20 Deviation from the Mean

In the previous chapter, we took it for granted that expectation is useful and 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 expectation.

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. But when we make an estimate by repeated sampling, we need to know how much confidence we should have that our estimate is OK, and how large a sample is needed to reach a given confidence level. The issue is fundamental to all experimental science. Because of random errors—*noise*—repeated measurements of the same quantity rarely come out exactly the same. Determining how much confidence to put in experimental measurements is a fundamental and universal scientific issue. Technically, judging sampling or measurement accuracy reduces to finding the probability that an estimate *deviates* by a given amount from its expected value.

Another aspect of this issue comes up in engineering. When designing a sea wall, you need to know how strong to make it to withstand tsunamis for, say, at least a century. If you're assembling a computer network, you might need to know how many component failures it should tolerate to likely operate without maintenance for at least a month. If your business is insurance, you need to know how large a financial reserve to maintain to be nearly certain of paying benefits for, say, the next three decades. Technically, such questions come down to finding the probability of *extreme* deviations from the mean.

This issue of *deviation from the mean* is the focus of this chapter.

20.1 Markov's Theorem

Markov's theorem gives a generally coarse estimate of the probability that a random variable takes a value *much larger* than its mean. It is an almost trivial result by

itself, but it actually leads fairly directly to much stronger results.

The idea behind Markov's Theorem can be explained by considering the quantity known as *intelligence quotient*, IQ, which remains in wide use despite doubts about its legitimacy. IQ was devised so that its average measurement would be 100. This immediately implies that at most 1/3 of 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 $(1/3) \cdot 300 = 100$. So, the probability that a randomly chosen person has an IQ of 300 or more is at most 1/3. By the same logic, we can also conclude that at most 2/3 of the population can have an IQ of 150 or more.

Of course, these are not very strong conclusions. No IQ of over 300 has ever been recorded; and while many IQ's of over 150 have been recorded, the fraction of the population that actually has an IQ that high is very much smaller than 2/3. But though these conclusions are weak, we reached them using just the fact that the average IQ is 100—along with another fact we took for granted, that IQ is never negative. Using only these facts, we can't derive smaller fractions, because there are nonnegative random variables with mean 100 that achieve these fractions. 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.

Theorem 20.1.1 (Markov's Theorem)**.** *If R is a nonnegative random variable, then for all $x > 0$*

$$\Pr[R \geq x] \leq \frac{\mathrm{Ex}[R]}{x}. \tag{20.1}$$

Proof. Let I_x be the indicator variable for the event $[R \geq x]$. Then

$$x I_x \leq R$$

holds for all values of R since $R \geq 0$. Taking expectations of both sides yields

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

and then dividing both sides of this inequality by x gives (20.1). ∎

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

Corollary 20.1.2. *If R is a nonnegative random variable, then for all $c \geq 1$*

$$\Pr[R \geq c \cdot \mathrm{Ex}[R]\,] \leq \frac{1}{c}. \tag{20.2}$$

This Corollary follows immediately from Markov's Theorem(20.1.1) by letting x be $c \cdot \mathrm{Ex}[R]$.

20.1.1 Applying Markov's Theorem

Let's go back to the Hat-Check problem of Section 19.5.2. 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 different 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 $\mathrm{Ex}[R] = 1$, 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 precisely right!

Unfortunately, Markov's Theorem is not always so accurate. For example, it gives the same $1/n$ upper limit for the probability that everyone gets their own hat back in the Hat-Check problem, where the probability is actually $1/(n!)$. So for Hat-Check, Markov's Theorem gives a probability bound that is way too large.

20.1.2 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:

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

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 on R (see Problem 20.3). 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.

20.2 Chebyshev's Theorem

We've seen that Markov's Theorem can give a better bound when applied to $R - b$ rather than R. More generally, a good trick for getting stronger bounds on a random variable R out of Markov's Theorem is to apply the theorem to some cleverly chosen function of R. Choosing functions that are powers of the absolute value of R turns out to be especially useful. In particular, since $|R|^z$ is nonnegative for any real number z, Markov's inequality also applies to the event $[|R|^z \geq x^z]$. But for positive $x, z > 0$ this event is equivalent to the event $[|R| \geq x]$ for , so we have:

Lemma 20.2.1. *For any random variable R and positive real numbers x, z,*

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

Rephrasing (20.2.1) in terms of $|R - \text{Ex}[R]|$, the random variable that measures R's deviation from its mean, we get

$$\Pr[\,|R - \text{Ex}[R]\,| \geq x] \leq \frac{\text{Ex}[(|R - \text{Ex}[R]|)^z]}{x^z}. \tag{20.3}$$

When z is positive and even, $(R - \text{Ex}[R])^z$ is nonnegative, so the absolute value on the right-hand side of the inequality (20.3) is redundant. The case when $z = 2$ turns out to be so important that the numerator of the right-hand side has been given a name:

Definition 20.2.2. The *variance* of a random variable R is:

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

Variance is also known as *mean square deviation*.

The restatement of (20.3) for $z = 2$ is known as *Chebyshev's Theorem*.[1]

Theorem 20.2.3 (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.

20.2.1 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:

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

The expected payoff is the same for both games, but the games are very different. This difference is not apparent in their expected value, but is captured by variance.

[1] There are Chebyshev Theorems in several other disciplines, but Theorem 20.2.3 is the only one we'll refer to.

We can compute the Var[A] by working "from the inside out" as follows:

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

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!

20.2.2 Standard Deviation

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. The happens because 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 20.2.4. The *standard deviation* σ_R of a random variable R is the square root of the variance:

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

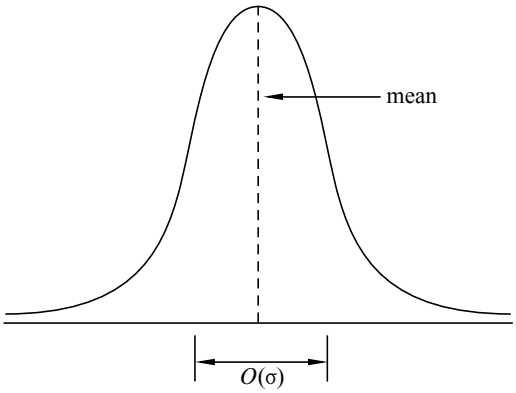


Figure 20.1 The standard deviation of a distribution indicates how wide the "main part" of it is.

So the standard deviation is the square root of the mean square 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.

Example 20.2.5. 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, so the standard deviation of 1416 describes this situation more closely than the value in the millions of the variance.

For bell-shaped distributions like the one illustrated in Figure 20.1, the standard deviation measures the "width" of the interval in which values are most likely to fall. This can be more clearly explained by rephrasing Chebyshev's Theorem in terms of standard deviation, which we can do by substituting $x = c\sigma_R$ in (20.1):

Corollary 20.2.6. *Let R be a random variable, and let c be a positive real number.*

$$\Pr[|R - \text{Ex}[R]| \geq c\sigma_R] \leq \frac{1}{c^2}. \tag{20.4}$$

Now 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.

The IQ Example

The standard standard deviation of IQ's regularly turns out to be about 15 even across different populations. This additional fact along with the national average IQ being 100 allows a better determination of the occurrence of IQ's of 300 or more.

Let the random variable R be the IQ of a random person. So $\text{Ex}[R] = 100$, $\sigma_R = 15$ and R is nonnegative. We want to compute $\Pr[R \geq 300]$.

We have already seen that Markov's Theorem 20.1.1 gives a coarse bound, namely,

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

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

$$\Pr[R \geq 300] = \Pr[|R - 100| \geq 200] \leq \frac{\text{Var}[R]}{200^2} = \frac{15^2}{200^2} \approx \frac{1}{178}.$$

So Chebyshev's Theorem implies that at most one person in 178 has an IQ of 300 or more. We have gotten a much tighter bound using additional information—the variance of R—than we could get knowing only the expectation.

20.3 Properties of Variance

Variance is the average *of the square* of the distance from the mean. For this reason, variance is sometimes called the "mean square deviation." Then we take its square root to get the standard deviation—which in turn is called "root mean square deviation."

But why bother squaring? Why not study the actual distance from the mean, namely, the absolute value of $R - \text{Ex}[R]$, instead of its root mean square? The answer is that variance and standard deviation have useful properties that make them much more important in probability theory than average absolute deviation. In this section, we'll describe some of those properties. In the next section, we'll see why these properties are important.

20.3.1 A Formula for Variance

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

Lemma 20.3.1.

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

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

$$\begin{aligned}
\text{Var}[R] &= \text{Ex}[(R - \text{Ex}[R])^2] && \text{(Def 20.2.2 of variance)}\\
&= \text{Ex}[(R-\mu)^2] && \text{(def of } \mu)\\
&= \text{Ex}[R^2 - 2\mu R + \mu^2]\\
&= \text{Ex}[R^2] - 2\mu\,\text{Ex}[R] + \mu^2 && \text{(linearity of expectation)}\\
&= \text{Ex}[R^2] - 2\mu^2 + \mu^2 && \text{(def of } \mu)\\
&= \text{Ex}[R^2] - \mu^2\\
&= \text{Ex}[R^2] - \text{Ex}^2[R]. && \text{(def of } \mu)
\end{aligned}$$

■

A simple and very useful formula for the variance of an indicator variable is an immediate consequence.

Corollary 20.3.2. *If B is a Bernoulli variable where $p ::= \Pr[B = 1]$ and $q ::= 1 - p$, then*

$$\text{Var}[B] = p - p^2 = pq. \tag{20.5}$$

Proof. By Lemma 19.4.2, $\text{Ex}[B] = p$. But B only takes values 0 and 1, so $B^2 = B$ and equation (20.5) follows immediately from Lemma 20.3.1. ■

20.3.2 Variance of Time to Failure

According to Section 19.4.6, the mean time to failure is $1/p$ for a process that fails during any given hour with probability p. What about the variance?

By Lemma 20.3.1,

$$\text{Var}[C] = \text{Ex}[C^2] - (1/p)^2 \tag{20.6}$$

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

Now $\text{Ex}[C^2] ::= \sum_{i\geq 1} i^2 q^{i-1} p$ by definition, and we could evaluate this series using methods from Chapter 14 or 16.

A simpler alternative appeals to conditional expectation much as we did in Section 19.4.6 to derive the formula for mean time to failure. Namely, the expected

value of C^2 is the probability p of failure in the first hour times 1^2, plus the probability q of non-failure in the first hour times the expected value of $(C+1)^2$. So

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

Combining this with (20.6) proves

Lemma 20.3.3. *If failures occur with probability p independently at each step, and C is the number of steps until the first failure,*[2] *then*

$$\operatorname{Var}[C] = \frac{q}{p^2}. \tag{20.7}$$

20.3.3 Dealing with Constants

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

Theorem 20.3.4. *[*Square Multiple Rule *for Variance] Let R be a random variable and a a constant. Then*

$$\operatorname{Var}[aR] = a^2 \operatorname{Var}[R]. \tag{20.8}$$

Proof. Beginning with the definition of variance and repeatedly applying linearity

[2] That is, C has the geometric distribution with parameter p according to Definition 19.4.7.

of expectation, we have:

$$\begin{aligned}
\operatorname{Var}[aR] &::= \operatorname{Ex}[(aR - \operatorname{Ex}[aR])^2] \\
&= \operatorname{Ex}[(aR)^2 - 2aR\operatorname{Ex}[aR] + \operatorname{Ex}^2[aR]] \\
&= \operatorname{Ex}[(aR)^2] - \operatorname{Ex}[2aR\operatorname{Ex}[aR]] + \operatorname{Ex}^2[aR] \\
&= a^2\operatorname{Ex}[R^2] - 2\operatorname{Ex}[aR]\operatorname{Ex}[aR] + \operatorname{Ex}^2[aR] \\
&= a^2\operatorname{Ex}[R^2] - a^2\operatorname{Ex}^2[R] \\
&= a^2\left(\operatorname{Ex}[R^2] - \operatorname{Ex}^2[R]\right) \\
&= a^2\operatorname{Var}[R] && \text{(Lemma 20.3.1)}
\end{aligned}$$

■

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

Theorem 20.3.5. *Let R be a random variable, and b a constant. Then*

$$\operatorname{Var}[R + b] = \operatorname{Var}[R]. \tag{20.9}$$

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 20.3.6.

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

20.3.4 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, such as Birthday Matching in Section 17.4, that involve variables that are pairwise independent but not mutually independent.

Theorem 20.3.7. *If R and S are independent random variables, then*

$$\operatorname{Var}[R + S] = \operatorname{Var}[R] + \operatorname{Var}[S]. \tag{20.10}$$

Proof. We may assume that $\operatorname{Ex}[R] = 0$, since we could always replace R by $R - \operatorname{Ex}[R]$ in equation (20.10); likewise for S. This substitution preserves the independence of the variables, and by Theorem 20.3.5, does not change the variances.

But for any variable T with expectation zero, we have $\text{Var}[T] = \text{Ex}[T^2]$, so we need only prove

$$\text{Ex}[(R+S)^2] = \text{Ex}[R^2] + \text{Ex}[S^2]. \tag{20.11}$$

But (20.11) follows from linearity of expectation and the fact that

$$\text{Ex}[RS] = \text{Ex}[R]\,\text{Ex}[S] \tag{20.12}$$

since R and S are independent:

$$\begin{aligned}
\text{Ex}[(R+S)^2] &= \text{Ex}[R^2 + 2RS + S^2] \\
&= \text{Ex}[R^2] + 2\,\text{Ex}[RS] + \text{Ex}[S^2] \\
&= \text{Ex}[R^2] + 2\,\text{Ex}[R]\,\text{Ex}[S] + \text{Ex}[S^2] \qquad \text{(by (20.12))} \\
&= \text{Ex}[R^2] + 2 \cdot 0 \cdot 0 + \text{Ex}[S^2] \\
&= \text{Ex}[R^2] + \text{Ex}[S^2].
\end{aligned}$$

■

It's easy to see that additivity of variance does not generally hold for variables that are not independent. For example, if $R = S$, then equation (20.10) becomes $\text{Var}[R+R] = \text{Var}[R] + \text{Var}[R]$. By the Square Multiple Rule, Theorem 20.3.4, this holds iff $4\,\text{Var}[R] = 2\,\text{Var}[R]$, which implies that $\text{Var}[R] = 0$. So equation (20.10) fails when $R = S$ and R has nonzero variance.

The proof of Theorem 20.3.7 carries over to the sum of any finite number of variables (Problem 20.19), so we have:

Theorem 20.3.8. *[Pairwise Independent Additivity of Variance] If $R_1, R_2, \ldots, R_n$ are* pairwise *independent random variables, then*

$$\text{Var}[R_1 + R_2 + \cdots + R_n] = \text{Var}[R_1] + \text{Var}[R_2] + \cdots + \text{Var}[R_n]. \tag{20.13}$$

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 pq by Corollary 20.3.2, so by linearity of variance, we have

Lemma 20.3.9 (Variance of the Binomial Distribution)**.** *If J has the (n, p)-binomial distribution, then*

$$\text{Var}[J] = n\,\text{Var}[I_k] = npq. \tag{20.14}$$

20.3.5 Matching Birthdays

We saw in Section 17.4 that in a class of 95 students, it is virtually certain that at least one pair of students will have the same birthday. In fact, several pairs of students are likely to have the same birthday. How many matched birthdays should we expect, and how likely are we to see that many matches in a random group of students?

Having matching birthdays for different pairs of students are *not* mutually independent events. If Alice matches Bob and Alice matches Carol, it's certain that Bob and Carol match as well! So the events that various pairs of students have matching birthdays are not even three-way independent.

But knowing that Alice's birthday matches Bob's tells us nothing about who Carol matches. This means that the events that a pair of people have matching birthdays are pairwise independent (see Problem 19.2). So pairwise independent additivity of variance, Theorem 20.3.8, will allow us to calculate the variance of the number of birthday pairs and then apply Chebyshev's bound to estimate the liklihood of seeing some given number of matching pairs.

In particular, suppose there are n students and d days in the year, and let M be the number of pairs of students with matching birthdays. 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 in our probability model, the B_i's are mutually independent variables, and 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 the number M of matching pairs of birthdays among the n choices is simply the sum of the $E_{i,j}$'s:

$$M = \sum_{1 \le i < j \le n} E_{i,j}. \tag{20.15}$$

Linearity of expectation make it easy to calculate the expected number of pairs of students with matching birthdays.

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

Similarly, pairwise independence makes it easy to calculate the variance.

$$\begin{aligned}
\mathrm{Var}[M] &= \mathrm{Var}\left[\sum_{1\le i<j\le n} E_{i,j}\right] \\
&= \sum_{1\le i<j\le n} \mathrm{Var}[E_{i,j}] && \text{(Theorem 20.3.8)} \\
&= \binom{n}{2}\cdot\frac{1}{d}\left(1-\frac{1}{d}\right). && \text{(Corollary 20.3.2)}
\end{aligned}$$

In particular, for a class of $n = 95$ students with $d = 365$ possible birthdays, we have $\mathrm{Ex}[M] \approx 12.23$ and $\mathrm{Var}[M] \approx 12.23(1-1/365) < 12.2$. So by Chebyshev's Theorem

$$\Pr[|M - \mathrm{Ex}[M]| \ge x] < \frac{12.2}{x^2}.$$

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 within 7 of 12.23, that is, between 6 and 19.

20.4 Estimation by Random Sampling

Massachusetts Democrats were astonished in 2010 when their early polls of sample voters showed Republican Scott Brown was favored by a majority of voters and so would win the special election to fill the Senate seat that the late Democrat Teddy Kennedy had occupied for over 40 years. Based on their poll results, they mounted an intense, but ultimately unsuccessful, effort to save the seat for their party.

20.4.1 A Voter Poll

Suppose at some time before the election that p was the fraction of voters favoring Scott Brown. We want to estimate this unknown fraction p. Suppose we have some random process for selecting voters from registration lists that selects each voter with equal probability. We can define an indicator 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[3] and

[3] We're choosing a random voter n times *with replacement*. 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! 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 for little gain.

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 independent. So formally, we model our estimation process by assuming we have mutually independent indicator 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,

$$S_n ::= \sum_{i=1}^{n} K_i. \tag{20.16}$$

The variable S_n/n describes the fraction of sampled voters who favor Scott Brown. Most people intuitively, and correctly, expect this sample fraction to give a useful approximation to the unknown fraction p.

So we will use the sample value S_n/n as our *statistical estimate* of p. We know that S_n has a binomial distribution with parameters n and p; we can choose n, but p is unknown.

How Large a Sample?

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

$$\Pr\left[\left|\frac{S_n}{n} - p\right| \leq 0.04\right] \geq 0.95\,. \tag{20.17}$$

So we'd better determine the number n of times we must poll voters so that inequality (20.17) will hold. Chebyshev's Theorem offers a simple way to determine such a n.

S_n is binomially distributed. Equation (20.14), combined with the fact that pq is maximized when $p = q$, that is, when $p = 1/2$ (check for yourself!), gives

$$\text{Var}[S_n] = n(pq) \leq n \cdot \frac{1}{4} = \frac{n}{4}. \tag{20.18}$$

Next, we bound the variance of S_n/n:

$$\begin{aligned}
\text{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \text{Var}[S_n] && \text{(Square Multiple Rule for Variance (20.8))}\\
&\leq \left(\frac{1}{n}\right)^2 \frac{n}{4} && \text{(by (20.18))}\\
&= \frac{1}{4n} && (20.19)
\end{aligned}$$

Using Chebyshev's bound and (20.19) we have:

$$\Pr\left[\left|\frac{S_n}{n} - p\right| \geq 0.04\right] \leq \frac{\mathrm{Var}[S_n/n]}{(0.04)^2} \leq \frac{1}{4n(0.04)^2} = \frac{156.25}{n} \tag{20.20}$$

To make our our estimate with 95% confidence, we want the right-hand side of (20.20) to be at most 1/20. So we choose n so that

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

that is,

$$n \geq 3,125.$$

Section 20.5.2 describes how to get tighter estimates of the tails of binomial distributions that lead to a bound on n that is about four times smaller than the one above. But working through this example using only the variance illustrates an approach to estimation that is applicable to arbitrary random variables, not just binomial variables.

20.4.2 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 called 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 20.4.1 (Pairwise Independent Sampling)**.** *Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean μ and deviation σ. Define*

$$S_n ::= \sum_{i=1}^{n} G_i. \tag{20.21}$$

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:

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

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

$$\begin{aligned}
\operatorname{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \operatorname{Var}[S_n] && \text{(Square Multiple Rule for Variance (20.8))}\\
&= \frac{1}{n^2}\operatorname{Var}\left[\sum_{i=1}^n G_i\right] && \text{(def of } S_n\text{)}\\
&= \frac{1}{n^2}\sum_{i=1}^n \operatorname{Var}[G_i] && \text{(pairwise independent additivity)}\\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. && (20.22)
\end{aligned}$$

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

$$\begin{aligned}
\Pr\left[\left|\frac{S_n}{n} - \mu\right| \geq x\right] &\leq \frac{\operatorname{Var}\left[S_n/n\right]}{x^2}. && \text{(Chebyshev's bound)}\\
&= \frac{\sigma^2/n}{x^2} && \text{(by (20.22))}\\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

■

The Pairwise Independent Sampling Theorem provides a quantitative 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:[4] by choosing a large enough sample size, we can get arbitrarily accurate estimates of the mean with confidence arbitrarily close to 100%.

[4]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.

Corollary 20.4.2. *[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 $\epsilon > 0$,

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

20.4.3 Confidence in an Estimation

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 influence the sample size. But our analysis shows that polling a little over 3000 people people is always sufficient, regardless of whether there are ten thousand, or a million, or a 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.*

As already discussed in Section 18.9, 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. To judge the credibility of the estimate, it may be important to examine how well this procedure was performed. More important, the confidence assertion above can be rephrased as

> **Either** the fraction of voters who prefer Brown is $1250/3125 \pm 0.04$
> **or** something unlikely (probability 1/20) happened.

If our experience led us to judge that having the preference fraction actually be in this particular interval was unlikely, then this level of confidence would justifiably remain unconvincing.

20.5 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.

20.5.1 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: it receives lots of forum posts that computer servers have to process. If any server is assigned more work than it can complete in a given interval, then it is overloaded and system performance suffers. That would be bad, because Fussbook users are *not* a tolerant bunch. So balancing the work load across mutliple servers is vital.

An early idea was to assign each server 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 from overload, the drawback of an *ad hoc* approach was clear: it's easy to miss something that will mess up your plan.

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, but approximation algorithms can come close. Unfortunately, in this case task lengths are not known in advance, which is typical of workload problems in the real world.

So the load balancing problem seems sort of hopeless, because there is no data available to guide decisions. So the programmers of Fussbook gave up and just randomly assigned posts to computers. Imagine their surprise when the system stayed up and hasn't crashed yet!

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

Specifically, Fussbook receives 24,000 forum posts in every 10-minute interval. Each post is assigned to one of several servers for processing, and each server works sequentially through its assigned tasks. It takes a server an average of $1/4$ second to process a post. Some posts, such as pointless grammar critiques and snide witticisms, are easier, but no post—not even the most protracted harangues—takes more than one full second.

Measuring workload in seconds, this means a server is overloaded when it is assigned more than 600 units of work in a given 600 second interval. Fussbook's average processing load of $24{,}000 \cdot 1/4 = 6000$ seconds per interval would keep 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 would 11 be enough? ... or 15, 20, 100? We'll answer that question with a new mathematical tool.

20.5.2 The Chernoff Bound

The Chernoff[5] 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 processor is just a bit below its capacity, then that processor is unlikely to be

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

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 20.5.6.

Theorem 20.5.1 (Chernoff Bound)**.** *Let $T_1, \ldots T_n$ 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 $c \geq 1$,*

$$\Pr[T \geq c \operatorname{Ex}[T]] \leq e^{-\beta(c)\operatorname{Ex}[T]} \tag{20.23}$$

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

The Chernoff bound applies only to distributions of sums of independent random variables that take on values in the real interval $[0, 1]$. The binomial distribution is the most well-known distribution that fits these criteria, but many others are possible, because the Chernoff bound allows the variables in the sum to have differing, arbitrary, or even unknown distributions over the range $[0, 1]$. Furthermore, there is no direct dependence on either 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!

20.5.3 Chernoff Bound for Binomial Tails

The Chernoff bound can be applied in easy steps, though the details can be daunting at first. Let's walk through a simple example to get the hang of it: bounding the probability that the number of heads that come up in 1000 independent tosses of a coin exceeds the 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 $\Pr[T \geq c \operatorname{Ex}[T]]$ where c = 1.2. To that end, we compute $\beta(c)$ as defined in the theorem:

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

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

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

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

The bound rapidly becomes much smaller 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 $\beta(c)$ rises from 0.0187 to about 0.0410, which may not seem significant, but because $\beta(c)$ appears in the exponent of the upper bound, the final probability decreases from around 1 in 10,000 to about 1 in a billion!

20.5.4 Chernoff Bound for a Lottery Game

Pick-4 is a lottery game in which 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. When 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—especially at the point where more than 2000 win and the lottery must 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[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.

[6] As we noted in Chapter 19, human choices are often not uniform and they can be highly dependent. For example, lots of people will pick an important date. The lottery folks should not get too much comfort from the analysis that follows, unless they assign random 4-digit numbers to each player.

Since 2000 winners would be twice the expected number, we choose $c = 2$, compute $\beta(c) = 0.386\ldots$, 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 more than it took in. In fact, the number of winners won't even be 10% higher than expected very often. To prove that, let $c = 1.1$, compute $\beta(c) = 0.00484\ldots$, 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 as to the outcome!

20.5.5 Randomized Load Balancing

Now let's return to Fussbook and its load balancing problem. Specifically, we need to determine a number m of servers that makes it very unlikely that any server is overloaded by being assigned more than 600 seconds of work in a given interval.

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, and otherwise T_i is the length of the task. So $T = \sum_{i=1}^{n} T_i$ is the total number of seconds of work 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 assignment of a post to a server is independent of the time required to process the post. The second condition is satisfied because we know that no post takes more than 1 second to process; this is why we chose to measure work in seconds.

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

$$\begin{aligned}\mathrm{Ex}[T] &= \frac{24{,}000 \text{ tasks} \cdot 1/4 \text{ second per task}}{m \text{ servers}} \\ &= 6000/m \text{ seconds.}\end{aligned} \tag{20.24}$$

So if there are fewer than 10 servers, then the expected load on the first server is greater than its capacity, and we can expect it to be overloaded. If there are exactly 10 servers, 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 based on the number of servers to bound the probability that the first server is overloaded. We have from (20.24)

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

so by the Chernoff bound

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

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 17.5.2. 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!

20.5.6 Proof of the Chernoff Bound

The proof of the Chernoff bound is somewhat involved. In fact, *Chernoff himself* couldn'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]

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

Proof. (of Theorem 20.5.1)

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:

$$\begin{aligned}
\Pr[T \geq c \operatorname{Ex}[T]] &= \Pr[c^T \geq c^{c \operatorname{Ex}[T]}] \\
&\leq \frac{\operatorname{Ex}[c^T]}{c^{c \operatorname{Ex}[T]}} && \text{(Markov Bound)} \\
&\leq \frac{e^{(c-1)\operatorname{Ex}[T]}}{c^{c \operatorname{Ex}[T]}} && \text{(Lemma 20.5.2 below)} \\
&= \frac{e^{(c-1)\operatorname{Ex}[T]}}{e^{c \ln(c) \operatorname{Ex}[T]}} = e^{-(c\ln(c)-c+1)\operatorname{Ex}[T]}.
\end{aligned}$$

■

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 of this supercharging 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 20.5.1.

Lemma 20.5.2.

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

Proof.

$$\begin{aligned}
\operatorname{Ex}\left[c^T\right] &= \operatorname{Ex}\left[c^{T_1+\cdots+T_n}\right] && \text{(def of } T\text{)} \\
&= \operatorname{Ex}\left[c^{T_1} \cdots c^{T_n}\right] \\
&= \operatorname{Ex}\left[c^{T_1}\right] \cdots \operatorname{Ex}[c^{T_n}] && \text{(independent product Cor 19.5.7)} \\
&\leq e^{(c-1)\operatorname{Ex}[T_1]} \ldots e^{(c-1)\operatorname{Ex}[T_n]} && \text{(Lemma 20.5.3 below)} \\
&= e^{(c-1)(\operatorname{Ex}[T_1]+\cdots+\operatorname{Ex}[T_n])} \\
&= e^{(c-1)\operatorname{Ex}[T_1+\cdots+T_n]} && \text{(linearity of Ex}[\cdot]\text{)} \\
&= e^{(c-1)\operatorname{Ex}[T]}.
\end{aligned}$$

The third equality depends on the fact that functions of independent variables are also independent (see Lemma 19.2.2). ■

Lemma 20.5.3.

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

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

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

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 real interval $[0, 1]$. ■

20.5.7 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

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

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

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

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

By Theorem 20.3.8, 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.

20.5.8 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*[8] 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.

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

Theorem 20.5.4 (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.

$$\begin{aligned}
\Pr[T = 0] &= \Pr[\overline{A}_1 \cap \overline{A}_2 \cap \ldots \cap \overline{A}_n] && (T = 0 \text{ iff no } A_i \text{ occurs}) \\
&= \prod_{i=1}^{n} \Pr[\overline{A}_i] && (\text{independence of } A_i) \\
&= \prod_{i=1}^{n} (1 - \Pr[A_i]) \\
&\leq \prod_{i=1}^{n} e^{-\Pr[A_i]} && (1 - x \leq e^{-x} \text{ for } 0 \leq x \leq 1) \\
&= e^{-\sum_{i=1}^{n} \Pr[A_i]} \\
&= e^{-\sum_{i=1}^{n} \mathrm{Ex}[T_i]} && (\text{since } T_i \text{ is an indicator for } A_i) \\
&= e^{-\mathrm{Ex}[T]} && (\text{linearity of expectation}) \quad \blacksquare
\end{aligned}$$

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 20.5.4 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 20.5.4 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.

Problems for Section 20.1

Practice Problems

Problem 20.1.
The vast majority of people have an above average number of fingers. Which of the following statements explain why this is true? Explain your reasoning.

1. Most people have a super secret extra bonus finger of which they are unaware.
2. A pedantic minority don't count their thumbs as fingers, while the majority of people do.
3. Polydactyly is rarer than amputation.
4. When you add up the total number of fingers among the world's population and then divide by the size of the population, you get a number less than ten.
5. This follows from Markov's Theorem, since no one has a negative number of fingers.
6. Missing fingers are more common than extra ones.

Class Problems

Problem 20.2.
A herd of cows is stricken by an outbreak of *cold cow disease*. The disease lowers a cow's body temperature from normal levels, 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) Use Markov's Bound 20.1.1 to prove that at most 3/4 of the cows could survive.

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

(c) Notice that the results of part (b) are purely arithmetic facts about averages, not about probabilities. But you verified the claim in part (a) by applying Markov's

bound on the deviation of a random variable. Justify this approach by regarding the temperature T of a cow as a random variable. Carefully specify the probability space on which T is defined: what are the sample points? what are their probabilities? Explain the precise connection between properties of T and average herd temperature that justifies the application of Markov's Bound.

Homework Problems

Problem 20.3.
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 b 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?

Exam Problems

Problem 20.4.
A herd of cows is stricken by an outbreak of *hot cow disease*. The disease raises the normal body temperature of a cow, and a cow will die if its temperature goes above 90 degrees. The disease epidemic is so intense that it raised the average temperature of the herd to 120 degrees. Body temperatures as high as 140 degrees, **but no higher**, were actually found in the herd.

(a) Use Markov's Bound 20.1.1 to prove that at most 2/5 of the cows could have survived.

(b) Notice that the conclusion of part (a) is a purely arithmetic facts about averages, not about probabilities. But you verified the claim of part (a) by applying Markov's bound on the deviation of a random variable. Justify this approach by explaining how to define a random variable T for the temperature of a cow. Carefully specify the probability space on which T is defined: what are the outcomes? what are their probabilities? Explain the precise connection between properties of T, average herd temperature, and fractions of the herd with various temperatures, that justify application of Markov's Bound.

Problems for Section 20.2

Exam Problems

Problem 20.5.

There is a herd of cows whose average body temperature turns out to be 100 degrees. Our thermometer produces such sensitive readings that no two cows have exactly the same body temperature. The herd is stricken by an outbreak of *wacky cow disease*, which will eventually kill any cow whose body temperature differs from the average by 10 degrees or more.

It turns out that the *collection-variance* of all the body temperatures is 20, where the *collection-variance* $\text{CVar}(A)$ of set A of numbers is

$$\text{CVar}(A) ::= \frac{\sum_{a\in A}(a-\mu)^2}{|A|},$$

where μ is the average value of the numbers in A.[9]

(a) Apply the Chebyshev bound to the temperature T of a random cow to show that at most 20% of the cows will be killed by this disease outbreak.

The conclusion of part (a) about a certain fraction of the herd was derived by bounding the deviation of a random variable. We can justify this approach by explaining how to define a suitable probability space in which, the temperature T of a cow is a random variable.

(b) Carefully specify the probability space on which T is defined: what are the outcomes? what are their probabilities?

(c) Explain why for this probability space, the fraction of cows with any given cow property P is the same as $\Pr[P]$.

(CONTINUED ON NEXT PAGE)

(d) Show that $\text{Var}[T]$ equals the collection variance of the temperatures in the herd.

[9] $\text{CVar}(A)$ is called A's *mean square deviation*.

Problems for Section 20.3

Practice Problems

Problem 20.6.
Suppose 120 students take a final exam and the mean of their scores is 90. You have no other information about the students and the exam, that is, you should not assume that the highest possible score is 100. You may, however, assume that exam scores are nonnegative.

(a) State the best possible upper bound on the number of students who scored at least 180.

(b) Now suppose somebody tells you that the lowest score on the exam is 30. Compute the new best possible upper bound on the number of students who scored at least 180.

Problem 20.7.
Suppose you flip a fair coin 100 times. The coin flips are all mutually independent.

(a) What is the expected number of heads?

(b) What upper bound does Markov's Theorem give for the probability that the number of heads is at least 70?

(c) What is the variance of the number of heads?

(d) What upper bound does Chebyshev's Theorem give for the probability that the number of heads is either less than 30 or greater than 70?

Problem 20.8.
Albert has a gambling problem. He plays 240 hands of draw poker, 120 hands of black jack, and 40 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. Let W be the expected number of hands that Albert wins in a day.

(a) What is $\mathrm{Ex}[W]$?

(b) What would the Markov bound be on the probability that Albert will win at least 216 hands on a given day?

(c) Assume the outcomes of the card games are pairwise independent. What is $\text{Var}[W]$? You may answer with a numerical expression that is not completely evaluated.

(d) What would the Chebyshev bound be on the probability that Albert will win at least 216 hands on a given day? You may answer with a numerical expression that includes the constant $v = \text{Var}[W]$.

Class Problems

Problem 20.9.

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 $\text{Ex}[X_i \cdot X_j]$ for $i \neq j$.

Hint: What is the probability that the second person got their hat back, given that the fifth person got their hat back, that is, $\Pr\left[X_2 = 1 \mid X_5 = 1\right]$?

(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) Use the Chebyshev bound to show that there is at most a 1% chance that more than 10 people get their own hat back.

Problem 20.10.

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,

$$\Pr[|R - \mu| \geq x] = \left(\frac{\sigma}{x}\right)^2. \tag{20.25}$$

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

Problem 20.11.
A computer program crashes at the end of each hour of use with probability $1/p$, if it has not crashed already. Let H be the number of hours until the first crash.

(a) What is the Chebyshev bound on

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

where $x > 0$?

(b) Conclude from part (a) 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 $H > a/p$.

(c) 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 (b).

Problem 20.12.
Let R be a positive integer-valued random variable.

(a) How large can $\mathrm{Ex}[1/R]$ be?

(b) How large can $\mathrm{Var}[R]$ be if the only values of R are 1 and 2?

(c) How large can $\mathrm{Var}[R]$ be if $\mathrm{Ex}[R] = 2$?

Problem 20.13.
A man has a set of n keys, one of which fits the door to his apartment. He tries the keys randomly throwing away each ill-fitting key that he tries until he finds the key that fits. That is, he chooses keys randomly from among those he has not yet tried. This way he is sure to find the right key within n tries.

Let T be the number of times he tries keys until he finds the right key. Problem 19.28 shows that

$$\mathrm{Ex}[T] = \frac{n+1}{2}.$$

Write a closed formula for $\mathrm{Var}[T]$.

Problem 20.14.
Let R be a positive integer valued random variable such that

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

where

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

(a) Prove that $\mathrm{Ex}[R]$ is finite.

(b) Prove that $\mathrm{Ex}[R^2]$ and therefore $\mathrm{Var}[R]$ are both infinite.

A joking way to phrase the point of this example is "the square root of infinity may be finite." Namely, let $T ::= R^2$; then part (b) implies that $\mathrm{Ex}[T] = \infty$ while $\mathrm{Ex}[\sqrt{T}] < \infty$ by (a).

Homework Problems

Problem 20.15.
A man has a set of n keys, one of which fits the door to his apartment. He tries a key at random, and if it does not fit the door, he simply puts it back; so he might try the same ill-fitting key several times. He continues until he finds the one right key that fits.

Let T be the number of times he tries keys until he finds the right key.

(a) Explain why

$$\mathrm{Ex}[T] = n \quad \text{and} \quad \mathrm{Var}[T] = n(n-1).$$

Let

$$f_n(a) ::= \Pr[T \geq an].$$

(b) Use the Chebyshev Bound to show that for any fixed $n > 1$,

$$f_n(a) = \Theta\left(\frac{1}{a^2}\right). \tag{20.26}$$

(c) Derive an upper bound for $f_n(a)$ that for any fixed $n > 1$ is asymptoticaly smaller than Chebyshev's bound (20.26).

You may assume that n is large enough to use the approximation

$$\left(1 - \frac{1}{n}\right)^{cn} \approx \frac{1}{e^c}$$

Problem 20.16.

There is a fair coin and a biased coin that flips heads with probability $3/4$. You are given one of the coins, but you don't know which. To determine which coin was picked, your strategy will be to choose a number n and flip the picked coin n times. If the number of heads flipped is closer to $(3/4)n$ than to $(1/2)n$, you will guess that the biased coin had been picked and otherwise you will guess that the fair coin had been picked.

(a) Use the Chebyshev Bound to find a value n so that with probability 0.95 your strategy makes the correct guess, no matter which coin was picked.

(b) Suppose you had access to a computer program that would generate, in the form of a plot or table, the full binomial-(n, p) probability density and cumulative distribution functions. How would you find the minimum number of coin flips needed to infer the identity of the chosen coin with probability 0.95? How would you expect the number n determined this way to compare to the number obtained in part(a)? (You do not need to determine the numerical value of this minimum n, but we'd be interested to know if you did.)

(c) Now that we have determined the proper number n, we will assert that the picked coin was the biased one whenever the number of Heads flipped is greater than $(5/8)n$, and we will be right with probability 0.95. What, if anything, does this imply about

$$\Pr\left[\text{picked coin was biased} \mid \#\text{ Heads flipped} \geq (5/8)n\right]?$$

Problem 20.17.

The *expected absolute deviation* of a real-valued random variable R with mean μ, is defined to be

$$\text{Ex}[\,|R - \mu|\,].$$

Prove that the expected absolute deviation is always less than or equal to the standard deviation σ. (For simplicity, you may assume that R is defined on a finite sample space.)

Hint: Suppose the sample space outcomes are $\omega_1, \omega_2, \ldots, \omega_n$, and let

$$\mathbf{p} ::= (p_1, p_2, \ldots, p_n) \quad \text{where } p_i = \sqrt{\Pr[\omega_i]},$$
$$\mathbf{r} ::= (r_1, r_2, \ldots, r_n) \quad \text{where } r_i = |R(\omega_i) - \mu| \sqrt{\Pr[\omega_i]}.$$

As usual, let $\mathbf{v} \cdot \mathbf{w} ::= \sum_{i=1}^{n} v_i u_i$ denote the dot product of n-vectors $\mathbf{v}, \mathbf{w}$, and let $|\mathbf{v}|$ be the norm of $\mathbf{v}$, namely, $\sqrt{\mathbf{v} \cdot \mathbf{v}}$.

Then verify that

$$|\mathbf{p}| = 1, \qquad |\mathbf{r}| = \sigma, \qquad \text{and} \qquad \text{Ex}[\,|R - \mu|\,] = \mathbf{r} \cdot \mathbf{p}.$$

Problem 20.18.
Prove the following "one-sided" version of the Chebyshev bound for deviation above the mean:

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

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

Hint: Let $S_a ::= (R - \text{Ex}[R] + a)^2$, for $0 \leq a \in \mathbb{R}$. So $R - \text{Ex}[R] \geq x$ 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 20.19.
Prove the pairwise independent additivity of variance Theorem 20.3.8: If $R_1, R_2, \ldots, R_n$ are pairwise independent random variables, then

$$\text{Var}[R_1 + R_2 + \cdots + R_n] = \text{Var}[R_1] + \text{Var}[R_2] + \cdots + \text{Var}[R_n]. \qquad (*)$$

Hint: Why is it OK to assume $\text{Ex}[R_i] = 0$?

Exam Problems

Problem 20.20.
You are playing a game where you get n turns. Each of your turns involves flipping a coin a number of times. On the first turn, you have 1 flip, on the second turn you have two flips, and so on until your nth turn when you flip the coin n times. All the flips are mutually independent.

The coin you are using is biased to flip Heads with probability p. You *win* a turn if you flip all Heads. Let W be the number of winning turns.

(a) Write a closed-form (no summations) expression for $\text{Ex}[W]$.

(b) Write a closed-form expression for $\text{Var}[W]$.

Problem 20.21.

Let K_n be the complete graph with n vertices. Each of the edges of the graph will be randomly assigned one of the colors red, green, or blue. The assignments of colors to edges are mutually independent, and the probability of an edge being assigned red is r, blue is b, and green is g (so $r + b + g = 1$).

A set of three vertices in the graph is called a *triangle*. A triangle is *monochromatic* if the three edges connecting the vertices are all the same color.

(a) Let m be the probability that any given triangle T is monochromatic. Write a simple formula for m in terms of r, b, and g.

(b) Let I_T be the indicator variable for whether T is monochromatic. Write simple formulas in terms of m, r, b and g for $\text{Ex}[I_T]$ and $\text{Var}[I_T]$.

$$\text{Ex}[I_T] =$$
$$\text{Var}[I_T] =$$

Let T and U be distinct triangles.

(c) What is the probability that T and U are both monochromatic if they do not share an edge?... if they do share an edge?

$$\textbf{Now assume } r = b = g = \frac{1}{3}\textbf{.}$$

(d) Show that I_T and I_U are independent random variables.

(e) Let M be the number of monochromatic triangles. Write simple formulas in terms of n and m for $\text{Ex}[M]$ and $\text{Var}[M]$.

$$\text{Ex}[M] =$$
$$\text{Var}[M] =$$

(f) Let $\mu ::= \text{Ex}[M]$. Use Chebyshev's Bound to prove that

$$\Pr\left[\,|M-\mu| > \sqrt{\mu \log \mu}\,\right] \leq \frac{1}{\log \mu}.$$

(g) Conclude that

$$\lim_{n\to\infty} \Pr\left[\,|M-\mu| > \sqrt{\mu \log \mu}\,\right] = 0$$

Problem 20.22.
You have a biased coin which flips Heads with probability p. You flip the coin n times. The coin flips are all mutually independent. Let H be the number of Heads.

(a) Write a closed-form (no summations) expression in terms of p and n for $\text{Ex}[H]$, the expected number of Heads. Briefly explain your answer.

(b) Write a closed-form expression in terms of p and n for $\text{Var}[H]$, the variance of the number of Heads. Briefly explain your answer.

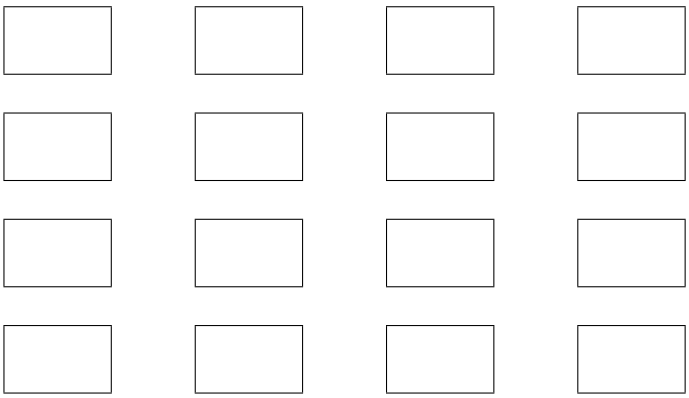
(c) Write a closed-form expression in terms of p for the upper bound that Markov's Theorem gives for the probability that the number of Heads is larger than the expected number by at least 1% of the number of flips, that is, by $n/100$.

(d) Show that the upper bound given by Chebyshev's Theorem for the probability that H differs from $\text{Ex}[H]$ by at least $n/100$ is

$$100^2 \frac{p(1-p)}{n}.$$

(e) The bound in part (d) implies that if you flip at least n times for a certain number n, then there is a 95% chance that the proportion of Heads among these n flips will be within 0.01 of p. Use the result from part (d) to write a simple expression for n in terms of p.

Problem 20.23.
A classroom has sixteen desks in a 4×4 arrangement as shown below.



If two desks are next to each other, vertically or horizontally, they are called an *adjacent pair*. So there are three horizontally adjacent pairs in each row, for a total of twelve horizontally adjacent pairs. Likewise, there are twelve vertically adjacent pairs. An adjacent pair D of desks is said to have a *flirtation* when there is a boy at one desk and a girl at the other desk.

(a) Suppose boys and girls are assigned to desks in some unknown probabilistic way. What is the Markov bound on the probability that the number of flirtations is at least 33 1/3% more than expected?

Suppose that boys and girls are actually assigned to desks mutually independently, with probability p of a desk being occupied by a boy, where $0 < p < 1$.

(b) Express the expected number of flirtations in terms of p.

Hint: Let I_D be the indicator variable for a flirtation at D.

Different pairs D and E of adjacent desks are said to *overlap* when they share a desk. For example, the first and second pairs in each row overlap, and so do the second and third pairs, but the first and third pairs do not overlap.

(c) Prove that if D and E overlap, and $p = 1/2$, then I_D and I_E are independent.

(d) When $p = 1/2$, what is the variance of the number of flirtations?

(e) What upper bound does Chebyshev's Theorem give on the probability that the number of heads is either less than 30 or greater than 70?

(f) Let D and E be pairs of adjacent desks that overlap. Prove that if $p \neq 1/2$, then F_D and F_E are *not* independent.

(g) Find four pairs of desks D_1, D_2, D_3, D_4 and explain why $F_{D_1}, F_{D_2}, F_{D_3}, F_{D_4}$ are *not* mutually independent (even if $p = 1/2$).

Problems for Section 20.4

Class Problems

Problem 20.24.
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 percentage of adult Americans who believe in evolution is 35 ± 3 percent?

Problem 20.25.
Let $B_1, B_2, \ldots, B_n$ be mutually independent random variables with a uniform distribution on the integer interval $[1..d]$. Let $E_{i,j}$ be the indicator variable for the event $[B_i = B_j]$.

Let M equal the number of events $[B_i = B_j]$ that are true, where $1 \le i < j \le n$. So

$$M = \sum_{1 \le i < j \le n} E_{i,j}.$$

It was observed in Section 17.4 (and proved in Problem 19.2) that $\Pr[B_i = B_j] = 1/d$ for $i \neq j$ and that the random variables $E_{i,j}$, where $1 \le i < j \le n$,

are pairwise independent.

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

(b) What are $\text{Ex}[M]$ and $\text{Var}[M]$?

(c) In a 6.01 class of 500 students, the youngest student was born 15 years ago and the oldest 35 years ago. Show that more than half the time, there will be will be between 12 and 23 pairs of students who have the same birth date. (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.)

Hint: Let D be the number of pairs of students in the class who have the same birth date. Note that $|D - \text{Ex}[D]| < 6$ IFF $D \in [12..23]$.

Problem 20.26.

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 20.27.

The proof of the Pairwise Independent Sampling Theorem 20.4.1 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 $\epsilon > 0$*,*

$$\Pr[|A_n - \mu_n| \geq \epsilon] \leq \frac{b}{\epsilon^2} \cdot \frac{1}{n}. \tag{20.27}$$

(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 20.28.

Let $G_1, G_2, G_3, \ldots$, be an infinite sequence of pairwise independent random variables with the same expectation μ and the same finite variance. Let

$$f(n,\epsilon) ::= \Pr\left[\left|\frac{\sum_{i=1}^n G_i}{n} - \mu\right| \leq \epsilon\right].$$

The Weak Law of Large Numbers can be expressed as a logical formula of the form:

$$\mathbf{Q}_0\ \mathbf{Q}_1\ \mathbf{Q}_2\ \mathbf{Q}_3.\ f(n,\epsilon) \geq 1 - \delta$$

where $\mathbf{Q}_0, \mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3$ is a sequence of four quantifiers from among:

$$\begin{array}{cccc} \forall n, & \exists n, & \forall n \geq n_0, & \exists n \geq n_0. \\ \forall n_0, & \exists n_0, & \forall n_0 \geq n, & \exists n_0 \geq n. \\ \forall \delta, & \exists \delta, & \forall \delta > 0, & \exists \delta > 0. \\ \forall \epsilon, & \exists \epsilon, & \forall \epsilon > 0, & \exists \epsilon > 0. \end{array}$$

Here the n, n_0 range over nonnegative integers, and δ, ϵ range over nonnegative real numbers.

Write out the proper sequence $\mathbf{Q}_0, \mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3$.

Exam Problems

Problem 20.29.

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 a random voter and ask them who they are going to vote for. You do this n times, with each voter selected with uniform probability and independently of other selections. Finally, you use the fraction P of voters who said 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. Indicate the true facts among the following:

1. Given a particular voter, the probability of that voter preferring the President is p.
2. The probability that some voter is chosen more than once in the random sample goes to one as n increases.
3. The probability that some voter is chosen more than once in the random sample goes to zero as the population of voters grows.
4. All voters are equally likely to be selected as the third in the random sample of n voters (assuming $n \geq 3$).
5. The probability that the second voter in the random sample will favor the President, given that the first voter prefers the President, is greater than p.
6. The probability that the second voter in the random sample will favor the President, given that the second voter 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. Among the following, Indicate the legitimate things you might say in a call to the President:

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.

Problem 20.30.
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 *random 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 random 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 *random sample* is defective is b.
3. All lines of code in the program are equally likely to be the third line chosen in the *random sample*.
4. Given that the first line chosen for the *random 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 *random sample* being buggy is b.

7. Given that the first two lines of code selected in the *random 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 *random sample* will be different.

Problems for Section 20.5

Practice Problems

Problem 20.31.
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*, but possibly *not* mutually, independent. What is the variance in the number of hands won per day? You may answer with a numerical expression that is not completely evaluated.

(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.

(e) Assuming outcomes of the card games are *mutually* independent, show that the probability that the gambler will win at least 108 hands on a given day is much smaller than the bound in part (d). *Hint:* $e^{1-2\ln 2} \leq 0.7$

Class Problems

Problem 20.32.
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: Use Chernoff's Bound, Theorem 20.5.1,. Note that $\beta(12) > 18$, where $\beta(c) ::= c \ln c - c + 1$.

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

a

Problem 20.33.
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) Give a tight upper bound on the probability that I forget one or more items when no independence assumption is made about forgetting different items.

(c) Use the Markov Bound to derive an upper bound on the probability that I forget 3 or more items.

(d) Now suppose that I forget each item of footwear independently. Use the Chebyshev Bound to derive an upper bound on the probability that I forget two or more items.

(e) Use Murphy's Law, Theorem 20.5.4, to derive a lower bound on 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 $\mathrm{Ex}[X] = 36$. 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 20.34.
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 20.35.
We have two coins: one is a fair coin, but the other produces heads with probability $3/4$. One of the two coins is picked, and this coin is tossed n times. Use the Chernoff Bound to determine the smallest n which allows determination of which coin was picked with 95% confidence.

Problem 20.36.
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

$$\sum_{n\in\mathbb{N}} \Pr[A_n] = \infty. \tag{20.28}$$

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 20.5.4, to B_k.

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

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.