0.0000000760\ldots \end{aligned}$$

These values suggest that a system with $m = 11$ machines might suffer immediate overload, $m = 12$ machines could fail in a few days, but $m = 13$ should be fine for a century or two!

18.7.5 Proof of the Chernoff Bound

The proof of the Chernoff bound is somewhat involved. Heck, even *Chernoff* didn't come up with it! His friend, Herman Rubin, showed him the argument. Thinking the bound not very significant, Chernoff did not credit Rubin in print. He felt pretty bad when it became famous![6]

Proof of Theorem 18.7.1. For clarity, we'll go through the proof "top down." That is, we'll use facts that are proved immediately afterward.

The key step is to exponentiate both sides of the inequality $T \geq c \operatorname{Ex}[T]$ and then apply the Markov bound:

$$\begin{aligned}
\Pr[T \geq c \operatorname{Ex}[T]] &= \Pr[c^T \geq c^{c \operatorname{Ex}[T]}] \\
&\leq \frac{\operatorname{Ex}[c^T]}{c^{c \operatorname{Ex}[T]}} && \text{(by Markov)} \\
&\leq \frac{e^{(c-1)\operatorname{Ex}[T]}}{c^{c \operatorname{Ex}[T]}} && \text{(by Lemma 18.7.2 below)} \\
&= \frac{e^{(c-1)\operatorname{Ex}[T]}}{e^{c \ln(c) \operatorname{Ex}[T]}} = e^{-(c \ln(c) - c + 1)\operatorname{Ex}[T]}.
\end{aligned}$$

■

Algebra aside, there is a brilliant idea in this proof: in this context, exponentiating somehow supercharges the Markov bound. This is not true in general! One unfortunate side-effect is that we have to bound some nasty expectations involving exponentials in order to complete the proof. This is done in the two lemmas below, where variables take on values as in Theorem 18.7.1.

Lemma 18.7.2.

$$\operatorname{Ex}[c^T] \leq e^{(c-1)\operatorname{Ex}[T]}.$$

[6] See "A Conversation with Herman Chernoff," *Statistical Science* 1996, Vol. 11, No. 4, pp 335–350.

Sub parts of the proof to prove

Proof.

$$\begin{aligned}
\operatorname{Ex}[c^T] &= \operatorname{Ex}[c^{T_1+\cdots+T_n}] && \text{(def of } T\text{)}\\
&= \operatorname{Ex}[c^{T_1}\cdots c^{T_n}]\\
&= \operatorname{Ex}[c^{T_1}]\cdots\operatorname{Ex}[c^{T_n}] && \text{(independent product Cor 17.5.7)}\\
&\le e^{(c-1)\operatorname{Ex}[T_1]}\ldots e^{(c-1)\operatorname{Ex}[T_n]} && \text{(by Lemma 18.7.3 below)}\\
&= e^{(c-1)(\operatorname{Ex}[T_1]+\cdots+\operatorname{Ex}[T_n])}\\
&= e^{(c-1)\operatorname{Ex}[T_1+\cdots+T_n]} && \text{(linearity of Ex[·])}\\
&= e^{(c-1)\operatorname{Ex}[T]}.
\end{aligned}$$

just rearranges + plugs back

■

Lemma 18.7.3.

$$\operatorname{Ex}[c^{T_i}] \le e^{(c-1)\operatorname{Ex}[T_i]}$$

Proof. All summations below range over values v taken by the random variable T_i, which are all required to be in the interval $[0, 1]$.

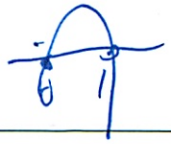
$$\begin{aligned}
\operatorname{Ex}[c^{T_i}] &= \sum c^v \Pr[T_i = v] && \text{(def of Ex[·])}\\
&\le \sum (1 + (c-1)v)\Pr[T_i = v] && \text{(convexity —see below)}\\
&= \sum \Pr[T_i = v] + (c-1)v\Pr[T_i = v]\\
&= \sum \Pr[T_i = v] + (c-1)\sum v\Pr[T_i = v]\\
&= 1 + (c-1)\operatorname{Ex}[T_i]\\
&\le e^{(c-1)\operatorname{Ex}[T_i]} && \text{(since } 1 + z \le e^z\text{)}.
\end{aligned}$$

The second step relies on the inequality

$$c^v \le 1 + (c-1)v,$$

which holds for all v in $[0, 1]$ and $c \ge 1$. This follows from the general principle that a convex function, namely c^v, is less than the linear function, $1 + (c-1)v$, between their points of intersection, namely $v = 0$ and 1. This inequality is why the variables T_i are restricted to the interval $[0, 1]$. ■

hmm never saw before but makes sense



or

still less than!

18.7.6 Comparing the Bounds

Suppose that we have a collection of mutually independent events $A_1, A_2, \ldots, A_n$, and we want to know how many of the events are likely to occur.

Let T_i be the indicator random variable for A_i and define

$$p_i = \Pr[T_i = 1] = \Pr[A_i]$$

for $1 \leq i \leq n$. Define

$$T = T_1 + T_2 + \cdots + T_n$$

to be the number of events that occur.

We know from Linearity of Expectation that

$$\begin{aligned} \operatorname{Ex}[T] &= \operatorname{Ex}[T_1] + \operatorname{Ex}[T_2] + \cdots + \operatorname{Ex}[T_n] \\ &= \sum_{i=1}^{n} p_i. \end{aligned}$$

This is true even if the events are *not* independent.

By Theorem 18.4.8, we also know that

$$\begin{aligned} \operatorname{Var}[T] &= \operatorname{Var}[T_1] + \operatorname{Var}[T_2] + \cdots + \operatorname{Var}[T_n] \\ &= \sum_{i=1}^{n} p_i(1 - p_i), \end{aligned}$$

and thus that

$$\sigma_T = \sqrt{\sum_{i=1}^{n} p_i(1 - p_i)}.$$

This is true even if the events are only pairwise independent.

Markov's Theorem tells us that for any $c > 1$,

$$\Pr[T \geq c \operatorname{Ex}[T]] \leq \frac{1}{c}.$$

Chebyshev's Theorem gives us the stronger result that

$$\Pr[|T - \operatorname{Ex}[T]| \geq c\sigma_T] \leq \frac{1}{c^2}.$$

The Chernoff Bound gives us an even stronger result, namely, that for any $c > 0$,

$$\Pr[T - \operatorname{Ex}[T] \geq c \operatorname{Ex}[T]] \leq e^{-(c \ln(c) - c + 1)\operatorname{Ex}[T]}.$$

In this case, the probability of exceeding the mean by $c \operatorname{Ex}[T]$ decreases as an exponentially small function of the deviation.

By considering the random variable $n - T$, we can also use the Chernoff Bound to prove that the probability that T is much lower than $\operatorname{Ex}[T]$ is also exponentially small.

'this only works so well for 'indicator RV Or some [0,1] RV

Linearity of Expectation if ind or not

Var for indicator/ Bernoulli

'the reverse'

wish still had G.O.M.I. book

18.7.7 Murphy's Law

If the expectation of a random variable is much less than 1, then Markov's Theorem implies that there is only a small probability that the variable has a value of 1 or more. On the other hand, a result that we call *Murphy's Law*[7] says that if a random variable is an independent sum of 0-1-valued variables and has a large expectation, then there is a huge probability of getting a value of at least 1.

Theorem 18.7.4 (Murphy's Law). *Let $A_1, A_2, \ldots, A_n$ be mutually independent events. Let T_i be the indicator random variable for A_i and define*

$$T ::= T_1 + T + 2 + \cdots + T_n$$

to be the number of events that occur. Then

$$\Pr[T = 0] \leq e^{-\mathrm{Ex}[T]}.$$

Proof.

$$\begin{aligned}
\Pr[T = 0] &= \Pr[\overline{A}_1 \wedge \overline{A}_2 \wedge \cdots \wedge \overline{A}_n] \\
&= \prod_{i=1}^{n} \Pr[\overline{A}_i] && \text{(by independence of } A_i\text{)} \\
&= \prod_{i=1}^{n} (1 - \Pr[A_i]) \\
&\leq \prod_{i=1}^{n} e^{-\Pr[A_i]} && \text{(since } 1 - x \leq e^{-x}\text{)} \\
&= e^{-\sum_{i=1}^{n} \Pr[A_i]} \\
&= e^{-\sum_{i=1}^{n} \mathrm{Ex}[T_i]} && \text{(since } T_i \text{ is an indicator for } A_i\text{)} \\
&= e^{-\mathrm{Ex}[T]} && \text{(linearity of expectation)}
\end{aligned}$$

■

For example, given any set of mutually independent events, if you expect 10 of them to happen, then at least one of them will happen with probability at least $1 - e^{-10}$. The probability that none of them happen is at most $e^{-10} < 1/22000$.

So if there are a lot of independent things that can go wrong and their probabilities sum to a number much greater than 1, then Theorem 18.7.4 proves that some of them surely will go wrong.

[7] This is in reference and deference to the famous saying that "If something can go wrong, it will go wrong."

This result can help to explain "coincidences," "miracles," and crazy events that seem to have been very unlikely to happen. Such events do happen, in part, because there are so many possible unlikely events that the sum of their probabilities is greater than one. For example, someone *does* win the lottery.

In fact, if there are 100,000 random tickets in Pick-4, Theorem 18.7.4 says that the probability that there is no winner is less than $e^{-10} < 1/22000$. More generally, there are literally millions of one-in-a-million possible events and so some of them will surely occur.

weird fact

18.8 Coping with Infinity

Section Blank?

Problems for Section 18.2

Class Problems

Problem 18.1.

A herd of cows is stricken by an outbreak of *cold cow disease*. The disease lowers the normal body temperature of a cow, and a cow will die if its temperature goes below 90 degrees F. The disease epidemic is so intense that it lowered the average temperature of the herd to 85 degrees. Body temperatures as low as 70 degrees, **but no lower**, were actually found in the herd.

(a) Prove that at most 3/4 of the cows could have survived.

Hint: Let T be the temperature of a random cow. Make use of Markov's bound.

(b) Suppose there are 400 cows in the herd. Show that the bound of part (a) is best possible by giving an example set of temperatures for the cows so that the average herd temperature is 85, and with probability 3/4, a randomly chosen cow will have a high enough temperature to survive.

Homework Problems

Problem 18.2.

If R is a nonnegative random variable, then Markov's Theorem gives an upper bound on $\Pr[R \geq x]$ for any real number $x > \mathrm{Ex}[R]$. If a constant $b \geq 0$ is a lower bound on R, then Markov's Theorem can also be applied to $R - b$ to obtain a possibly different bound on $\Pr[R \geq x]$.

(a) Show that if $b > 0$, applying Markov's Theorem to $R - b$ gives a smaller upper bound on $\Pr[R \geq x]$ than simply applying Markov's Theorem directly to R.

next pg

(b) What value of $b \geq 0$ in part (a) gives the best bound?

Problems for Section 18.4

Practice Problems

Problem 18.3.
A gambler plays 120 hands of draw poker, 60 hands of black jack, and 20 hands of stud poker per day. He wins a hand of draw poker with probability 1/6, a hand of black jack with probability 1/2, and a hand of stud poker with probability 1/5.

(a) What is the expected number of hands the gambler wins in a day?

(b) What would the Markov bound be on the probability that the gambler will win at least 108 hands on a given day?

(c) Assume the outcomes of the card games are pairwise independent. What is the variance in the number of hands won per day?

(d) What would the Chebyshev bound be on the probability that the gambler will win at least 108 hands on a given day? You may answer with a numerical expression that is not completely evaluated.

Problem 18.4. **(a)** A computer program crashes at the end of each hour of use with probability $1/p$, if it has not crashed already. If H is the number of hours until the first crash, we know

$$\text{Ex}[H] = \frac{1}{p}, \qquad \text{(Equation (17.8))}$$

$$\text{Var}[H] = \frac{q}{p^2} \qquad \text{(Equation (18.8)),}$$

where $q ::= 1 - p$.

(b) What is the Chebyshev bound on

$$\Pr[|H - (1/p)| > x/p]$$

where $x > 0$?

(c) Conclude from part (b) that for $a \geq 2$,

$$\Pr[H > a/p] \leq \frac{1-p}{(a-1)^2}$$

Hint: Check that $|H - (1/p)| > (a-1)/p$ iff $H > a/p$.

For example, given any set of mutually independent events, if you expect 10 of them to happen, then at least one of them will happen with probability at least $1 - e^{-10}$. The probability that none of them happen is at most $e^{-10} < 1/22000$.

So if there are a lot of independent things that can go wrong and their probabilities sum to a number much greater than 1, then Theorem 18.7.4 proves that some of them surely will go wrong.

This result can help to explain "coincidences," "miracles," and crazy events that seem to have been very unlikely to happen. Such events do happen, in part, because there are so many possible unlikely events that the sum of their probabilities is greater than one. For example, someone *does* win the lottery.

In fact, if there are 100,000 random tickets in Pick-4, Theorem 18.7.4 says that the probability that there is no winner is less than $e^{-10} < 1/22000$. More generally, there are literally millions of one-in-a-million possible events and so some of them will surely occur.

18.8 Really Great Expectations

Making independent tosses of a fair coin until some desired pattern comes up is a simple process you should feel solidly in command of by now, right? So how about a bet about the simplest such process —tossing until a head comes up? Ok, you're wary of betting with us, but how about this: we'll let *you set the odds.*

18.8.1 Repeating Yourself

Here's the bet: you make independent tosses of a fair coin until a head comes up. Then you will repeat the process. If a second head comes up in the same or fewer tosses than the first, you have to start over yet again. You keep starting over until you finally toss a run of tails longer than your first one. The payment rules are that you will pay me 1 cent each time you start over. When you win by finally getting a run of tails longer than your first one, I will pay you some generous amount. And by the way, you're certain to win —whatever your initial run of tails happened to be, a longer run will occur again with probability 1!

For example, if your first tosses are TTTH, then you will keep tossing until you get a run of 4 tails. So your winning flips might be

TTTHTHTTHHTTHTHTTTHTHHTTTT.

1 1 1 5 4 5 3 3 Doe

In this run there are 10 heads, which means you had to start over 9 times. So you would have paid me 9 cents by the time you finally won by tossing 4 tails. Now

you've won, and I'll pay you generously —how does 25 cents sound? Maybe you'd rather have \$1? How about \$10?

Of course there's a trap here. Let's calculate your expected winnings.

Suppose your initial run of tails had length k. After that, each time a head comes up, you have to start over and try to get $k+1$ tails in a row. If we regard your getting $k+1$ tails in a row as a "failed" try, and regard your having to start over because a head came up too soon as a "successful" try, then the number of times you have to start over is the number of tries till the first failure. So the expected number of tries will be the mean time to failure, which is 2^{k+1}, because the probability of tossing $k+1$ tails in a row is $2^{-(k+1)}$.

Let T be the length of your initial run of tails. So $T = k$ means that your initial tosses were $\mathrm{T}^k\mathrm{H}$. Let R be the number of times you repeat trying to beat your original run of tails. The number of cents you expect to finish with is the number of cents in my generous payment minus $\mathrm{Ex}[R]$. It's now easy to calculate $\mathrm{Ex}[R]$ by conditioning on the value of T:

$$\mathrm{Ex}[R] = \sum_{k\in\mathbb{N}} \mathrm{Ex}[R \mid T = k] \cdot \Pr[T = k] = \sum_{k\in\mathbb{N}} 2^{k+1} \cdot 2^{-(k+1)} = \sum_{k\in\mathbb{N}} 1 = \infty.$$

So you can expect to pay me an infinite number of cents before winning my "generous" payment. No amount of generosity can make this bet fair!

We haven't faced infinite expectations until now, but they just popped up in a very simple way. In fact this particular example is a special case of an astonishingly general one worked out in Problem 18.23: the expected waiting time for *any* random variable to achieve a larger value is infinite.

18.8.2 The St. Petersburg Paradox

One of the simplest casino bets is on "red" or "black" at the roulette table. In each play at roulette, a small ball is set spinning around a roulette wheel until it lands in a red, black, or green colored slot. The payoff for a bet on red or black matches the bet; for example, if you bet \$10 on red and the ball lands in a red slot, you get back your original \$10 bet plus another matching \$10.

In the US, a roulette wheel has two green slots among 18 black and 18 red slots, so the probability of red is $18/38 \approx 0.473$. In Europe, where roulette wheels have only one green slot, the odds for red are a little better —that is, $18/37 \approx 0.486$ —but still less than even.

There is a notorious gambling strategy allegedly used against the casino in St. Peterburg way back in czarist days: bet \$10 on red, and keep doubling the bet until a red comes up. This strategy implies that a player will leave the game as a net winner of \$10 as soon as the red first appears.

Suppose you had the good fortune to gamble against a fair roulette wheel. Then whatever your bet on a spin of the wheel, you are equally likely to win or lose, and your expected win is 0. This also means that the expected win after any given number of spins remains zero, so even playing the St. Peterburg strategy it seems your expected win would be 0.

But wait a minute. As long as there is a fixed, positive probability of red appearing on each spin of the wheel, it's *certain* that red will eventually come up. That is, you can be certain of leaving the casino having won \$10. This implies that even against an *unfair* roulette wheel, your expected win is \$10, contradicting the idea that you can't expect to win in a game that's biased against you.

This is paradoxical and something's obviously wrong here. In fact, there are two things wrong.

The first thing that's wrong is the argument claiming that the expectation is 0. It would be 0 if the number of bets had a fixed bound. If you could only make n bets, then your expectation in the fair game would be the sum of your expected wins on each of the bets, namely, $n \cdot 0 = 0$. But there is no such fixed bound, and that changes things.

To explain this carefully, let C_i be the number of dollars won on the ith spin. So $C_i = 2^{i-1}$ when red comes up for the first time on the ith spin, and $C_i = -2^{i-1}$, when the first red spin comes up after the ith spin. We can define C_i to be 0 if the first red comes up before the ith spin. This means

$$\mathrm{Ex}[C_i] = 0.$$

Also, the total of your winnings is

$$C ::= \sum_{i \in \mathbb{Z}^+} C_i.$$

The conclusion that $\mathrm{Ex}[C] = 10$ follows from Total Expectation, conditioning on the number of spins till a red first occurs. Namely, if the first red occurs on the ith spin, the amount won is

$$-10 \cdot (1 + 2 + 2^2 + \cdots + 2^{i-2}) + 10 \cdot 2^{i-1} = 10.$$

Then by Total Expectation,

$$\begin{aligned}
\mathrm{Ex}[C] &= \sum_{i \in \mathbb{Z}^+} \mathrm{Ex}[C \mid \text{first red on } i\text{th spin}] \cdot \Pr[\text{first red on } i\text{th spin}] \\
&= \sum_{i \in \mathbb{Z}^+} 10 \cdot 2^{-i} \\
&= 10 \cdot \sum_{i \in \mathbb{Z}^+} 2^{-i} = 10 \cdot 1 = 10.
\end{aligned}$$

So sure enough,

$$\mathrm{Ex}[C] ::= \mathrm{Ex}[\sum_{i\in\mathbb{Z}^+} C_i] = 10. \tag{18.25}$$

But since $\mathrm{Ex}[C_i] = 0$,

$$\sum_{i\in\mathbb{Z}^+} \mathrm{Ex}[C_i] = \sum_{i\in\mathbb{Z}^+} 0 = 0. \tag{18.26}$$

It seems that (18.26) and (18.25) contradict each other, but they don't. The apparent contradiction comes from applying infinite linearity to conclude

False Claim.

$$\mathrm{Ex}\left[\sum_{i\in\mathbb{Z}^+} C_i\right] = \sum_{i\in\mathbb{Z}^+} \mathrm{Ex}[C_i].$$

But this is a case where the convergence conditions required for infinite linearity don't hold. Even though the left hand sum converges (to 10) and the right hand sum converges (to 0), the infinite linearity Theorem (17.5.5) requires that the sum of expectations of *absolute values* converges. That is, infinite linearity would follow if the sum

$$\sum_{i\in\mathbb{Z}^+} \mathrm{Ex}[\,|C_i|\,] \tag{18.27}$$

converged. But

$$\begin{aligned}\mathrm{Ex}[\,|C_i|\,] &= (|10\cdot 2^{i-1}|)\cdot \Pr[\text{1st red in } i\text{th spin}]\\ &\quad + (|-10\cdot 2^{i-1}|)\cdot \Pr[\text{1st red after } i\text{th spin}]\\ &\quad + 0\cdot \Pr[\text{1st red before the } i\text{th spin}]\\ &= (10\cdot 2^{i-1})\cdot 2^{-(i)} + (10\cdot 2^{i-1})\cdot 2^{-(i)} + 0 = 10,\end{aligned}$$

so the sum (18.27) diverges —rapidly.

Probability theory truly leads to this absurd conclusion: a game entailing an unbounded number of fair bets may not be fair in the end. In fact, even against an *unfair* wheel, as long as there is some fixed positive probability of red on each spin, you are certain to win \$10 playing the St. Petersburg strategy!

This brings us to the second thing that's wrong here: you may wind up losing a lot of money before you catch up with your net win of \$10. Let L be the number of dollars you need to have in order to keep betting until the wheel finally spins red. If red first comes up on the ith spin, then L would equal

$$10(1+2+4+\cdots+2^i) = 10(2^{i+1}-1)$$

By Total Expectation,

$$\begin{aligned}\mathrm{Ex}[L] &= \sum_{i\in\mathbb{Z}^+} \mathrm{Ex}[L \mid 1\text{st red in } i\text{th spin}] \cdot \Pr[1\text{st red in } i\text{th spin}] \\ &= \sum_{i\in\mathbb{Z}^+} (10\cdot(2^{i+1}-1))\cdot 2^{-i} \geq \sum_{i\in\mathbb{Z}^+} 10 = \infty.\end{aligned}$$

That is, you can expect to lose an infinite amount of money before finally winning \$10 —giving you a percentage profit of 0.

So yes, probability theory leads to the absurd conclusion that, even with the odds heavily against you, you're certain to win playing roulette, but only if you make the absurd assumption that you have an infinite bankroll. We can't fault the theory for reaching an absurd conclusion from an absurd assumption.

Problems for Section 18.2

Class Problems

Problem 18.1.
A herd of cows is stricken by an outbreak of *cold cow disease*. The disease lowers the normal body temperature of a cow, and a cow will die if its temperature goes below 90 degrees F. The disease epidemic is so intense that it lowered the average temperature of the herd to 85 degrees. Body temperatures as low as 70 degrees, **but no lower**, were actually found in the herd.

(a) Prove that at most 3/4 of the cows could have survived.

Hint: Let T be the temperature of a random cow. Make use of Markov's bound.

(b) Suppose there are 400 cows in the herd. Show that the bound of part (a) is best possible by giving an example set of temperatures for the cows so that the average herd temperature is 85, and with probability 3/4, a randomly chosen cow will have a high enough temperature to survive.

Homework Problems

Problem 18.2.
If R is a nonnegative random variable, then Markov's Theorem gives an upper bound on $\Pr[R \geq x]$ for any real number $x > \mathrm{Ex}[R]$. If a constant $b \geq 0$ is a lower bound on R, then Markov's Theorem can also be applied to $R - b$ to obtain a possibly different bound on $\Pr[R \geq x]$.

(a) Show that if $b > 0$, applying Markov's Theorem to $R - b$ gives a smaller upper bound on $\Pr[R \geq x]$ than simply applying Markov's Theorem directly to R.

(b) What value of $b \geq 0$ in part (a) gives the best bound?

Problems for Section 18.4

Practice Problems

Problem 18.3.
A gambler plays 120 hands of draw poker, 60 hands of black jack, and 20 hands of stud poker per day. He wins a hand of draw poker with probability 1/6, a hand of black jack with probability 1/2, and a hand of stud poker with probability 1/5.

(a) What is the expected number of hands the gambler wins in a day?

(b) What would the Markov bound be on the probability that the gambler will win at least 108 hands on a given day?

(c) Assume the outcomes of the card games are pairwise independent. What is the variance in the number of hands won per day?

(d) What would the Chebyshev bound be on the probability that the gambler will win at least 108 hands on a given day? You may answer with a numerical expression that is not completely evaluated.

Problem 18.4. **(a)** A computer program crashes at the end of each hour of use with probability $1/p$, if it has not crashed already. If H is the number of hours until the first crash, we know

$$\mathrm{Ex}[H] = \frac{1}{p}, \qquad \text{(Equation (17.8))}$$

$$\mathrm{Var}[H] = \frac{q}{p^2} \qquad \text{(Equation (18.8)),}$$

where $q ::= 1 - p$.

(b) What is the Chebyshev bound on

$$\Pr[|H - (1/p)| > x/p]$$

where $x > 0$?

(c) Conclude from part (b) that for $a \geq 2$,

$$\Pr[H > a/p] \leq \frac{1-p}{(a-1)^2}$$

Hint: Check that $|H - (1/p)| > (a-1)/p$ iff $H > a/p$.

(d) What actually is

$$\Pr[H > a/p]?$$

Conclude that for any fixed $p > 0$, the probability that $H > a/p$ is an asymptotically smaller function of a than the Chebyshev bound of part (c).

Class Problems

Problem 18.5.

The hat-check staff has had a long day serving at a party, and at the end of the party they simply return the n checked hats in a random way such that the probability that any particular person gets their own hat back is $1/n$.

Let X_i be the indicator variable for the ith person getting their own hat back. Let S_n be the total number of people who get their own hat back.

(a) What is the expected number of people who get their own hat back?

(b) Write a simple formula for $\text{Ex}[X_i X_j]$ for $i \neq j$. *Hint:* What is $\Pr[X_j = 1 \mid X_i = 1]$?

(c) Explain why you cannot use the variance of sums formula to calculate $\text{Var}[S_n]$.

(d) Show that $\text{Ex}[S_n^2] = 2$. *Hint:* $X_i^2 = X_i$.

(e) What is the variance of S_n?

(f) Show that there is at most a 1% chance that more than 10 people get their own hat back. Try to give an intuitive explanation of why the chance remains this small regardless of n.

Problem 18.6.

For any random variable, R, with mean, μ, and standard deviation, σ, the Chebyshev Bound says that for any real number $x > 0$,

$$\Pr[|R - \mu| \geq x] \leq \left(\frac{\sigma}{x}\right)^2.$$

Show that for any real number, μ, and real numbers $x \geq \sigma > 0$, there is an R for which the Chebyshev Bound is tight, that is,

$$\Pr[|R| \geq x] = \left(\frac{\sigma}{x}\right)^2. \tag{18.28}$$

Hint: First assume $\mu = 0$ and let R only take values $0, -x$, and x.

Homework Problems

Problem 18.7.
There is a "one-sided" version of Chebyshev's bound for deviation above the mean:

Lemma (One-sided Chebyshev bound)**.**

$$\Pr[R - \mathrm{Ex}[R] \geq x] \leq \frac{\mathrm{Var}[R]}{x^2 + \mathrm{Var}[R]}.$$

Hint: Let $S_a ::= (R - \mathrm{Ex}[R] + a)^2$, for $0 \leq a \in \mathbb{R}$. So $R - \mathrm{Ex}[R] \geq x$ implies $S_a \geq (x+a)^2$. Apply Markov's bound to $\Pr[S_a \geq (x+a)^2]$. Choose a to minimize this last bound.

Problem 18.8.
A man has a set of n keys, one of which fits the door to his apartment. He tries the keys until he finds the correct one. Give the expectation and variance for the number of trials until success if

(a) he tries the keys at random (possibly repeating a key tried earlier)

(b) he chooses keys randomly from among those he has not yet tried.

Problems for Section 18.6

Practice Problems

Problem 18.9.
You work for the president and you want to estimate the fraction p of voters in the entire nation that will prefer him in the upcoming elections. You do this by random sampling. Specifically, you select n voters independently and randomly, ask them who they are going to vote for, and use the fraction P of those that say they will vote for the President as an estimate for p.

(a) Our theorems about sampling and distributions allow us to calculate how confident we can be that the random variable, P, takes a value near the constant, p. This calculation uses some facts about voters and the way they are chosen. Which of the following facts are true?

1. Given a particular voter, the probability of that voter preferring the President is p.
2. Given a particular voter, the probability of that voter preferring the President is 1 or 0.

3. The probability that some voter is chosen more than once in the sequence goes to zero as n increases.
4. All voters are equally likely to be selected as the third in our sequence of n choices of voters (assuming $n \geq 3$).
5. The probability that the second voter chosen will favor the President, given that the first voter chosen prefers the President, is greater than p.
6. The probability that the second voter chosen will favor the President, given that the second voter chosen is from the same state as the first, may not equal p.

(b) Suppose that according to your calculations, the following is true about your polling:

$$\Pr[|P - p| \leq 0.04] \geq 0.95.$$

You do the asking, you count how many said they will vote for the President, you divide by n, and find the fraction is 0.53. You call the President, and ... what do you say?

1. Mr. President, $p = 0.53$!
2. Mr. President, with probability at least 95 percent, p is within 0.04 of 0.53.
3. Mr. President, either p is within 0.04 of 0.53 or something very strange (5-in-100) has happened.
4. Mr. President, we can be 95% confident that p is within 0.04 of 0.53.

Class Problems

Problem 18.10.

A recent Gallup poll found that 35% of the adult population of the United States believes that the theory of evolution is "well-supported by the evidence." Gallup polled 1928 Americans selected uniformly and independently at random. Of these, 675 asserted belief in evolution, leading to Gallup's estimate that the fraction of Americans who believe in evolution is $675/1928 \approx 0.350$. Gallup claims a margin of error of 3 percentage points, that is, he claims to be confident that his estimate is within 0.03 of the actual percentage.

(a) What is the largest variance an indicator variable can have?

(b) Use the Pairwise Independent Sampling Theorem to determine a confidence level with which Gallup can make his claim.

(c) Gallup actually claims greater than 99% confidence in his estimate. How might he have arrived at this conclusion? (Just explain what quantity he could calculate; you do not need to carry out a calculation.)

(d) Accepting the accuracy of all of Gallup's polling data and calculations, can you conclude that there is a high probability that the number of adult Americans who believe in evolution is 35 ± 3 percent?

Problem 18.11.

Let $B_1, B_2, \ldots, B_n$ be mutually independent random variables with a uniform distribution on the integer interval $[1, d]$. Let D equal to the number of events $[B_i = B_j]$ that happen where $i \neq j$. It was observed in Section 16.7 (and proved in Problem 17.2) that $\Pr[B_i = B_j] = 1/d$ for $i \neq j$ and that the events $[B_i = B_j]$ are pairwise independent.

Let $E_{i,j}$ be the indicator variable for the event $[B_i = B_j]$.

(a) What are $\mathrm{Ex}[E_{i,j}]$ and $\mathrm{Var}[E_{i,j}]$ for $i \neq j$?

(b) What are $\mathrm{Ex}[D]$ and $\mathrm{Var}[D]$?

(c) In a 6.01 class of 500 students, the youngest student was born 15 years ago and the oldest 35 years ago. Let D be the number of students in the class who were born on exactly the same date. What is the probability that $4 \leq S \leq 32$? (For simplicity, assume that the distribution of birthdays is uniform over the 7305 days in the two decade interval from 35 years ago to 15 years ago.)

Problem 18.12.

A defendent in traffic court is trying to beat a speeding ticket on the grounds that—since virtually everybody speeds on the turnpike—the police have unconstitutional discretion in giving tickets to anyone they choose. (By the way, we don't recommend this defense :-).)

To support his argument, the defendent arranged to get a random sample of trips by 3,125 cars on the turnpike and found that 94% of them broke the speed limit at some point during their trip. He says that as a consequence of sampling theory (in particular, the Pairwise Independent Sampling Theorem), the court can be 95% confident that the actual percentage of all cars that were speeding is $94 \pm 4\%$.

The judge observes that the actual number of car trips on the turnpike was never considered in making this estimate. He is skeptical that, whether there were a thousand, a million, or 100,000,000 car trips on the turnpike, sampling only 3,125

is sufficient to be so confident.

Suppose you were were the defendent. How would you explain to the judge why the number of randomly selected cars that have to be checked for speeding *does not depend on the number of recorded trips*? Remember that judges are not trained to understand formulas, so you have to provide an intuitive, nonquantitative explanation.

Problem 18.13.

The proof of the Pairwise Independent Sampling Theorem 18.5.1 was given for a sequence $R_1, R_2, \ldots$ of pairwise independent random variables with the same mean and variance.

The theorem generalizes straighforwardly to sequences of pairwise independent random variables, possibly with *different* distributions, as long as all their variances are bounded by some constant.

Theorem (Generalized Pairwise Independent Sampling). *Let $X_1, X_2, \ldots$ be a sequence of pairwise independent random variables such that* $\text{Var}[X_i] \leq b$ *for some* $b \geq 0$ *and all* $i \geq 1$. *Let*

$$A_n ::= \frac{X_1 + X_2 + \cdots + X_n}{n},$$

$$\mu_n ::= \text{Ex}[A_n].$$

Then for every $\epsilon > 0$,

$$\Pr[|A_n - \mu_n| > \epsilon] \leq \frac{b}{\epsilon^2} \cdot \frac{1}{n}. \tag{18.29}$$

(a) Prove the Generalized Pairwise Independent Sampling Theorem.

(b) Conclude that the following holds:

Corollary (Generalized Weak Law of Large Numbers). *For every* $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr[|A_n - \mu_n| \leq \epsilon] = 1.$$

Problem 18.14.

An *International Journal of Epidemiology* has a policy of publishing papers about drug trial results only if the conclusion about the drug's effectiveness (or lack thereof) holds at the 95% confidence level. The editors and reviewers carefully check that any trial whose results they publish was *properly performed and accurately reported*. They are also careful to check that trials whose results they publish have been conducted independently of each other.

The editors of the Journal reason that under this policy, their readership can be confident that at most 5% of the published studies will be mistaken. Later, the editors are embarrassed —and astonished —to learn that *every one* of the 20 drug trial results they published during the year was wrong. The editors thought that because the trials were conducted independently, the probability of publishing 20 wrong results was negligible, namely, $(1/20)^{20} < 10^{-25}$.

Write a brief explanation to these befuddled editors explaining what's wrong with their reasoning and how it could be that all 20 published studies were wrong.

Exam Problems

Problem 18.15.

Yesterday, the programmers at a local company wrote a large program. To estimate the fraction, b, of lines of code in this program that are buggy, the QA team will take a small sample of lines chosen randomly and independently (so it is possible, though unlikely, that the same line of code might be chosen more than once). For each line chosen, they can run tests that determine whether that line of code is buggy, after which they will use the fraction of buggy lines in their sample as their estimate of the fraction b.

The company statistician can use estimates of a binomial distribution to calculate a value, s, for a number of lines of code to sample which ensures that with 97% confidence, the fraction of buggy lines in the sample will be within 0.006 of the actual fraction, b, of buggy lines in the program.

Mathematically, the *program* is an actual outcome that already happened. The *sample* is a random variable defined by the process for randomly choosing s lines from the program. The justification for the statistician's confidence depends on some properties of the program and how the sample of s lines of code from the program are chosen. These properties are described in some of the statements below. Indicate which of these statements are true, and explain your answers.

1. The probability that the ninth line of code in the *program* is buggy is b.

2. The probability that the ninth line of code chosen for the *sample* is defective, is b.

3. All lines of code in the program are equally likely to be the third line chosen in the *sample*.

4. Given that the first line chosen for the *sample* is buggy, the probability that the second line chosen will also be buggy is greater than b.

5. Given that the last line in the *program* is buggy, the probability that the next-to-last line in the program will also be buggy is greater than b.

6. The expectation of the indicator variable for the last line in the *sample* being buggy is b.

7. Given that the first two lines of code selected in the *sample* are the same kind of statement—they might both be assignment statements, or both be conditional statements, or both loop statements,...—the probability that the first line is buggy may be greater than b.

8. There is zero probability that all the lines in the *sample* will be different.

Problems for Section 18.7

Class Problems

Problem 18.16.
We want to store 2 billion records into a hash table that has 1 billion slots. Assuming the records are randomly and independently chosen with uniform probability of being assigned to each slot, two records are expected to be stored in each slot. Of course under a random assignment, some slots may be assigned more than two records.

(a) Show that the probability that a given slot gets assigned more than 23 records is less than e^{-36}.

Hint: For $c = 12$, the value of $c \ln c - c + 1$ is greater than 18.

(b) Show that the probability that there is a slot that gets assigned more than 23 records is less than e^{-15}. This is less than $1/3,000,000$. *Hint:* $\ln 10^9 < 21$.

Problem 18.17.
Sometimes I forget a few items when I leave the house in the morning. For example, here are probabilities that I forget various pieces of footwear:

left sock	0.2
right sock	0.1
left shoe	0.1
right shoe	0.3

(a) Let X be the number of these that I forget. What is $\mathrm{Ex}[X]$?

(b) Upper bound the probability that I forget one or more items. Make no independence assumptions.

(c) Use the Markov Inequality to upper bound the probability that I forget 3 or more items.

(d) Now suppose that I forget each item of footwear independently. Use Chebyshev's Inequality to upper bound the probability that I forget two or more items.

(e) Use Theorem 18.7.4 to lower bound the probability that I forget one or more items.

(f) I'm supposed to remember many other items, of course: clothing, watch, backpack, notebook, pencil, kleenex, ID, keys, etc. Let X be the total number of items I remember. Suppose I remember items mutually independently and $\text{Ex}[X] = 36$. Use Chernoff's Bound to give an upper bound on the probability that I remember 48 or more items.

(g) Give an upper bound on the probability that I remember 108 or more items.

Problem 18.18.
Reasoning based on the Chernoff bound goes a long way in explaining the recent subprime mortgage collapse. A bit of standard vocabulary about the mortgage market is needed:

- A **loan** is money lent to a borrower. If the borrower does not pay on the loan, the loan is said to be in **default**, and collateral is seized. In the case of mortgage loans, the borrower's home is used as collateral.
- A **bond** is a collection of loans, packaged into one entity. A bond can be divided into **tranches**, in some ordering, which tell us how to assign losses from defaults. Suppose a bond contains 1000 loans, and is divided into 10 tranches of 100 bonds each. Then, all the defaults must fill up the lowest tranche before the affect others. For example, suppose 150 defaults happened. Then, the first 100 defaults would occur in tranche 1, and the next 50 defaults would happen in tranche 2.
- The lowest tranche of a bond is called the **mezzanine tranche**.
- We can make a "super bond" of tranches called a **collateralized debt obligation (CDO)** by collecting mezzanine tranches from different bonds. This

super bond can then be itself separated into tranches, which are again ordered to indicate how to assign losses.

(a) Suppose that 1000 loans make up a bond, and the fail rate is 5% in a year. Assuming mutual independence, give an upper bound for the probability that there are one or more failures in the second-worst tranche. What is the probability that there are failures in the best Tranche?

(b) Now, do not assume that the loans are independent. Give an upper bound for the probability that there are one or more failures in the second tranche. What is an upper bound for the probability that the entire bond defaults? Show that it is a tight bound. *Hint:* Use Markov's theorem.

(c) Given this setup (and assuming mutual independence between the loans), what is the expected failure rate in the mezzanine tranche?

(d) We take the mezzanine tranches from 100 bonds and create a CDO. What is the expected number of underlying failures to hit the CDO?

(e) We divide this CDO into 10 tranches of 1000 bonds each. Assuming mutual independence, give an upper bound on the probability of one or more failures in the best tranche. The third tranche?

(f) Repeat the previous question without the assumption of mutual independence.

Homework Problems

Problem 18.19.

An infinite version of Murphy's Law is that if an infinite number of mutually independent events are expected to happen, then the probability that only finitely many happen is 0. This is known as the first *Borel-Cantelli lemma*.

(a) Let $A_0, A_1, \ldots$ be any infinite sequence of mutually independent events such that

$$\sum_{n\in\mathbb{N}} \Pr[A_n] = \infty. \tag{18.30}$$

Prove that $\Pr[\text{no } A_n \text{ occurs}] = 0$.

Hint: B_k the event that no A_n with $n \leq k$ occurs. So the event that no A_n occurs is

$$B ::= \bigcap_{k\in\mathbb{N}} B_k.$$

Apply Murphy's Law, Theorem 18.7.4, to B_k.

(b) Conclude that Pr[only finitely many A_n's occur] $= 0$.

Hint: Let C_k be the event that no A_n with $n \geq k$ occurs. So the event that only finitely many A_n's occur is

$$C ::= \bigcup_{k \in \mathbb{N}} C_k.$$

Apply part (a) to C_k.

Problems for Section 18.8

Practice Problems

Problem 18.20.
Let R be a positive integer valued random variable such that

$$\text{PDF}_R(n) = \frac{1}{cn^3},$$

where

$$c ::= \sum_{n=1}^{\infty} \frac{1}{n^3}.$$

(a) Prove that Ex$[R]$ is finite.

(b) Prove that Var$[R]$ is infinite.

Problem 18.21.
Let T be a positive integer valued random variable such that

$$\text{PDF}_T(n) = \frac{1}{an^2},$$

where

$$a ::= \sum_{n \in \mathbb{Z}^+} \frac{1}{n^2}.$$

(a) Prove that Ex$[T]$ is infinite.

(b) Prove that Ex$[\sqrt{T}]$ is finite.

Class Problems

Problem 18.22.
You have a biased coin with nonzero probability $p < 1$ of coming up heads. You toss until a head comes up, and then, as in Section 18.8, you keep tossing until you get a long run of tails, but this time let "long run" mean a run of tails that is at least $k - 10$ when your initial run was length k. Prove that the expected number of times you toss a head and start over is still infinite.

Problem 18.23.
Let $T_0, T_1, \ldots$ be a sequence of mutually independent random variables with the same distribution. Let

$$R ::= \min\{k > 0 \mid T_k > T_0\}.$$

(a) Suppose the range of the T_0 is the set $\{t_0 < t_1 < t_2 < \cdots\}$. Explain why the following Theorem implies that $\mathrm{Ex}[R] = \infty$.

Theorem 18.8.1. *If $p_0 + p_1 + p_2 + \cdots = 1$ and all $p_i \geq 0$, then the sum*

$$\Omega ::= \sum_{k \in \mathbb{N}} \frac{p_k}{p_{k+1} + p_{k+2} + \cdots}.$$

diverges.

(b) Let

$$S_k ::= p_k + p_{k+1} + \ldots,$$

and

$$a_k ::= \frac{S_k}{S_{k+1}} - 1.$$

Prove that

$$\Omega = \sum_{k \in \mathbb{N}} a_k. \tag{18.31}$$

(c) Prove that

$$\prod_{k \leq n} (a_k + 1) = \frac{1}{S_{n+1}}.$$

(d) Conclude from part (c) that

$$\prod_{k \in \mathbb{N}} (a_k + 1) = \infty. \tag{18.32}$$

(e) Conclude that $e^{\Omega} = \infty$ and hence $\Omega = \infty$.

19 Random Processes

Random Walks are used to model situations in which an object moves in a sequence of steps in randomly chosen directions. For example in Physics, three-dimensional random walks are used to model Brownian motion and gas diffusion. In this chapter we'll examine two examples of random walks. First, we'll model gambling as a simple 1-dimensional random walk —a walk along a straight line. Then we'll explain how the Google search engine used random walks through the graph of world-wide web links to determine the relative importance of websites.

19.1 Gamblers' Ruin

Suppose a gambler starts with an initial stake of n dollars and makes a sequence of \$1 bets. If he wins an individual bet, he gets his money back plus another \$1. If he loses the bet, he loses the \$1.

We can model this scenario as a random walk between integer points on the real line. The position on the line at any time corresponds to the gambler's cash-on-hand or *capital*. Walking one step to the right (left) corresponds to winning (losing) a \$1 bet and thereby increasing (decreasing) his capital by \$1. The gambler plays until either he is bankrupt or increases his capital to a target amount of T dollars. If he reaches his target, then he is called an overall *winner*, and his *intended profit*, m, will be $T - n$ dollars. If his capital reaches zero dollars before reaching his target, then we say that he is "ruined" or *goes broke*. We'll assume that the gambler has the same probability, p, of winning each individual \$1 bet and that the bets are mutually independent. We'd like to find the probability that the gambler wins.

The gambler's situation as he proceeds with his \$1 bets is illustrated in Figure 19.1. The random walk has boundaries at 0 and T. If the random walk ever reaches either of these boundary values, then it terminates.

In a *fair game*, the gambler is equally likely to win or lose each bet, that is $p = 1/2$. The corresponding random walk is called *unbiased*. The gambler is more likely to win if $p > 1/2$ and less likely to win if $p < 1/2$; these random walks are called *biased*. We want to determine the probability that the walk terminates at boundary T, namely, the probability that the gambler is a winner. We'll do this in Section 19.1.1, but before we derive the probability, let's just look at what it turns out to be.

Let's begin by supposing the coin is fair, the gambler starts with 100 dollars, and

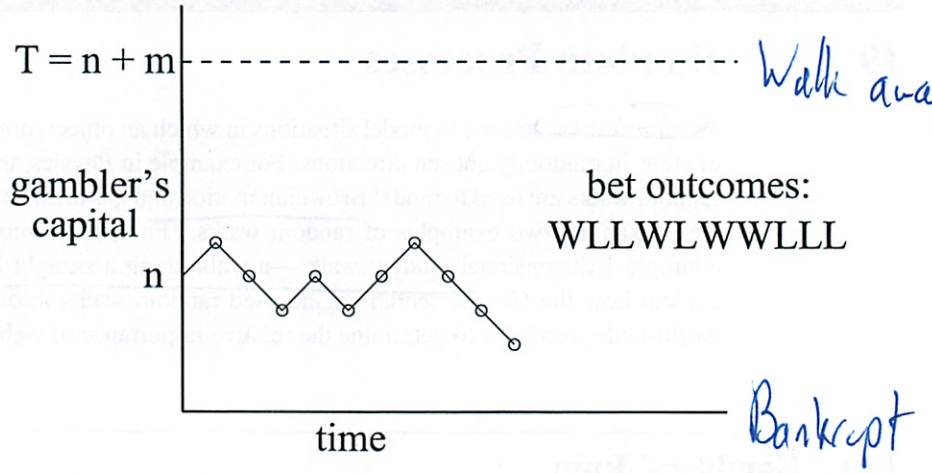


Figure 19.1 *A graph of the gambler's capital versus time for one possible sequence of bet outcomes. At each time step, the graph goes up with probability p and down with probability $1 - p$. The gambler continues betting until the graph reaches either 0 or T.*

he wants to double his money. That is, he plays until he goes broke or reaches a target of 200 dollars. Since he starts equidistant from his target and bankruptcy, it's clear by symmetry that his probability of winning in this case is 1/2.

We'll show below that starting with n dollars and aiming for a target of $T \geq n$ dollars, the probability the gambler reaches his target before going broke is n/T. For example, suppose he want to win the same \$100, but instead starts out with \$500. Now his chances are pretty good: the probability of his making the 100 dollars is $5/6$. And if he started with one million dollars still aiming to win \$100 dollars he almost certain to win: the probability is $1M/(1M + 100) > .9999$.

So in the fair game, the larger the initial stake relative to the target, the higher the probability the gambler will win, which makes some intuitive sense. But note that although the gambler now wins nearly all the time, the game is still fair. When he wins, he only wins \$100; when he loses, he loses big: \$1M. So the gambler's average win is actually zero dollars.

Another way to describe this scenario is as a game between two playets. Say Albert starts with \$500, and Eric starts with \$100. They flip a fair coin, and every time a Head appears, Albert wins \$1 from Eric, and vice versa for Tails. They play this game until one person goes bankrupt. What is the probability of Albert winning?

This problem is identical to the Gambler's Ruin problem with $n = 500$ and

$T = 100 + 500 = 600$. The probability of Albert winning is $500/600 = 5/6$, namely, the ratio of his wealth to the combined wealth. Eric's chances of winnning are $1/6$.

Now suppose instead that the gambler chooses to play roulette in an American casino, always betting \$1 on red. This game is slightly biased against the gambler: the probability of winning a single bet is $p = 18/38 \approx 0.47$. (It's the two green numbers that slightly bias the bets and give the casino an edge.) Still, the bets are almost fair, and you might expect that starting with \$500, the gambler has a reasonable chance of winning \$100 —the 5/6 probability of winning in the unbiased game surely gets reduced, but perhaps not too drastically.

Not so! The gambler's odds of winning \$100 making one dollar bets against the "slightly" unfair roulette wheel are less than 1 in 37,000. If that seems surprising, listen to this: *no matter how much money* the gambler has to start —\$5000, \$50,000, $\$5 \cdot 10^{12}$ —his odds are still less than 1 in 37,000 of winning a mere 100 dollars!

Moral: Don't play!

The theory of random walks is filled with such fascinating and counter-intuitive conclusions.

19.1.1 The Probability of Avoiding Ruin

We will determine the probability that the gambler wins using an idea of Pascal's dating back to the beginnings of the subject of probability.

Pascal viewed the walk as a two-player game between Albert and Eric as described above. Albert starts with a stack of n chips and Eric starts with a stack of $m = T - n$ chips. At each bet, Albert wins Eric's top chip with probabillity p and loses his top chip to Eric with probabillity $q ::= 1 - p$. They play this game until one person goes bankrupt.

Pascal's ingenious idea was to alter the value of the chips to make the game fair. Namely, Albert's bottom chip will be given payoff value r where $r ::= q/p$, and the successive chips *up* his stack will be worth $r^2, r^3, \ldots$ up to his top chip with payoff value r^n. Eric's top chip will be worth r^{n+1} and the successive chips *down* his stack will be worth $r^{n+2}, r^{n+3}, \ldots$ down to his bottom chip worth r^{n+m}.

Now the expected change in Albert's chip values on the first bet is

$$r^{n+1} \cdot p - r^n \cdot q = (r^n \cdot \frac{q}{p}) \cdot p - r^n \cdot q = 0,$$

so this payoff makes the bet fair. Moreover, whether Albert wins or loses the bet, the successive chip values counting up Albert's stack and then down Eric's remain $r, r^2, \ldots, r^n, \ldots, r^{n+m}$, ensuring by the same reasoning that every bet payoff remains fair. So Albert's expected payoff at the end of the game is the sum of the

expectations of his payoffs of each bet, namely 0. Here we're legitimately appealing to infinite linearity, since the payoff amounts remain bounded independent of the number of bets.

When Albert wins all of Eric's chips his total payoff gain is $\sum_{i=n+1}^{n+m} r^i$, and when he loses all his chips to Eric, he total payoff loss is $\sum_{i=1}^{n} r^i$. Letting w_n be Albert's probability of winning, we now have

$$0 = \text{Ex[Albert's payoff]} = \left(\sum_{i=n+1}^{n+m} r^i\right) \cdot w_n - \left(\sum_{i=1}^{n} r^i\right) \cdot (1 - w_n).$$

In the truly fair game when $r = 1$, we have $0 = mw_n - n(1 - w_n)$, so $w_n = n/(n+m)$, proving the claim above.

In the biased game with $r \neq 1$, we have

$$0 = r \cdot \frac{r^{n+m} - r^n}{r - 1} \cdot w_n - r \cdot \frac{r^n - 1}{r - 1} \cdot (1 - w_n).$$

Solving for w_n gives

$$w_n = \frac{r^n - 1}{r^{n+m} - 1} = \frac{r^n - 1}{r^T - 1} \tag{19.1}$$

We have now proved

Theorem 19.1.1. *In the Gambler's Ruin game with initial capital, n, target, T, and probability p of winning each individual bet,*

$$\Pr[\textit{the gambler is a winner}] = \begin{cases} \dfrac{n}{T} & \text{for } p = \dfrac{1}{2}, \\ \dfrac{r^n - 1}{r^T - 1} & \text{for } p \neq \dfrac{1}{2}, \end{cases} \tag{19.2}$$

where $r ::= q/p$.

The expression (19.1) for the probability that the Gambler wins in the biased game is a little hard to interpret. There is a simpler upper bound which is nearly tight when the gambler's starting capital is large and the game is biased *against* the gambler. Then $r > 1$, both the numerator and denominator in (19.1) are positive, and the numerator is smaller. This implies that

$$w_n < \frac{r^n}{r^T} = r^{n-T}$$

and gives:

Corollary 19.1.2. *In the Gambler's Ruin game with initial capital, n, target, T, and probability $p < 1/2$ of winning each individual bet,*

$$\Pr[\textit{the gambler is a winner}] < \left(\frac{p}{1-p}\right)^{T-n} \tag{19.3}$$

So the gambler gains his intended profit before going broke with probability at most $p/(1-p)$ raised to the intended-profit power. Notice that this upper bound does not depend on the gambler's starting capital, but only on his intended profit. This has the amazing consequence we announced above: *no matter how much money he starts with*, if he makes \$1 bets on red in roulette aiming to win \$100, the probability that he wins is less than

$$\left(\frac{18/38}{20/38}\right)^{100} = \left(\frac{9}{10}\right)^{100} < \frac{1}{37,648}.$$

The bound (19.3) is exponential in the intended profit. So, for example, doubling his intended profit will square his probability of winning. In particular, the probability that the gambler's stake goes up 200 dollars before he goes broke playing roulette is at most

$$(9/10)^{200} = ((9/10)^{100})^2 = \left(\frac{1}{37,648}\right)^2,$$

which is about 1 in 70 billion.

19.1.2 Intuition

Why is the gambler so unlikely to make money when the game is slightly biased against him? Intuitively, there are two forces at work. First, the gambler's capital has random upward and downward *swings* due to runs of good and bad luck. Second, the gambler's capital will have a steady, downward *drift*, because the negative bias means an average loss of a few cents on each \$1 bet. The situation is shown in Figure 19.2.

Our intuition is that if the gambler starts with, say, a billion dollars, then he is sure to play for a very long time, so at some point there should be a lucky, upward swing that puts him \$100 ahead. The problem is that his capital is steadily drifting downward. If the gambler does not have a lucky, upward swing early on, then he is doomed. After his capital drifts downward a few hundred dollars, he needs a huge upward swing to save himself. And such a huge swing is extremely improbable. As a rule of thumb, *drift dominates swings* in the long term.

We can quantify these drifts and swings. After k rounds for $k \leq \min(m,n)$, the number of wins by our player has a binomial distribution with parameters $p < 1/2$

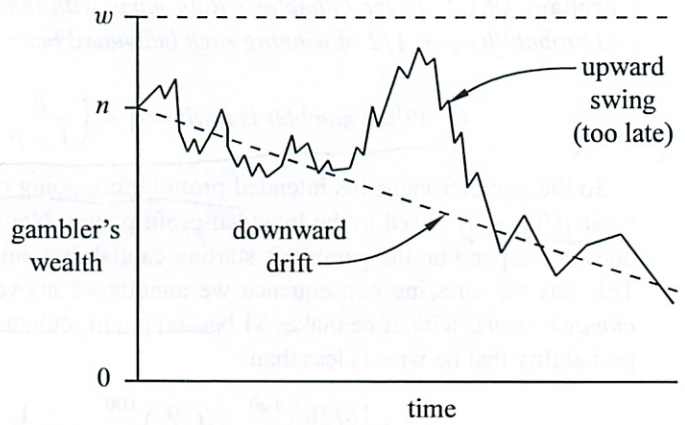


Figure 19.2 In a biased random walk, the downward drift usually dominates swings of good luck.

and k. His expected win on any single bet is $p - q = 2p - 1$ dollars, so his expected capital is $n - k(1-2p)$. Now to be a winner, his actual number of wins must exceed the expected number by $m + k(1-2p)$. But we saw before that the binomial distribution has a standard deviation of only $\sqrt{kp(1-p)}$. So for the gambler to win, he needs his number of wins to deviate by

$$\frac{m + k(1-2p)}{\sqrt{kp(1-2p)}} = \Theta(\sqrt{k})$$

times its standard deviation. In our study of binomial tails, we saw that this was extremely unlikely.

In a fair game, there is no drift; swings are the only effect. In the absence of downward drift, our earlier intuition is correct. If the gambler starts with a trillion dollars then almost certainly there will eventually be a lucky swing that puts him \$100 ahead.

19.1.3 Quit While You Are Ahead

Suppose that the gambler never quits while he is ahead. That is, he starts with $n > 0$ dollars, ignores any target T, but plays until he is flat broke. Then it turns out that if the game is not favorable, that is, $p \leq 1/2$, the gambler is sure to go broke. In particular, even in a "fair" game with $p = 1/2$ he is sure to go broke.

Lemma 19.1.3. *If the gambler starts with one or more dollars and plays a fair game until he is broke, then he will go broke with probability 1.*

Proof. If the gambler has initial capital n and goes broke in a game without reaching a target T, then he would also go broke if he were playing and ignored the target. So the probability that he will lose if he keeps playing without stopping at any target T must be at least as large as the probability that he loses when he has a target $T > n$.

But we know that in a fair game, the probability that he loses is $1 - n/T$. This number can be made arbitrarily close to 1 by choosing a sufficiently large value of T. Hence, the probability of his losing while playing without any target has a lower bound arbitrarily close to 1, which means it must in fact be 1. ■

So even if the gambler starts with a million dollars and plays a perfectly fair game, he will eventually lose it all with probability 1. But there is good news: if the game is fair, he can "expect" to play forever:

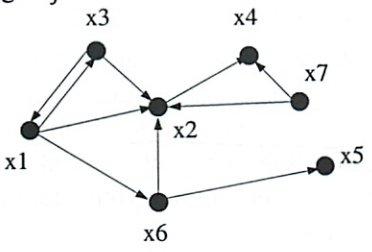
Lemma 19.1.4. *If the gambler starts with one or more dollars and plays a fair game until he goes broke, then his expected number of plays is infinite.*

A proof appears in Problem 19.1.

So even starting with just one dollar, the expected number of plays before going broke is infinite! Of course, this does not mean that the gambler is *likely* to play for long —there is even a 50% chance he will lose the very first bet and go broke right away.

19.2 Random Walks on Graphs

The hyperlink structure of the World Wide Web can be described as a digraph. The vertices are the web pages with a directed edge from vertex x to vertex y if x has a link to y. For example, in the following graph the vertices $x_1, \ldots, x_n$ correspond to web pages and $\langle x_i \to x_j \rangle$ is a directed edge when page x_i contains a hyperlink to page x_j.



The web graph is an enormous graph with many billions and probably even trillions of vertices. At first glance, this graph wouldn't seem to be very interesting.

But in 1995, two students at Stanford, Larry Page and indexBrin, Sergey Sergey Brin realized that the structure of this graph could be very useful in building a search engine. Traditional document searching programs had been around for a long time and they worked in a fairly straightforward way. Basically, you would enter some search terms and the searching program would return all documents containing those terms. A relevance score might also be returned for each document based on the frequency or position that the search terms appeared in the document. For example, if the search term appeared in the title or appeared 100 times in a document, that document would get a higher score. So if an author wanted a document to get a higher score for certain keywords, he would put the keywords in the title and make it appear in lots of places. You can even see this today with some bogus web sites.

This approach works fine if you only have a few documents that match a search term. But on the web, there are billions of documents and millions of matches to a typical search.

For example, a few years ago a search on Google for "math for computer science notes" gave 378,000 hits! How does Google decide which 10 or 20 to show first? It wouldn't be smart to pick a page that gets a high keyword score because it has "math math ... math" across the front of the document.

One way to get placed high on the list is to pay Google an advertising fees — and Google gets an enormous revenue stream from these fees. Of course an early listing is worth a fee only if an advertiser's target audience is attracted to the listing. But an audience does get attracted to Google listings because its ranking method is really good at determining the most relevant web pages. For example, Google demonstrated its accuracy in our case by giving first rank to the Fall 2002 open courseware page for 6.042 :-) . So how did Google know to pick 6.042 to be first out of 378,000?

Well back in 1995, Larry and Sergey got the idea to allow the digraph structure of the web to determine which pages are likely to be the most important.

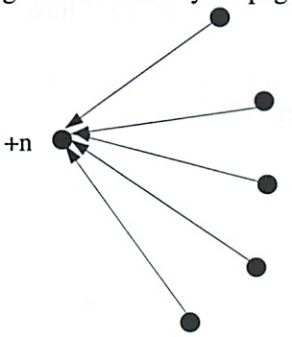
19.2.1 A First Crack at Page Rank

Looking at the web graph, any idea which vertex/page might be the best to rank 1st? Assume that all the pages match the search terms for now. Well, intuitively, we should choose x_2, since lots of other pages point to it. This leads us to their first idea: try defining the *page rank* of x to be the number of links pointing to x, that is, indegree(x). The idea is to think of web pages as voting for the most important page —the more votes, the better rank.

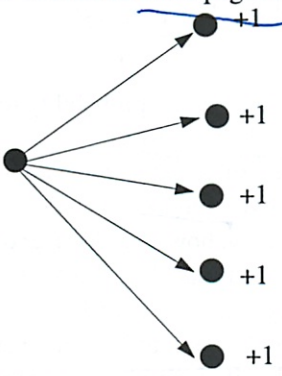
Of course, there are some problems with this idea. Suppose you wanted to have your page get a high ranking. One thing you could do is to create lots of dummy

lots of people do this

pages with links to your page.



There is another problem —a page could become unfairly influential by having lots of links to other pages it wanted to hype.



So this strategy for high ranking would amount to, "vote early, vote often," which is no good if you want to build a search engine that's worth paying fees for. So, admittedly, their original idea was not so great. It was better than nothing, but certainly not worth billions of dollars.

19.2.2 Random Walk on the Web Graph

But then Sergey and Larry thought some more and came up with a couple of improvements. Instead of just counting the indegree of a vertex, they considered the probability of being at each page after a long random walk on the web graph. In particular, they decided to model a user's web experience as following each link on a page with uniform probability. That is, they assigned each edge $x \to y$ of the web graph with a probability conditioned on being on page x:

$$\Pr[\text{follow link } \langle x \to y \rangle \mid \text{at page } x] ::= \frac{1}{\text{outdegree}(x)}.$$

The user experience is then just a random walk on the web graph.

one of the many out degrees

more out links is bad

is this still how it works?

For example, if the user is at page x, and there are three links from page x, then each link is followed with probability $1/3$.

We can also compute the probability of arriving at a particular page, y, by summing over all edges pointing to y. We thus have

$$\begin{aligned}\Pr[\text{go to } y] &= \sum_{\text{edges } \langle x \to y \rangle} \Pr\left[\text{follow link } \langle x \to y \rangle \mid \text{at page } x\right] \cdot \Pr[\text{at page } x] \\ &= \sum_{\text{edges } \langle x \to y \rangle} \frac{\Pr[\text{at } x]}{\text{outdegree}(x)} \qquad (19.4)\end{aligned}$$

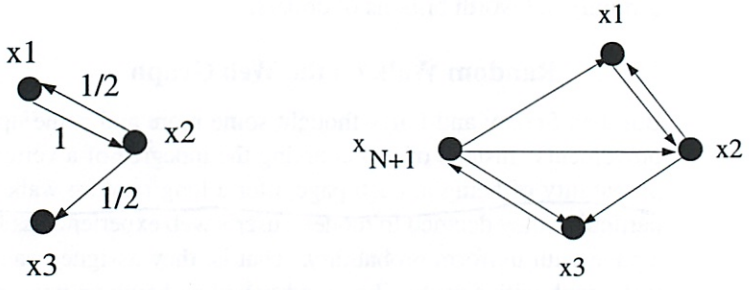
For example, in our web graph, we have

$$\Pr[\text{go to } x_4] = \frac{\Pr[\text{at } x_7]}{2} + \frac{\Pr[\text{at } x_2]}{1}.$$

One can think of this equation as x_7 sending half its probability to x_2 and the other half to x_4. The page x_2 sends all of its probability to x_4.

There's one aspect of the web graph described thus far that doesn't mesh with the user experience —some pages have no hyperlinks out. Under the current model, the user cannot escape these pages. In reality, however, the user doesn't fall off the end of the web into a void of nothingness. Instead, he restarts his web journey.

To model this aspect of the web, Sergey and Larry added a supervertex to the web graph and had every page with no hyperlinks point to it. Moreover, the supervertex points to every other vertex in the graph, allowing you to restart the walk from a random place. For example, below left is a graph and below right is the same graph after adding the supervertex x_{N+1}.



The addition of the supervertex also removes the possibility that the value $1/\text{outdegree}(x)$ might involve a division by zero.

interesting solution
I would have tried something "non distructive"
but wrong intuition

19.2.3 Stationary Distribution & Page Rank

The basic idea of page rank is just a stationary distribution over the web graph, so let's define a stationary distribution.

Suppose each vertex is assigned a probability that corresponds, intuitively, to the likelihood that a random walker is at that vertex at a randomly chosen time. We assume that the walk never leaves the vertices in the graph, so we require that

$$\sum_{\text{vertices } x} \Pr[\text{at } x] = 1. \tag{19.5}$$

Definition 19.2.1. An assignment of probabililties to vertices in a digraph is a *stationary distribution* if for all vertices x

$$\Pr[\text{at } x] = \Pr[\text{go to } x \text{ at next step}]$$

Sergey and Larry defined their page ranks to be a stationary distribution. They did this by solving the following system of linear equations: find a nonnegative number, $\text{PR}(x)$, for each vertex, x, such that

$$\text{PR}(x) = \sum_{\text{edges } \langle y \to x \rangle} \frac{\text{PR}(y)}{\text{outdegree}(y)}, \tag{19.6}$$

corresponding to the intuitive equations given in (19.4). These numbers must also satisfy the additional constraint corresponding to (19.5):

$$\sum_{\text{vertices } x} \text{PR}(x) = 1. \tag{19.7}$$

So if there are n vertices, then equations (19.6) and (19.7) provide a system of $n + 1$ linear equations in the n variables, $\text{PR}(x)$. Note that constraint (19.7) is needed because the remaining constraints (19.6) could be satisfied by letting $\text{PR}(x) ::= 0$ for all x, which is useless.

Sergey and Larry were smart fellows, and they set up their page rank algorithm so it would always have a meaningful solution. Their addition of a supervertex ensures there is always a *unique* stationary distribution. Moreover, starting from *any* vertex and taking a sufficiently long random walk on the graph, the probability of being at each page will get closer and closer to the stationary distribution. Note that general digraphs without supervertices may have neither of these properties: there may not be a unique stationary distribution, and even when there is, there may be starting points from which the probabilities of positions during a random walk do not converge to the stationary distribution. Examles of this appear in some problems below.

Now just keeping track of the digraph whose vertices are billions of web pages is a daunting task. That's why Google is building power plants. Indeed, Larry and Sergey named their system Google after the number 10^{100} —which called a "googol" —to reflect the fact that the web graph is so enormous.

Anyway, now you can see how 6.042 ranked first out of 378,000 matches. Lots of other universities used our notes and presumably have links to the 6.042 open courseware site, and the university sites themselves are legitimate, which ultimately leads to 6.042 getting a high page rank in the web graph.

Very clever idea

their prob reflects back

Problems for Section 19.1

Class Problems

Problem 19.1.

In gambler's ruin scenario, the gambler makes independent \$1 bets, where the probability of winning a bet p and of losing is $q ::= 1 - p$. The gambler keeps betting until he goes broke or reaches a target of T dollars.

Suppose $T = \infty$, that is, the gambler keeps playing until he goes broke. Let r be the probability that starting with $n > 0$ dollars, the gambler's stake ever gets reduced to $n - 1$ dollars.

(a) Explain why

$$r = q + pr^2.$$

(b) Conclude that if $p \leq 1/2$, then $r = 1$.

(c) Conclude that even in a fair game, the gambler is sure to get ruined *no matter how much money he starts with*!

Hint: If r_n is probability of ruin starting with stake n, then $r_n = r_{n+1}p + r_{n-1}q$, so

$$r_{n+1} = \frac{r_n}{p} - r_{n-1}\frac{q}{p}. \tag{19.8}$$

(d) Let t be the expected time for the gambler's stake to go down by 1 dollar. Verify that

$$t = q + p(1 + 2t).$$

Conclude that starting with a 1 dollar stake in a fair game, the gambler can expect to play forever!

Are these called Markov Chains in 6.041?

Problems for Section 19.2

Class Problems

Problem 19.2. **(a)** Find a stationary distribution for the random walk graph in Figure 19.3.

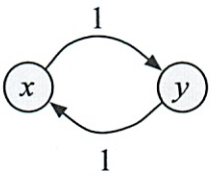


Figure 19.3

(b) If you start at node x in Figure 19.3 and take a (long) random walk, does the distribution over nodes ever get close to the stationary distribution? Explain.

(c) Find a stationary distribution for the random walk graph in Figure 19.4.

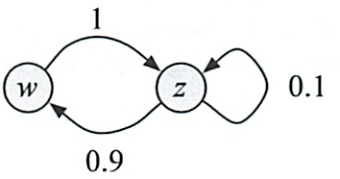


Figure 19.4

(d) If you start at node w Figure 19.4 and take a (long) random walk, does the distribution over nodes ever get close to the stationary distribution? You needn't prove anything here, just write out a few steps and see what's happening.

(e) Find a stationary distribution for the random walk graph in Figure 19.5.

(f) If you start at node b in Figure 19.5 and take a long random walk, the probability you are at node d will be close to what fraction? Explain.

Problem 19.3.
We use random walks on a digraph, G, to model the typical movement pattern of a Math for CS student right after the final exam.

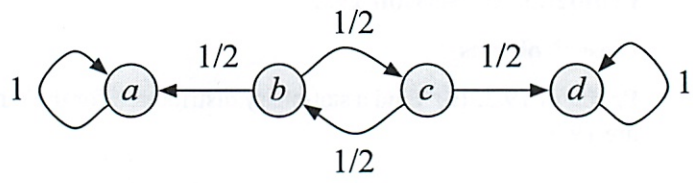


Figure 19.5

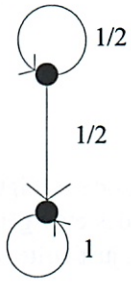
The student comes out of the final exam located on a particular node of the graph, corresponding to the exam room. What happens next is unpredictable, as the student is in a total haze. At each step of the walk, if the student is at node u at the end of the previous step, they pick one of the edges $\langle u \to v \rangle$ uniformly at random from the set of all edges directed out of u, and then walk to the node v.

Let $n ::= |V(G)|$ and define the vector $P^{(j)}$ to be

$$P^{(j)} ::= (p_1^{(j)}, \ldots, p_n^{(j)})$$

where $p_i^{(j)}$ is the probability of being at node i after j steps.

(a) We will start by looking at a simple graph. If the student starts at node 1 (the top node) in the following graph, what is $P^{(0)}$, $P^{(1)}$, $P^{(2)}$? Give a nice expression for $P^{(n)}$.



(b) Given an arbitrary graph, show how to write an expression for $p_i^{(j)}$ in terms of the $p_k^{(j-1)}$'s.

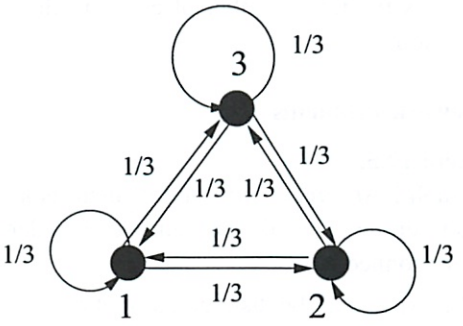
(c) Does your answer to the last part look like any other system of equations you've seen in this course?

(d) Let the *limiting distribution* vector, π, be

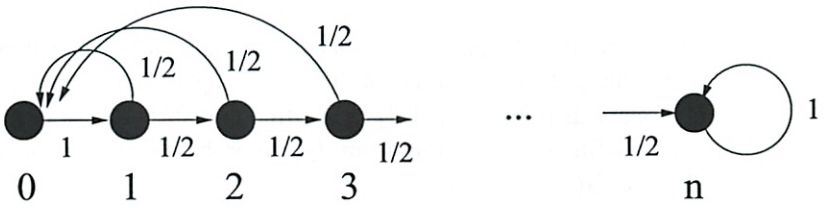
$$\lim_{k \to \infty} \frac{\sum_{i=1}^{k} P^{(i)}}{k}.$$

What is the limiting distribution of the graph from part a? Would it change if the start distribution were $P^{(0)} = (1/2, 1/2)$ or $P^{(0)} = (1/3, 2/3)$?

(e) Let's consider another directed graph. If the student starts at node 1 with probability 1/2 and node 2 with probability 1/2, what is $P^{(0)}, P^{(1)}, P^{(2)}$ in the following graph? What is the limiting distribution?



(f) Now we are ready for the real problem. In order to make it home, the poor Math for student is faced with n doors along a long hall way. Unbeknownst to him, the door that goes outside to paradise (that is, freedom from the class and more importantly, vacation!) is at the *very end*. At each step along the way, he passes by a door which he opens up and goes through with probability 1/2. Every time he does this, he gets teleported back to the exam room. Let's figure out how long it will take the poor guy to escape from the class. What is $P^{(0)}, P^{(1)}, P^{(2)}$? What is the limiting distribution?



(g) Show that the expected number, $T(n)$, of teleportations you make back to the exam room before you escape to the outside world is $2^{n-1} - 1$.

Problem 19.4.

A Google-graph is a random-walk graph such that every edge leaving any given vertex has the same probability. That is, the probability of each edge $\langle v \to w \rangle$ is $1/\text{out-degree}(v)$.

A directed graph is *symmetric* if, whenever $\langle v \to w \rangle$ is an edge, so is $\langle w \to v \rangle$. Given any finite, symmetric Google-graph, let

$$d(v) ::= \frac{\text{out-degree}(v)}{e},$$

where e is the total number of edges in the graph. Show that d is a stationary distribution.

Homework Problems

Problem 19.5.

A digraph is *strongly connected* iff there is a directed path between every pair of distinct vertices. In this problem we consider a finite random walk graph that is strongly connected.

(a) Let d_1 and d_2 be distinct distributions for the graph, and define the *maximum dilation*, γ, of d_1 over d_2 to be

$$\gamma ::= \max_{x \in V} \frac{d_1(x)}{d_2(x)}.$$

Call a vertex, x, *dilated* if $d_1(x)/d_2(x) = \gamma$. Show that there is an edge, $\langle y \to z \rangle$, from an undilated vertex y to a dilated vertex, z. *Hint:* Choose any dilated vertex, x, and consider the set, D, of dilated vertices connected to x by a directed path (going to x) that only uses dilated vertices. Explain why $D \neq V$, and then use the fact that the graph is strongly connected.

(b) Prove that the graph has *at most one* stationary distribution. (There always *is* a stationary distribution, but we're not asking you prove this.) *Hint:* Let d_1 be a stationary distribution and d_2 be a different distribution. Let z be the vertex from part (a). Show that starting from d_2, the probability of z changes at the next step. That is, $\widehat{d_2}(z) \neq d_2(z)$.

Exam Problems

Problem 19.6.

For which of the graphs in Figure 19.6 is the uniform distribution over nodes a stationary distribution? The edges are labeled with transition probabilities. Explain your reasoning.

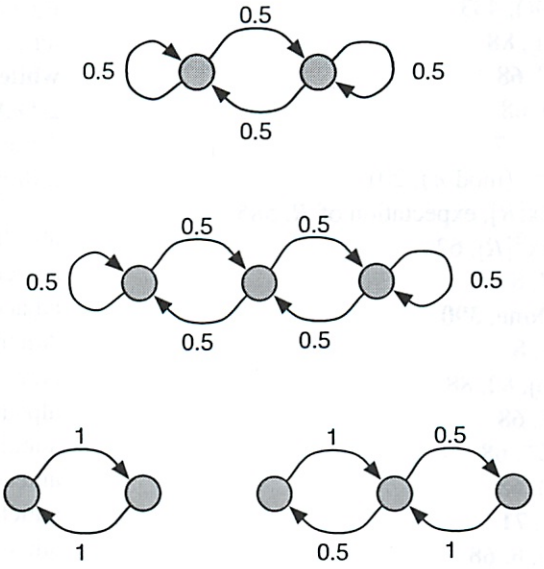


Figure 19.6 Which ones have uniform stationary distribution?

Index

Glossary of Symbols

symbol	meaning
$::=$	is defined to be
$\wedge$	and
$\vee$	or
$\longrightarrow$	implies, if ..., then $\cdots$
$\longrightarrow$	state transition
$\neg P, \overline{P}$	not P
$\longleftrightarrow$	iff, equivalent
$\oplus$	xor, exclusive-or
$\exists$	exists
$\forall$	for all
$\in$	is a member of, is in
$\subseteq$	is a (possibly =) subset of
$\subset$	is a proper (not =) subset of
$\cup$	set union
$\cap$	set intersection
$\overline{A}$	complement of set A
$-$	set difference
$\mathcal{P}(A)$	powerset of set, A
$\emptyset$	the empty set, $\{\}$
$\mathbb{Z}$	integers
$\mathbb{N}, \mathbb{Z}^{\geq 0}$	nonnegative integers
$\mathbb{Z}^{+}$	positive integers
$\mathbb{Z}^{-}$	negative integers
$\mathbb{Q}$	rational numbers
$\mathbb{R}$	real numbers
$\mathbb{C}$	complex numbers
$R(X)$	image of set X under binary relation R
R^{-1}	inverse of binary relation R
$R^{-1}(X)$	inverse image of set X under relation R

symbol	meaning
λ	the empty string/list
A^*	the finite strings over alphabet A
$\text{rev}(s)$	the reversal of string s
$s \cdot t$	concatenation of strings s, t; $\text{append}(s, t)$
$\#_c(s)$	number of occurrences of character c in string s
$m \mid n$	integer m divides integer n; m is a factor of n
gcd	greatest common divisor
(k, n)	$\{i \mid k < i < n\}$
$[k, n)$	$\{i \mid k \leq i < n\}$
$(k, n]$	$\{i \mid k < i \leq n\}$
$[k, n]$	$\{i \mid k \leq i \leq n\}$
$\langle u \to v \rangle$	directed edge from vertex u to vertex v
Id_A	identity relation on set A: $a \text{Id}_A a'$ iff $a = a'$
R^*	path relation of relation R; reflexive transitive closure of R
R^+	positive path relation of R; transitive closure of R
$\langle u — v \rangle$	undirected edge connecting vertices u $neq v$
$E(G)$	the edges of graph G
$V(G)$	the vertices of graph G
C_n	the length-n undirected cycle
L_n	the length-n line graph
K_n	the n-vertex complete graph
$L(G)$	the "left" vertices of bipartite graph G
$R(G)$	the "right" vertices of bipartite graph G
$K_{n,m}$	the complete bipartite graph with n left and m right vertices
H_n	the nth Harmonic number $\sum_{i=1}^{n} 1/i$
$\sim$	asymptotic equality
$n!$	n factorial $::= n \cdot (n-1) \cdots 2 \cdot 1$
$o()$	asymptotic notation "little oh"
$O()$	asymptotic notation "big oh"
$\Theta()$	asymptotic notation "Theta"
$\Omega()$	asymptotic notation "big Omega"
$\omega()$	asymptotic notation "little omega"
$\Pr[A]$	probability of event A
$\Pr[A \mid B]$	conditional probability of A given B
$\text{Ex}[R]$	expectation of random variable R
$\text{Ex}[R \mid A]$	conditional expectation of R given event A
$\text{Var}[R]$	variance of R
σ_R	standard deviation of R