19 Deviations

In some cases, a random variable is likely to be very close to its expected value. For example, if we flip 100 fair, mutually-independent coins, it is very likely that we will get about 50 heads. In fact, we proved in Section 17.5 that the probability of getting fewer than 25 or more than 75 heads are each less than $3 \cdot 10^{-7}$. In such cases, the mean provides a lot of information about the random variable.

In other cases, a random variable is likely to be *far* from its expected value. For example, suppose we flipped 100 fair coins that are glued together so that they all come out "heads" or they call all come out "tails." In this case, the expected value of the number of heads is still 50, but the actual number of heads is guaranteed to be far from this value—it will be 0 or 100, each with probability $1/2$.

Mathematicians have developed a variety of measures and methods to help us understand how a random variable performs in comparison to its mean. The simplest and most widely used measure is called the *variance* of the random variable. The variance is a single value associated with the random variable that is large for random variables that are likely to deviate significantly from the mean and that is small otherwise.

19.1 Variance

19.1.1 Definition and Examples

Consider the following two gambling games:

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

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

Which game would you rather play? 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:

$$\begin{aligned}
\text{Ex}[A] &= 2 \cdot \frac{2}{3} + (-1) \cdot \frac{1}{3} = 1, \\
\text{Ex}[B] &= 1002 \cdot \frac{2}{3} + (-2001) \cdot \frac{1}{3} = 1.
\end{aligned}$$

The expected payoff is the same for both games, but they are obviously very different! The stakes are a lot higher for Game B and so it is likely to deviate much farther from its mean than is Game A. This fact is captured by the notion of *variance*.

Definition 19.1.1. The *variance* Var$[R]$ of a random variable R is

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

In words, the variance of a random variable R is the expectation of the square of the amount by which R differs from its expectation.

Yikes! That's a mouthful. Try saying that 10 times in a row!

Let's look at this definition more carefully. We'll start with $R - \text{Ex}[R]$. That's the amount by which R differs from its expectation and it is obviously an important measure. Next, we square this value. More on why we do that in a moment. Finally, we take the the expected value of the square. If the square is likely to be large, then the variance will be large. If it is likely to be small, then the variance will be small. That's just the kind of statistic we are looking for. Let's see how it works out for our two gambling games.

We'll start with Game A:

$$\begin{aligned}
A - \text{Ex}[A] &= \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ -2 & \text{with probability } \frac{1}{3} \end{cases} \\
(A - \text{Ex}[A])^2 &= \begin{cases} 1 & \text{with probability } \frac{2}{3} \\ 4 & \text{with probability } \frac{1}{3} \end{cases} \\
\text{Ex}[(A - \text{Ex}[A])^2] &= 1 \cdot \frac{2}{3} + 4 \cdot \frac{1}{3} \\
\text{Var}[A] &= 2. \qquad (19.1)
\end{aligned}$$

For Game B, we have

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

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!

Why Bother Squaring?

The variance is the average *of the square* of the deviation from the mean. For this reason, variance is sometimes called the "mean squared deviation." But why bother squaring? Why not simply compute the average deviation from the mean? That is, why not define variance to be $\text{Ex}[R - \text{Ex}[R]]$?

The problem with this definition is that the positive and negative deviations from the mean exactly cancel. By linearity of expectation, we have:

$$\text{Ex}\left[R - \text{Ex}[R]\right] = \text{Ex}[R] - \text{Ex}\left[\text{Ex}[R]\right]$$

Since $\text{Ex}[R]$ is a constant, its expected value is itself. Therefore

$$\text{Ex}\left[R - \text{Ex}[R]\right] = \text{Ex}[R] - \text{Ex}[R] = 0.$$

By this definition, every random variable would have zero variance, which would not be very useful! Because of the square in the conventional definition, both positive and negative deviations from the mean increase the variance, and they do not cancel.

Of course, we could also prevent positive and negative deviations from canceling by taking an absolute value. In other words, we could compute $\text{Ex}[\,|R - \text{Ex}[R]|\,]$. But this measure doesn't have the many useful properties that variance has, and so mathematicians went with squaring.

19.1.2 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 these reasons, people often describe the deviation of a random variable using *standard deviation* instead of variance.

Definition 19.1.2. The *standard deviation* σ_R of a random variable R is the square root of the variance:

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

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 roughly canceling the square on the inside.

For example, the standard deviations for A and B are

$$\begin{aligned}\sigma_A &= \sqrt{\text{Var}[A]} = \sqrt{2} \approx 1.41,\\ \sigma_B &= \sqrt{\text{Var}[B]} = \sqrt{2,004,002} \approx 1416.\end{aligned}$$

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.

19.1.3 An Alternative Formulation

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

Lemma 19.1.3. *For any random variable R,*

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

Here we use the notation $\text{Ex}^2[R]$ as shorthand for $(\text{Ex}[R])^2$. Remember that $\text{Ex}[R^2]$ is generally not equal to $\text{Ex}^2[R]$. We know the expected value of a product is the product of the expected values for independent variables, but not in general. And R is not independent of itself unless it is constant.

Proof of Lemma 19.1.3. Let $\mu = \text{Ex}[R]$. Then

$$\begin{aligned}
\text{Var}[R] &= \text{Ex}[(R - \text{Ex}[R])^2] && \text{(Definition 19.1.1 of variance)}\\
&= \text{Ex}[(R - \mu)^2] && \text{(definition of } \mu\text{)}\\
&= \text{Ex}[R^2 - 2\mu R + \mu^2] \\
&= \text{Ex}[R^2] - 2\mu\,\text{Ex}[R] + \mu^2 && \text{(linearity of expectation)}\\
&= \text{Ex}[R^2] - 2\mu^2 + \mu^2 && \text{(definition of } \mu\text{)}\\
&= \text{Ex}[R^2] - \mu^2 \\
&= \text{Ex}[R^2] - \text{Ex}^2[R]. && \text{(definition of } \mu\text{)} \quad \blacksquare
\end{aligned}$$

For example, let's take another look at Game A from Section 19.1 where you win \$2 with probability $2/3$ and lose \$1 with probability $1/3$. Then

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

and

$$\text{Ex}[A^2] = 4 \cdot\!\leftarrow \frac{2}{3} + 1 \cdot\!\leftarrow \frac{1}{3} = 3.$$

By Lemma 19.1.3, this means that

$$\text{Var}[A] = \text{Ex}[A^2] - \text{Ex}^2[A] = 3 - 1^2 = 2,$$

confirming the result in Equation 19.1.

The alternate formulation of variance given in Lemma 19.1.3 has a cute implication:

Corollary 19.1.4. *If R is a random variable, then* $\text{Ex}[R^2] \ge \text{Ex}^2[R]$.

Proof. We defined $\text{Var}[R]$ as an average of a squared expression, so $\text{Var}[R]$ is nonnegative. Then we proved that $\text{Var}[R] =\!\leftarrow\!\text{Ex}[R^2] - \text{Ex}^2[R]$. This implies that $\text{Ex}[R^2] - \text{Ex}^2[R]$ is nonnegative. Therefore, $\text{Ex}[R^2] \ge \text{Ex}^2[R]$. ■

In words, the expectation of a square is at least the square of the expectation. The two are equal exactly when the variance is zero:

$$\text{Ex}[R^2] = \text{Ex}^2[R] \quad \text{iff} \quad \text{Ex}[R^2] - \text{Ex}^2[R] = 0 \quad \text{iff} \quad \text{Var}[R] = 0.$$

This happens precisely when

$$\Pr\big[R = \text{Ex}[R]\big] \Big(= 1; \leftarrow$$

namely, when R is a constant.[1]

[1] Technically, R could deviate from its mean on some sample points with probability 0, but we are ignoring events of probability 0 when computing expectations and variances.

19.1.4 Indicator Random Variables

Computing the variance of an indicator random variable is straightforward given Lemma 19.1.3.

Lemma 19.1.5. *Let B be an indicator random variable for which* $\Pr[B = 1] = p$. *Then*

$$\text{Var}[B] = p - p^2 = p(1 - p). \tag{19.2}$$

Proof. By Lemma 18.1.3, $\text{Ex}[B] = p$. But since B only takes values 0 and 1, $B^2 = B$. So

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

as claimed. ■

For example, let R be the number of heads when you flip a single fair coin. Then

$$\text{Var}[R] = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4} \tag{19.3}$$

and

$$\sigma_R = \sqrt{\frac{1}{4}} = \frac{1}{2}.$$

19.1.5 Mean Time to Failure

As another example, consider the mean time to failure problem, described in Section 18.1.4. If the system crashes at each step with probability p, then we already know that the mean time to failure is $1/p$. In other words, if C is the number of steps up to and including the step when the first crash occurs, then

$$\text{Ex}[C] = \frac{1}{p}.$$

What about the variance of C? To use Lemma 19.1.3, we need to compute $\text{Ex}[C^2]$. As in Section 18.1.4, we can do this by summing over all the sample points or we can use the Law of Total Expectation. The latter approach is simpler, so we’ll do that. The analysis breaks into two cases: the system crashes in the first step or it doesn’t. Hence,

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

Simplifying, we find that

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

and that

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

Using Lemma 19.1.3, we conclude that

$$\begin{aligned}
\operatorname{Var}[C] &= \operatorname{Ex}[C^2] - \operatorname{Ex}^2[C] \\
&= \frac{2-p}{p^2} - \frac{1}{p^2} \\
&= \frac{1-p}{p^2}.
\end{aligned}$$

19.1.6 Uniform Random Variables

Computing the variance of a uniform random variable is also straightforward given Lemma 19.1.3. For example, we can compute the variance of the outcome of a fair die R as follows:

$$\begin{aligned}
\operatorname{Ex}[R^2] &= \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6}, \\
\operatorname{Ex}^2[R] &= \left(3\frac{1}{2}\right)^2 = \frac{49}{4}, \\
\operatorname{Var}[R] &= \operatorname{Ex}[R^2] - \operatorname{Ex}^2[R] = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.
\end{aligned}$$

For a general uniform random variable R on $\{1, 2, 3, \ldots n\}$, the variance can be computed as follows:

$$\begin{aligned}
\text{Ex}[R] &= \frac{1}{n}(1 + 2 + \cdots + n) \\
&= \frac{1}{n} \cdot \frac{n(n+1)}{2} \\
&= \frac{n+1}{2}.
\end{aligned}$$

$$\begin{aligned}
\text{Ex}[R^2] &= \frac{1}{n}(1^2 + 2^2 + \cdots + n^2) \\
&= \frac{1}{n} \cdot \frac{(2n+1)n(n+1)}{6} \\
&= \frac{(2n+1)(n+1)}{6}.
\end{aligned}$$

$$\begin{aligned}
\text{Var}[R] &= \text{Ex}[R^2] - \text{Ex}^2[R] \\
&= \frac{(2n+1)(n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\
&= \frac{n^2-1}{12}.
\end{aligned}$$

19.1.7 Dealing with Constants

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

Theorem 19.1.6. *Let R be a random variable, and let a and b be constants. Then*

$$\text{Var}[aR + b] = a^2 \,\text{Var}[R]. \tag{19.4}$$

Proof. Beginning with Lemma 19.1.3 and repeatedly applying linearity of expectation, we have:

$$\begin{aligned}
\text{Var}[aR] &= \text{Ex}[(aR+b)^2] - \text{Ex}^2[aR+b] \\
&= \text{Ex}[a^2R^2 + 2abR + b^2] - (a\,\text{Ex}[R] + b)^2 \\
&= a^2\,\text{Ex}[R^2] + 2ab\,\text{Ex}[R] + b^2 - a^2\,\text{Ex}^2[R] - 2ab\,\text{Ex}[R] - b^2 \\
&= a^2\,\text{Ex}[R^2] - a^2\,\text{Ex}^2[R] \\
&= a^2\left(\text{Ex}[R^2] - \text{Ex}^2[R]\right) \\
&= a^2\,\text{Var}[R] \qquad \text{(by Lemma 19.1.3).}
\end{aligned}$$

■

Corollary 19.1.7.

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

19.1.8 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* random variables. In fact, *mutual* independence is not necessary: *pairwise* independence will do.

Theorem 19.1.8. *If R_1 and R_2 are independent random variables, then*

$$\operatorname{Var}[R_1 + R_2] = \operatorname{Var}[R_1] + \operatorname{Var}[R_2]. \tag{19.5}$$

Proof. As with the proof of Theorem 19.1.6, this proof uses repeated applications of Lemma 19.1.3 and Linearity of Expectation.

$$\begin{aligned}
\operatorname{Var}[R_1 + R_2] &= \operatorname{Ex}[(R_1 + R_2)^2] - \operatorname{Ex}^2[R_1 + R_2] \\
&= \operatorname{Ex}[R_1^2 + 2R_1R_2 + R^2] - (\operatorname{Ex}[R_1] + \operatorname{Ex}[R_2])^2 \\
&= \operatorname{Ex}[R_1^2] + 2\operatorname{Ex}[R_1R_2] + \operatorname{Ex}[R_2^2] \\
&\qquad - \operatorname{Ex}^2[R_1] - 2\operatorname{Ex}[R_1]\operatorname{Ex}[R_2] - \operatorname{Ex}^2[R_2] \\
&= \operatorname{Var}[R_1] + \operatorname{Var}[R_2] + 2(\operatorname{Ex}[R_1R_2] - \operatorname{Ex}[R_1]\operatorname{Ex}[R_2]) \\
&= \operatorname{Var}[R_1] + \operatorname{Var}[R_2].
\end{aligned}$$

The last step follows because

$$\operatorname{Ex}[R_1R_2] = \operatorname{Ex}[R_1]\operatorname{Ex}[R_2]$$

when R_1 and R_2 are independent. ■

Note that Theorem 19.1.8 does not necessarily hold if R_1 and R_2 are dependent since then it would generally not be true that

$$\operatorname{Ex}[R_1R_2] = \operatorname{Ex}[R_1]\operatorname{Ex}[R_2] \tag{19.6}$$

in the last step of the proof. For example, suppose that $R_1 = R_2 = R$. Then Equation 19.6 holds only if R is essentially constant.

The proof of Theorem 19.1.8 carries over straightforwardly to the sum of any finite number of variables.

Theorem 19.1.9 (Pairwise Independent Additivity of Variance)**.** *If $R_1, R_2, \ldots, R_n$ are* pairwise *independent random variables, then*

$$\operatorname{Var}[R_1 + R_2 + \cdots + R_n] = \operatorname{Var}[R_1] + \operatorname{Var}[R_2] + \cdots + \operatorname{Var}[R_n]. \tag{19.7}$$

Unfortunately, there is no product rule for computing variances, even if the random variables are mutually independent. However, we can use Theorem 19.1.9 to quickly compute the variance of a random variable with a general binomial distribution.

19.1.9 Binomial Distributions

Lemma 19.1.10 (Variance of the Binomial Distribution)**.** *If J has a binomial distribution with parameters n and p, then*

$$\text{Var}[J] = np(1-p). \tag{19.8}$$

Proof. From the definition of the binomial distribution, we can think of J as being the number of "heads" when you flip n mutually independent coins, each of which is "heads" with probability p. Thus J can be expressed as the sum of n mutually independent indicator variables J_i where

$$\Pr[J_i = 1] = p$$

for $1 \le i \le n$. From Lemma 19.1.5, we know that

$$\text{Var}[J_i] = p(1-p).$$

By Theorem 19.1.9, this means that

$$\text{Var}[J] = \sum_{i=1}^{n} \text{Var}[J_i] = np(1-p). \qquad \blacksquare$$

For example, suppose we flip n mutually independent[2] fair coins. Let R be the number of heads. Then Theorem 19.1.9 tells us that

$$\text{Var}[R] = n\left(\frac{1}{2}\right)\left(1-\frac{1}{2}\right) = \frac{n}{4}.$$

Hence,

$$\sigma_R = \frac{\sqrt{n}}{2}.$$

This value is small compared with

$$\text{Ex}[R] = \frac{n}{2},$$

which should not be surprising since we already knew from Section 17.5 that R is unlikely to stray very far from its mean.

[2] Actually, we only need to assume pairwise independence for this to be true using Theorem 19.1.9.

19.2 Markov's Theorem

The variance of a random variable gives us a rough idea of the amount by which a random variable is likely to deviate from its mean. But it does not directly give us specific bounds on the probability that the deviation exceeds a specified threshold. To obtain such specific bounds, we'll need to work a little harder.

In this section, we derive a famous result known as Markov's Theorem that gives an upper bound on the probability that a random variable exceeds a specified threshold. In the next section, we give a similar but stronger result known as Chebyshev's Theorem. The difference between these results is that Markov's Theorem depends only on the mean of the random variable, whereas Chebyshev's Theorem makes use of the mean *and* the variance. Basically, the more you know about a random variable, the better bounds you can derive on the probability that it deviates from its mean.

19.2.1 A Motivating Example

The idea behind Markov's Theorem can be explained with a simple example involving *intelligence quotients*, or IQs. This quantity was devised so that the average IQ measurement would be 100. 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 at least 300, then the average IQ 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 since no IQ over 200 has ever been recorded.

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 over 150 have certainly been recorded, although a much smaller fraction than 2/3 of the population actually has an IQ that high.

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.

Markov's Theorem characterizes the bounds that can be achieved with this kind of analysis

19.2.2 The Theorem

Theorem 19.2.1 (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 $x > 0$

$$\begin{aligned}
\mathrm{Ex}[R] &= \sum_{y \in \mathrm{range}(R)} y \Pr[R = y] \\
&\geq \sum_{\substack{y \geq x, \\ y \in \mathrm{range}(R)}} y \Pr[R = y] && \text{(because } R \geq 0\text{)} \\
&\geq \sum_{\substack{y \geq x, \\ y \in \mathrm{range}(R)}} x \Pr[R = y] \\
&= x \sum_{\substack{y \geq x, \\ y \in \mathrm{range}(R)}} \Pr[R = y] \\
&= x \Pr[R \geq x]. && (19.9)
\end{aligned}$$

Hence,

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

Corollary 19.2.2. *If R is a nonnegative random variable, then for all $c \geq 1$,*

$$\Pr\big[R \geq c \cdot \mathrm{Ex}[R]\big] \leq \frac{1}{c}. \tag{19.10}$$

Proof. Set $x = c\,\mathrm{Ex}[R]$ in Theorem 19.2.1. ■

As an example, suppose we flip 100 fair coins and use Markov's Theorem to compute the probability of getting all heads:

$$\Pr[\text{heads} \geq 100] \leq \frac{\mathrm{Ex}[\text{heads}]}{100} = \frac{50}{100} = \frac{1}{2}.$$

If the coins are mutually independent, then the actual probability of getting all heads is a minuscule 1 in 2^{100}. In this case, Markov's Theorem looks very weak. However, in applying Markov's Theorem, we made no independence assumptions. In fact, if all the coins are glued together, then probability of throwing all heads is exactly $1/2$. In this nasty case, Markov's Theorem is actually tight!

The Chinese Appetizer Problem

Suppose that n people are seated at a circular table and that each person has an appetizer in front of them on a rotating Chinese banquet tray. Just as everyone is about to dig in, some joker spins the tray so that each person receives a random appetizer. We are interested in the number of people R that get their same appetizer as before, assuming that the n appetizers are all different.

Each person gets their original appetizer with probability $1/n$. Hence, by Linearity of Expectation,

$$\text{Ex}[R] = n \cdot \frac{1}{n} = 1.$$

What is the probability that all n people get their original appetizer back? Markov's Theorem tells us that

$$\Pr[R = n] = \Pr[R \ge n] \le \frac{\text{Ex}[R]}{n} = \overset{1}{\underset{n}{\leftarrow}}.$$

In fact, this bound is tight sine everyone gets their original appetizers back if and only if the rotating tray returns to its original configuration, which happens with probability $1/n$.

The Chinese Appetizer problem is similar to the Hat Check problem that we studied in Section 18.3.2, except that no distribution was specified in the Hat Check problem—we were told only that each person gets their correct hat back with probability $1/n$. If the hats are scrambled according to uniformly random permutations, then the probability that everyone gets the right hat back is $1/n!$, which is much less than the $1/n$ upper bound given by Markov's Theorem. So, in this case, the bound given by Markov's Theorem is not close to the actual probability.

What is the probability that at least two people get their right hats back? Markov's Theorem tells us that

$$\Pr[R \ge 2] \le \frac{\text{Ex}[R]}{2} = \overset{1}{\underset{2}{\leftarrow}}.$$

In this case, Markov's Theorem is not too far off from the right answer if the hats are distributed according to a random permutation[3] but it is not very close to the correct answer $1/n$ for the case when the hats are distributed as in the Chinese Appetizer problem.

Why R Must be Nonnegative

Remember that Markov's Theorem applies only to nonnegative random variables! Indeed, the theorem is false if this restriction is removed. For example, let R be -10

[3] Proving this requires some effort.

with probability $1/2$ and 10 with probability $1/2$. Then

$$\text{Ex}[R] = -10 \cdot \frac{1}{2} + 10 \cdot \frac{1}{2} = 0.$$

Suppose that we now tried to compute $\Pr[R \geq 5]$ using Markov's Theorem:

$$\Pr[R \geq 5] \leq\leftarrow \frac{\text{Ex}[R]}{5} = \frac{0}{5} = 0.$$

This is the wrong answer! Obviously, R is at least 5 with probability $1/2$.

On the other hand, we can still apply Markov's Theorem indirectly to derive a bound on the probability that an arbitrary variable like R is 5 or more. For example, given any random variable, R with expectation 0 and values ≥ -10, we can conclude that $\Pr[R \geq 5] \leq 2/3$. To prove this fact, we define $T ::= R + 10$. Then T is a nonnegative random variable with expectation $\text{Ex}[R + 10] = \text{Ex}[R] + 10 = 10$, so Markov's Theorem applies and tells us that $\Pr[T \geq 15] \leq 10/15 = 2/3$. But $T \geq 15$ iff $R \geq 5$, so $\Pr[R \geq 5] \leq 2/3$, as claimed.

19.2.3 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. That's because if R is the IQ of a random MIT student, then

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

But let's also suppose that no MIT student has an IQ less than 100 (which may be true). 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:

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

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

More generally, we can get better bounds applying Markov's Theorem to $R - l$ instead of R for any lower bound l on R, even when l is negative.

Theorem 19.2.3. *Let R be a random variable for which $R \geq l$ for some $l \in \mathbb{R}$. Then for all $x \geq l$,*

$$\Pr[R \geq x] \leq\leftarrow \frac{\text{Ex}[R] - l}{x - l}.$$

Proof. Define

$$T ::= R - l.$$

Then T is a nonnegative random variable with mean

$$\text{Ex}[T] = \text{Ex}[R - l] = \text{Ex}[R] - l.$$

Hence, Markov's Theorem implies that

$$\begin{aligned} \Pr[T \geq x - l] &\leq\leftarrow \frac{\text{Ex}[T]}{x - l} \\ &=\leftarrow \frac{\text{Ex}[R] - l}{x - l}. \end{aligned}$$

The result then follows from the fact that

$$\begin{aligned} \Pr[R \geq x] &= \Pr[R - l \geq x - l] \\ &= \Pr[T \geq x - l]. \end{aligned}$$ ■

19.2.4 Deviations Below the Mean

Markov's Theorem says that a random variable is unlikely to greatly exceed the mean. Correspondingly, there is a variation of Markov's Theorem that says a random variable is unlikely to be much smaller than its mean.

Theorem 19.2.4. *Let $u \in \mathbb{R}$ and let R be a random variable such that $R \leq u$. Then for all $x < u$,*

$$\Pr[R \leq x] \leq\leftarrow \frac{u - \text{Ex}[R]}{u - x}.$$

Proof. The proof is similar to that of Theorem 19.2.3. Define

$$S ::= u - R.$$

Then S is a nonnegative random variable with mean

$$\text{Ex}[S] = \text{Ex}[u - R] = u - \text{Ex}[R].$$

Hence, Markov's Theorem implies that

$$\Pr[S \geq u - x] \leq\leftarrow \frac{\text{Ex}[S]}{u - x} =\leftarrow \frac{u - \text{Ex}[R]}{u - x}.$$

The result then follows from the fact that

$$\Pr[R \leq x] = \Pr[u - S \leq x] = \Pr[S \geq u - x].$$ ■

For example, suppose that the class average on a midterm was 75/100. What fraction of the class scored below 50?

There is not enough information here to answer the question exactly, but Theorem 19.2.4 gives an upper bound. Let R be the score of a random student. Since 100 is the highest possible score, we can set $u = 100$ to meet the condition in the theorem that $R \le u$. Applying Theorem 19.2.4, we find:

$$\Pr[R \le 50] \le \frac{100 - 75}{100 - 50} = \frac{1}{2}.$$

That is, at most half of the class scored 50 or worse. This makes sense; if more than half of the class scored 50 or worse, then the class average could not be 75, even if everyone else scored 100. As with Markov’s Theorem, Theorem 19.2.4 often gives weak results. In fact, based on the data given, the *entire* class could have scored *above* 50.

19.2.5 Using Markov’s Theorem to Analyze Non-Random Events

In the previous example, we used a theorem about a random variable to conclude facts about non-random data. For example, we concluded that if the average score on a test is 75, then at most $1/2$ the class scored 50 or worse. There is no randomness in this problem, so how can we apply Theorem 19.2.4 to reach this conclusion?

The explanation is not difficult. For any set of scores $S = \{s_1, s_2, \dots, s_n\}$, we introduce a random variable R such that

$$\Pr[R = s_i] = \frac{(\text{\# of students with score } s_i)}{n}.$$

We then use Theorem 19.2.4 to conclude that $\Pr[R \le 50] \le 1/2$. To see why this means (with certainty) that at most $1/2$ of the students scored 50 or less, we observe that

$$\begin{aligned}
\Pr[R \le 50] &= \sum_{s_i \le 50} \Pr[R = s_i] \\
&= \sum_{s_i \le 50} \frac{(\text{\# of students with score } s_i)}{n} \\
&= \frac{1}{n}(\text{\# of students with score 50 or less}).
\end{aligned}$$

So, if $\Pr[R \le 50] \le 1/2$, then the number of students with score 50 or less is at most $n/2$.

19.3 Chebyshev's Theorem

As we have just seen, Markov's Theorem can be extended by applying it to functions of a random variable R such as $R - l$ and $u - R$. Even stronger results can be obtained by applying Markov's Theorem to powers of R.

Lemma 19.3.1. *For any random variable R, $\alpha \in \mathbb{R}^+$, and $x > 0$,*

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

Proof. The event $|R| \geq x$ is the same as the event $|R|^\alpha \geq x^\alpha$. Since $|R|^\alpha$ is nonnegative, the result follows immediately from Markov's Theorem. ■

Similarly,

$$\Pr[|R - \operatorname{Ex}[R]| \geq x] \leq \frac{\operatorname{Ex}[(R - \operatorname{Ex}[R])^\alpha]}{x^\alpha}. \qquad (19.11)$$

The restatement of Equation 19.11 for $\alpha = 2$ is known as *Chebyshev's Theorem*.

Theorem 19.3.2 (Chebyshev)**.** *Let R be a random variable and $x \in \mathbb{R}^+$. Then*

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

Proof. Define

$$T ::= R - \operatorname{Ex}[R].$$

Then

$$\begin{aligned}
\Pr\big[\,|R - \operatorname{Ex}[R]| \geq x\big] &= \Pr[|T| \geq x] \\
&= \Pr[T^2 \geq x^2] \\
&\leq \frac{\operatorname{Ex}[T^2]}{x^2} && \text{(by Markov's Theorem)} \\
&= \frac{\operatorname{Ex}[(R - \operatorname{Ex}[R])^2]}{x^2} \\
&= \frac{\operatorname{Var}[R]}{x^2}. && \text{(by Definition 19.1.1)}
\end{aligned}$$

■

Corollary 19.3.3. *Let R be a random variable, and let c be a positive real number.*

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

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

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

As an 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 19.2.1 gives a coarse bound, namely,

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

Now we apply Corollary 19.3.3 to the same problem:

$$\Pr[R \geq 300] \leq \Pr\big[\,|R - 100| \geq 20\sigma_R\big] \leq \frac{1}{400}. \tag{19.12}$$

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 standard deviation of R, than we could get knowing only the expectation.

More generally, Corollary 19.3.3 tells us that a random variable is never likely to stray by more than a few standard deviations from its mean. For example, plugging $c = 3$ into Corollary 19.3.3, we find that the probability that a random variable strays from the mean by more than 3σ is at most $1/9$.

This fact has a nice pictorial characterization for pdf's with a "bell-curve" shape; namely, the width of the bell is $O(\sigma)$, as shown in Figure 19.1.

19.3.1 Bounds on One-Sided Errors

Corollary 19.3.3 gives bounds on the probability of deviating from the mean in *either* direction. If you only care about deviations in one direction, as was the case in the IQ example, then slightly better bounds can be obtained.

Theorem 19.3.4. *For any random variable R and any $c > 0$,*

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

and

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

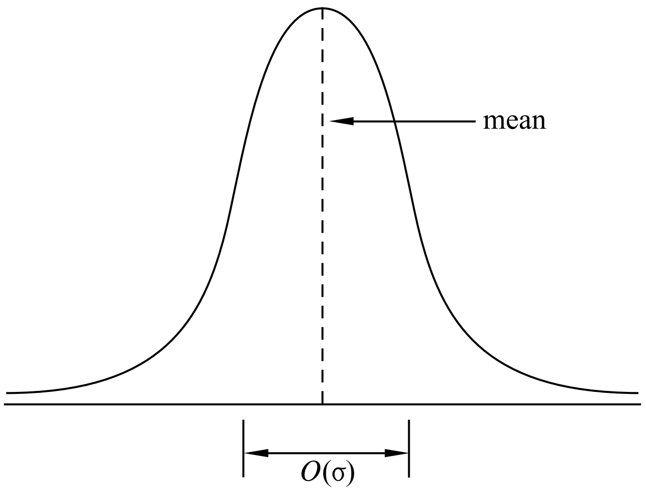

Figure 19.1 If the pdf of a random variable is "bell-shaped," then the width of the bell is $O(\sigma)$.

The proof of Theorem 19.3.4 is trickier than the proof of Chebyshev's Theorem and we will not give the details here. Nor will we prove the fact that the bounds in Theorem 19.3.4 are the best bounds that you can obtain if you know only the mean and standard deviation of the random variable R.

Returning to the IQ example, Theorem 19.3.4 tells us that

$$\Pr[R \geq 300] \leq \Pr[R - 100 \geq 20\sigma_R] \leq \xleftarrow[401]{1},$$

which is a *very slight* improvement over Equation 19.12.

As another example, suppose we give an exam. What fraction of the class can score more than 2 standard deviations from the average? If R is the score of a random student, then

$$\Pr[|R - \text{Ex}[R]| \geq 2\sigma_R] \leq \xleftarrow[4]{1}.$$

For one-sided error, the fraction that could be 2 standard deviations or more above the average is at most

$$\frac{1}{2^2 + 1} = \frac{1}{5}.$$

This results holds no matter what the test scores are, and is again a deterministic fact derived using probabilistic tools.

19.4 Bounds for 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 $T_1, T_2, \dots,$ T_n where $0 \le T_i \le 1$. A random variable with a binomial distribution is just one of many examples of such a T. Here is another.

19.4.1 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.

19.4.2 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 19.4.3.

Theorem 19.4.1 (Chernoff Bound)**.** *Let $T_1, \dots T_n$ 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 $c \ge 1$,*

$$\Pr[T \ge c \operatorname{Ex}[T]] \le e^{-k \operatorname{Ex}[T]} \tag{19.13}$$

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!

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

More Examples

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.

What is the probability that the number of heads that come up in 1000 independent tosses of a fair coin exceeds the expectation by 20% or more? Let T_i be an indicator variable for the event that the i-th 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 fact, this example is similar to many applications of the Chernoff bound in that every T_i is *either* 0 or 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 $\Pr[T \geq c \operatorname{Ex}[T]]$ where c = 1.2. To that end, we compute k as defined in the theorem:

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

Plugging this value into the Chernoff bound gives:

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

So the probability of getting 20% or more extra heads on 1000 coins is less than 1 in 10,000.[5]

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 pretty darn small.

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 seem insignificant.

[5] Since we are analyzing a binomial distribution here, we can get somewhat better bounds using the methods from Section 17.5, but it is much easier to use the Chernoff bounds, and they provide results that are nearly as good.

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!

Pick-4

Pick-4 is a lottery game where you 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. And if 10 million people play, then the expected number of winners is 1000. The lottery operator's nightmare is that the number of winners is much greater; say, 2000 or more. What is the probability that will happen?

Let T_i be an indicator for the event that the i-th player wins. Then $T = T_1 + \leftarrow \cdots + T_n$ is the total number of winners. If we assume[6] 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:

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

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:

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

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;

[6] As we noted in Chapter 18, 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.

that is, assigned to do more than 10 minutes of work in a 10-minute interval.

To begin, let’s find the probability that the first server is overloaded. Let T_i be the number of seconds that the first server spends on the i-th task. So T_i is zero if the task is assigned to another machine, and otherwise T_i is the length of the task. Then $T = \sum_{i=1}^{n} T_i$ is the total length of tasks assigned to the server, where $n = 24{,}000$. We need an upper bound on $\Pr[T \geq 600]$; that is, the probability that the first server is assigned more than 600 seconds (or, equivalently, 10 minutes) of work.

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:

$$\begin{aligned} \text{Ex}[T] &= \frac{24{,}000 \text{ tasks} \cdot 1/4 \text{ second per task}}{m \text{ machines}} \\ &= 6000/m \text{ seconds.} \end{aligned} \tag{19.14}$$

For example, if there are $m = 10$ machines, then the expected load on the first server is 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:

$$\begin{aligned} \Pr\left[T \geq 600\right] &= \Pr\left[T \geq \frac{m}{10}\,\text{Ex}[T]\right] \\ &= \Pr\left[T \geq c\,\text{Ex}[T]\right] \\ &\leq e^{-(c\ln(c)-c+1)\cdot 6000/m}, \end{aligned}$$

where $c = m/10$. The first equality follows from Equation 19.14.

The probability that *some* server is overloaded is at most m times the probability that the first server is overloaded by the Sum Rule in Section 14.4.2. 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:

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

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!

19.4.3 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![7]

Here is the theorem again, for reference:

Theorem 19.4.2 (Chernoff Bound)**.** *Let $T_1, \dots, T_n$ be mutually independent random variables such that $0 \leq T_i \leq 1$ for all i. Let $T = T_1 + \cdots + T_n$. Then for all $c \geq 1$,*

$$\Pr[T \geq c \operatorname{Ex}[T]] \leq e^{-k \operatorname{Ex}[T]} \tag{19.13}$$

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

Proof. 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:

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

In the third step, the numerator is rewritten using the inequality

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

which is proved below in Lemma 19.4.3. The final step is simplification, using the fact that c^c is equal to $e^{c \ln(c)}$. ■

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

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 19.4.1.

Lemma 19.4.3.

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

Proof.

$$\begin{aligned}
\operatorname{Ex}[c^T] &= \operatorname{Ex}[c^{T_1+\cdots+T_n}] \\
&= \operatorname{Ex}[c^{T_1}\cdots c^{T_n}] \\
&= \operatorname{Ex}[c^{T_1}]\cdots\operatorname{Ex}[c^{T_n}] \\
&\le e^{(c-1)\operatorname{Ex}[T_1]}\cdots e^{(c-1)\operatorname{Ex}[T_n]} \\
&= e^{(c-1)(\operatorname{Ex}[T_1]+\cdots+\operatorname{Ex}[T_n])} \\
&= e^{(c-1)\operatorname{Ex}[T_1+\cdots+T_n]} \\
&= e^{(c-1)\operatorname{Ex}[T]}.
\end{aligned}$$

The first step uses the definition of T, and the second is just algebra. The third step uses the fact that the expectation of a product of independent random variables is the product of the expectations. This is where the requirement that the T_i be independent is used. Then we bound each term using the inequality

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

which is proved in Lemma 19.4.4. The last steps are simplifications using algebra and linearity of expectation. ■

Lemma 19.4.4.

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

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].

$$\begin{aligned}
\text{Ex}[c^{T_i}] &= \sum_v c^v \Pr[T_i = v] \\
&\le \sum_v (1 + (c-1)v) \Pr[T_i = v] \\
&= \sum_v \Pr[T_i = v] + (c-1)v \Pr[T_i = v] \\
&= \sum_v \Pr[T_i = v] + \sum_v (c-1)v \Pr[T_i = v] \\
&= 1 + (c-1) \sum_v v \Pr[T_i = v] \\
&= 1 + (c-1) \text{Ex}[T_i] \\
&\le e^{(c-1)\text{Ex}[T_i]}.
\end{aligned}$$

The first step uses the definition of expectation. 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]. We then multiply out inside the summation and split into two sums. The first sum adds the probabilities of all possible outcomes, so it is equal to 1. After pulling the constant $c - 1$ out of the second sum, we're left with the definition of $\text{Ex}[T_i]$. The final step uses the standard inequality $1 + z \le e^z$, which holds for all $z > 0$. ■

19.5 Mutually Independent Events

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

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

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

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

to be the number of events that occur.

We know from Linearity of Expectation that

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

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

By Theorem 19.1.9, we also know that

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

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 $c > 1$,

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

Chebyshev’s Theorem gives us the stronger result that

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

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

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

In this case, the probability of exceeding the mean by $c\,\text{Ex}[T]$ 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 $\text{Ex}[T]$ is also exponentially small.

19.5.1 Murphy's Law

Suppose we want to know the probability that at least 1 event occurs. If $\text{Ex}[T] < 1$, then Markov's Theorem tells us that

$$\Pr[T \geq 1] \leq \text{Ex}[T].$$

On the other hand, if $\text{Ex}[T] \geq 1$, then we can obtain a lower bound on $\Pr[T \geq 1]$ using a result that we call Murphy's Law[8].

Theorem 19.5.1 (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

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

Proof.

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

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 19.5.1 proves that some of them surely will go wrong.

[8] 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 19.5.1 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.

19.5.2 Another Magic Trick

Theorem 19.5.1 is surprisingly powerful. In fact, it is so powerful that it can enable us to read your mind. Here's how.

You choose a secret number n from 1 to 9. Then we randomly shuffle an ordinary deck of 52 cards and display the cards one at a time. You watch as we reveal the cards and when we reveal the nth card, that card becomes your *secret card*. If the card is an Ace, a 10, or a face card, then you assign that card a *value* of 1. Otherwise, you assign that card a value that is its number. For example, the $J\heartsuit$ gets assigned a value $v_1 = 1$ and the $4\diamondsuit$ gets assigned a value $v_1 = 4$. You do all of this in your mind so that we can't tell when the nth card shows up.

We keep revealing the cards, and when the $(n + v_1)$th card shows up, that card becomes your *new* secret card. You compute its value v_2 using the same scheme as for v_1. For example, if your new secret card is the $10\clubsuit$, then $v_2 = \leftarrow 1$. The $(n + v_1 + v_2)$th card will then become your next secret card, and so forth.

We proceed in this fashion until all 52 cards have been revealed, whereupon we read your mind by predicting your last secret card! How is this possible?

For the purposes of illustration, suppose that your secret number was $n = 3$ and the deck consisted of the 11 cards:

$$3\diamondsuit\leftarrow 5\spadesuit\leftarrow 2\diamondsuit\leftarrow 3\clubsuit\leftarrow 10\clubsuit\leftarrow Q\diamondsuit\leftarrow 3\heartsuit\leftarrow 7\spadesuit\leftarrow 6\clubsuit\leftarrow 4\diamondsuit\leftarrow 2\heartsuit.$$

Then your secret cards would be

$$2\diamondsuit,\ 10\clubsuit,\ Q\diamondsuit,\ 3\heartsuit,\ 4\diamondsuit\leftarrow$$

since $v_1 = 2,\ v_2 = 1,\ v_3 = 1,\ v_4 = 3$, and $v_5 = 4$. In this example, your last secret card is the $4\diamondsuit$.

To make the trick work, we follow the same rules as you, except that we start with $n = 1$. With the 11-card deck shown above, our secret cards would be

$$3\diamondsuit,\ 3\clubsuit,\ 3\heartsuit,\ 4\diamondsuit.$$

We have the same last secret card as you do! That is *not* a coincidence. In fact, this is how we predict your last card—we just guess that it is the same as our last card. And, we will be right with probability greater than 90%.

To see why the trick is likely to work, you need to notice that if we ever share a secret card, then we will surely have the same *last* secret card. That's because we will perform exactly the same steps as the cards are revealed.

Each time we get a new secret card, there is always a chance that it was one of your secret cards. For any given step, the chance of a match is small but we get a lot of chances. In fact, the number of chances will typically outweigh the inverse of the probability of a match on any given step and so, at least informally, Murphy's Law suggests that we are likely to eventually get a match, whereupon we can read your mind.

The details of the proof are complicated and we will not present them here. One of the main complications is that when you are revealing cards from a deck without replacement, the probability of getting a match on a given step is conditional based on the cards that have already been revealed.

19.5.3 The Subprime Mortgage Disaster

Throughout the last few chapters, we have seen many examples where powerful conclusions can be drawn about a collection of events if the events are independent. Of course, such conclusions are totally invalid if the events have dependencies. Unforeseen dependencies can result in disaster in practice. For example, misguided assumptions about the independence of loans (combined with a large amount of greed) triggered the global financial meltdown in 2008–2009.

In what follows, we'll explain some of what went wrong. You may notice that we have changed the names of the key participants. That is not to protect the innocent, since innocents are few and far between in this sordid tale. Rather, we changed the names to protect ourselves.[9] In fact, just to be on the safe side, we'll forget about what really happened here on Earth and instead tell you a fairy tale that took place in a land far, far away.

The central players in our story are the major Wall Street firms, of which Golden Scoundrels (commonly referred to as "Golden") is the biggest and most aggressive. Firms such as Golden ostensibly exist to make markets; they purport to create an open and orderly market in which buyers and sellers can be brought together and through which capitalism can flourish. It all sounds good, but the fees that can be had from facilitating transactions in a truly open and orderly market are often just not enough to satisfy the ever-increasing need to make more. So the employees at

[9] For a much more detailed accounting of these events (and one that does name names), you may enjoy reading *The Big Short* by Michael Lewis.

such firms are always trying to figure out a way to create new opportunities to make even more money.

One day, they came up with a whopper. Suppose they bought a collection of 1000 (say) subprime mortgage loans from all around the country and packaged them up into a single entity called a *bond*. A *mortgage loan* is a loan to a homeowner using the house as collateral; if the homeowner stops paying on the loan (in which case the loan is said to be in *default*), then the owner of the loan takes ownership of the house. A mortgage loan is classified as *subprime* if the homeowner does not have a very good credit history. Subprime loans are considered to be more risky than *prime* loans since they are more likely to default. Defaults are bad for everyone; the homeowner loses the home and the loan owner gets stuck trying to sell the house, which can take years and often results in very high losses.

Of course, a bond consisting of 1000 subprime loans doesn’t sound very appealing to investors, so to dress it up, Golden sells the bond in *tranches*. The idea behind the tranches is to provide a way to assign losses from defaults. In a typical scenario, there would be 10 tranches and they are prioritized from 1 to 10. The defaults are assessed against the lowest tranches first. For example, suppose that there were 150 defaults in the collection of 1000 loans (an impossibly high number of defaults according to Golden). Then the lowest tranche would absorb the first 100 defaults (effectively wiping them out since all 100 of “their” loans would be in default) and the second-lowest tranche would be assigned the next 50 defaults, (wiping out half of their investment). The remaining 8 tranches would be doing great—none of “their” loans would be in default.

Because they are taking on more risk, the lower tranches would get more of the interest payments. The top tranche would get the lowest rate of return and would also be the safest. The lowest tranche would get the most interest, but also be the most exposed.

But how much should you pay for a tranche? Suppose the probability that any given loan defaults in a year is 1%. In other words, suppose you expect 10 of the 1000 loans to default in each year. If the defaults are independent, then we can use the Chernoff bound to conclude that the chances of more than 100 defaults (10%) in the 1000-loan collection is exceedingly tiny. This means that every tranche but the lowest is essentially risk-free. That is excellent news for Golden since they can buy 1000 cheap[10] subprime loans and then sell the top 9 tranches at premium rates, thereby making a large and instant profit on 900 of the 1000 loans. It is like turning a bunch of junk into a bunch of gold with a little junk left over.

There remains the problem of the lowest tranche, which is expected to have 10 defaults in a pool of 100 loans for a default rate of 10%. This isn’t so good

[10] They are *subprime* loans after all.

so the first thing to do is to give the tranche a better sounding name than "lowest tranche." "Mezzanine" tranche sounds much less ominous and so that is what they used.

By the Chernoff bound, the default rate in the Mezzanine tranche is very unlikely to be much greater than 10%, and so the risk of owning this tranche can be addressed in part by increasing the interest payments for the tranche by 10%. But Golden had an even better idea (whopper number two)—rather than pay the extra 10%, why not collect together a bunch of mezzanine tranches from a bunch of bonds and then package them together into a "super bond" and then create tranches in the super-bond? The technical name for such a super bond is a *collateralized debt obligation* or CDO. This way, 90% of the mezzanine tranches instantly became essentially "risk-free," or so Golden claimed as they were marketing them.

The only problem now is getting the pension funds and other big investors to buy the CDOs at the same price as if they were AAA-rated "risk-free" bonds. This was a little tricky because 1) it was virtually impossible for the buyer to figure out exactly what loans they were effectively buying since they were buying a tranche of a collection of tranches, and 2) if you could ever figure out what it was, you would discover that it was the junk of the junk when it comes to loans.

The solution was to enlist the help of the big bond-rating agencies: Substandard and Prevaricators (S&P) and Mopey's. If Golden could get AAA ratings[11] on their tranches, then the pension funds and other big investors would buy them at premium rates.

It turned out to be easier than you might think (or hope) to convince S&P and Mopey's to give high ratings to the CDO tranches. After all, the ratings agencies are trying to make money too and they make money by rating bonds. And Golden was only going to pay them if their bonds and CDOs got good ratings. And, since defaults were assumed to be essentially independent, there was a good argument as to why all but the mezzanine tranche of a bond or CDO would be essentially risk-free.[12]

So the stage is set for Golden to make a bundle of money. Cheap junk loans come in the back door and exit as expensive AAA-rated bonds and CDOs out the front door. The remaining challenge is to ramp up the new money-making machine. That

[11] AAA ratings are the best you can get and are supposed to imply that there is virtually no chance of default.

[12] The logic gets a little fuzzy when you keep slicing and dicing the tranches—after a few iterations, you should be able to conclude that the mezzanine tranche of a CDO is sure to have 100% defaults, but it required effort to see what was going on under the covers and effort costs money, and so the ratings agencies considered the risk of the mezzanine tranche of one CDO to be the same as the mezzanine tranche of any other, even though they could have wildly different probabilities of sustaining large numbers of defaults.

means creating more (preferably, many, many more) junk loans to fuel the machine.

This is where Joe enters the scene. Joe is a migrant laborer earning $15,000 per year. Joe's credit history is not great (since he has never had a loan or credit card) but it is also not bad (since he has never missed a payment on a credit card and never defaulted on a loan). In short, Joe is a perfect candidate for a subprime mortgage loan on a $750,000 home.

When Loans-Я-Us approaches Joe for a home loan,[13] Joe dutifully explains that while he would love to own a $750,000 home, he doesn't have enough money to pay for food, let alone the interest payments on the mortgage. "No problem!" replies Loans-Я-Us. It is Joe's lucky day. The interest rates are super-low for the first 2 years and Joe can take out a second loan to cover them during that period. "What happens after 2 years?" Joe wants to know. "No problem!" replies Loans-Я-Us. Joe can refinance—his home will surely be worth more in 2 years. Indeed, Joe can even make money while he enjoys the comforts of his new home. If all goes well, he can even ease off on the laborer work, and maybe even by a second home. Joe is sold. In fact, millions of Joes are sold and, before long, the subprime loan business is booming.

It turns out that there were a few folks out there who really did their math homework when they were in college. They were running hedge funds and, as the money-making machine was cranking away, they realized that a disaster was looming. They knew that loan defaults are not independent—in fact, they are very dependent. Once home values stop rising, or a recession hits, or it comes time for Joe to refinance, defaults will occur at much higher rates than projected and the CDOs and many tranches of the underlying bonds will become worthless. And there is so much money invested in these bonds and CDOs that the economy could be ruined.

Unfortunately, the folks who figured out what was going to happen didn't alert anyone. They didn't go to the newspapers. They didn't call the See no Evil Commission. They didn't even call 911. Instead, they worked with Golden to find a new way to make even more money—betting against the CDO market.

If you think a stock is going to decline, you can profit from the decline by borrowing the stock and selling it. After the stock declines in value, you buy it back and return it to the person that lent it to you. Your profit is the decline in price. This process is called *shorting* the stock.

So the hedge funds wanted to short the CDOs. Unfortunately, there was no established way to borrow a tranche of a CDO. Always looking for a new way to make money, the investment houses came up with an even bigger whopper than the

[13] Yes, we know it is supposed to go the other way around—Joe is supposed to approach the loan company—but these are extraordinary times.

CDO—they invented the *credit default swap*.

The idea behind the credit default swap is to provide a kind of insurance against the event that a bond or CDO suffers a certain number of defaults. Since the hedge funds believe that the CDOs were going to have lots of defaults, they want to buy the insurance. The trick is to find someone dumb enough to sell the insurance. That's where the world's largest insurance company, Awful Insurance Group (AIG), enters the fray. AIG sells insurance on just about anything and they, too, are looking for new ways to make money, so why not sell insurance on CDO defaults?

Golden has a new business! They buy the CDO insurance from AIG for an astonishingly low price (about $2 annually for every $1000 of CDO value) and sell it to the hedge funds for a much higher price (about $20 annually for every $1000 of CDO value). If a CDO sustains defaults, then AIG needs to pay the value of the CDO ($1000 in this hypothetical example) to the hedge funds who own the insurance. Until that time, the hedge funds are paying the annual fee for the insurance, 90% of which is pocketed by Golden. This is a great business; Golden pockets 90% of the money and AIG takes all the risk. The only risk that Golden has is if AIG goes down, but AIG is "too big to fail. . . ."

Golden's new credit default swap business is even better than the CDO business. The only trouble now is that there are only so many Joes out there who can take out subprime loans. This means that there is a hard limit on how many billions Golden can make. This challenge led to whopper number four.

If the hedge funds want to buy insurance and AIG wants to sell it, who really cares if there is only one insurance policy per loan or CDO? Indeed, why not just sell lots of credit default swaps on the same set of junk CDOs? This way, the profits could be unlimited! And so it went. "Synthetic" CDOs were created and soon the "insurance" quickly turned into a very high-stakes (and very stupid, at least for AIG) bet. The odds were weighted heavily in favor of the folks who did their math homework (the hedge funds); the hedge funds had figured out that the failure of the CDOs was a virtual certainty, whereas AIG believed that failure was virtually impossible.

Of course, we all know how the story ends. The holders of the CDOs and subprime debt and the sellers of insurance got wiped out, losing hundreds of billions of dollars. Since many of these folks were deemed by the Government as "too big to fail," they were bailed out using nearly a trillion dollars of taxpayer money. The executives who presided over the disaster were given huge bonuses because, well, that's how it works for executives in the land far, far away. The story also ends well for the hedge funds that bought the insurance—they made many, many billions of dollars.

So everyone involved in the disaster ends up very rich. Everyone except Joe, of course. Joe got kicked out of his home and lost his job in the recession.

Too bad for Joe that it isn't just a fairy tale.

6.042J / 18.062J Mathematics for Computer Science
Fall 2010