18 Deviations

~~17~~ Deviation from ~~the Mean~~

— INSERT A goes here

~~17.1 Why the Mean?~~

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

~~The most important reason to care about the mean value comes from its connection to~~

INSERT A

In some cases, a random variable is likely to be very to be close to its ~~mean~~ expected value. For example, if we flip 100 fair, mutually-independent coins; ~~we~~ it is very likely that ~~the number of~~ we will get ~~heads will be clo~~ about 50 heads. In fact, we proved ~~that th~~ in section ___ that the probability of getting ~~we will get~~ ~~there is~~ fewer than 25 ~~tee~~ or more than 75 heads is less than ______. ~~In such~~ In such cases, the mean provides a lot of information about the random variable.

In other cases, a random ~~very~~ variable is likely to be ~~very~~ far from its expected value. For example, suppose we flipped 100 fair coins that are glued together so that they all come out "heads" or they all come out "tails". In this case, ~~we~~ the expected value of the number of heads is still 50, but ~~we~~ ~~are guaranteed to~~ the actual number of heads is guaranteed to be far from this value—

it will be 0 or 100, each with probability 1/2.

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

~~we will study [illegible] this measure at~~

18.1 Variance

18.1.1 Definitions and Examples

Consider the following two gambling games:

— INSERT 2A goes here —
(it is text on pp1137-1138)

no # The stakes are alot higher for Game B and so it is likely to deviate much farther from its mean than is Game A. This fact is captured by the notion of <u>variance</u>.

——— INSERT 3B goes here ———
(text from p 11 36)

In words, the variance of a random variable R is the expectation of the square of the amount by which R differs from its expectation. ~~That~~ ~~That was~~ That's a mouthful.

Yikes! ~~That's a mouthful~~ Try saying that 10 times in a row! ~~Let's break it up~~

Let's look at this definition more carefully. First, we look at R - Ex[R]. That's the amount by which ~~the ε~~ R differs from its expectation and it is obviously an important measure. ¶Next, we square this value. ~~We~~ more on why we do that in a moment. Finally, ~~do that so the difference will be positive. otherwise, Otherwise, the differences will cancel out and we we won't get a meaningful result. Of course, we could have also taken the absolute value, but the math doesn't work out as nicely if we instead of squaring, but the resulting definition doesn't work so won't have as nice properties if we do that.~~

~~Lastly,~~ we take the expected value of the square. If it is ~~large, then we expect~~ likely to be large, then the variance will be large. If it is likely to be small, then the variance will be small. That's just the kind of statistic we are looking for. Let's see ~~how this~~ ~~what~~ how it works out for our two gambling games.

We'll start with Game A:

——— INSERT ZC goes here ———
(text on pp 1138-1139)

~~Wy why the s~~

Why Bother Squaring? ← sub subsection

——— INSERT ZD goes here ———
(text on pp 1145-1146)

——— INSERT ZE goes here ———
(text on pp 1140-1141)

——— INSERT ZF goes here ———
(text on pp 1146-1148)

~~For example, for Game A in Subsection 18.1.1,~~

~~$\text{Ex}[A] =$~~

For example, let's take another look at Game A from Subsection 18.1.1 where you with $2 with probability 2/3 and lose ~~$2~~ $1 ~~with~~ with probability 1/3. Then

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

and

~~$\text{Ex}^2[A]$~~

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

~~so~~ By Lemma 17.4.1, this means that

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

confirming the result in Equation A7.

—— INSERT ZG goes here ——
(text on pp 1154 - 1156)

18.1.4 Indicator Random Variables

Computing the variance of an indicator random variable is straightforward given Lemma 17.4.1.

— INSERT A H goes here —
(text on p1149)

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

$$\text{Var}[R] = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4} \qquad \text{(eqn H1)}$$

and

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

~~For~~ Mean time to Failure ⟵ subsubsection

As another example, consider the mean time to failure problem described in Section 17.1.5 (please check this # is right). ~~We know that~~ If the system crashes at each ~~on step a~~ step with probability p, then we already know that the mean time to failure is $\frac{1}{p}$. In other words, if C is the number of steps up to and including the step ~~@~~ when the first crash occurs, then

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

What about the variance of c? To use Lemma 17.4.1, we need to compute $\mathrm{Ex}[c^2]$. ~~We can do this~~ As in Section 17.1.5, we can do this by summing over all the sample points or we can use the law of Total Expectation. The latter approach is simpler, so we'll do that. The analysis breaks into two cases: the ~~q~~system crashes in the first step or it doesn't. ~~$\mathrm{Ex}[c^2] =$~~ Hence,

$$\mathrm{Ex}[c^2] = 1^2 \cdot p + \mathrm{Ex}[(c+1)^2](1-p)$$

$$= p + \mathrm{Ex}[c^2]\overset{(1-p)+}{\cancel{*}}\, 2\,\mathrm{Ex}[\cancel{p}c](1-p)$$

$$+ (1-p)$$

$$= 1 + \mathrm{Ex}[c^2](1-p) + 2\,\frac{(1-p)}{p}.$$

Simplifying, we find that

$$p\,\mathrm{Ex}[c^2] = \cancel{\frac{p^2+2-2p}{p^2}}\quad \frac{2-p}{p}$$

and that

$$\mathrm{Ex}[c^2] = \cancel{\frac{p^2-2p+2}{p^3}}\quad \frac{2-p}{p^2}$$

using Lemma 17.4.1, we conclude that

$$\mathrm{Var}[C] = \mathrm{Ex}[C^2] - \mathrm{Ex}^2[C]$$

$$= ~~\frac{p^2-2p+2}{p^3} - \frac{1}{p^2}~~ \quad \frac{2-p}{p^2} - \frac{1}{p^2}$$

$$= ~~\frac{p^2-3p+2}{p^3}~~ \quad \frac{1-p}{p^2}.$$

18.1.5 Uniform Random Variables

Computing the expected value of a uniform random variable is also ~~ex~~ straightforward given Lemma 17.4.1 For

— INSERT A1 goes here —
(text on pp 1152)

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

$$\text{Ex}[R] = \frac{1}{n}(1+2+\cdots+n)$$

$$= \frac{1}{n}\,\frac{n(n+1)}{2}$$

$$= \cancel{\frac{n}{2}}\,\frac{n+1}{2}$$

$$\text{Ex}[R^2] = \frac{1}{n}\left(0^2+1^2+2^2+\cdots+n^2\right)$$

$$= \frac{1}{n}\,\frac{(2n+1)\,n(n+1)}{6} \qquad \text{(Equation } \text{—)}$$

$$= \frac{(2n+1)(n+1)}{6}.$$

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

$$= \frac{(2n+1)(n+1)}{6} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{4n^2+2n-3n^2}{12}$$

$$= \cancel{\frac{n^2+2n}{2}}\ \frac{n^2-1}{12}.$$

A-~~8~~10

~~18.1.6 Var [aR+b]~~

—— INSERT AJ goes here ——
(text on pp1156 - 1159)

Proof. ~~By Lemma 17.4.1,~~ As with the proof of theorem 17.4.4, this proof uses ~~the~~ repeated applications of Lemma 17.4.1 and Linearity of Expectation.

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

The last step follows because ~~since~~ $\mathrm{Ex}[R_1R_2] = \mathrm{Ex}[R_1]\,\mathrm{Ex}[R_2]$ when R_1 and R_2 are independent. □

~~Note that Theorem 17.4.7 holds if and only if R_1 and R_2 are independent~~

Note that Theorem 17.4.7 does not necessarily hold if R_1 and R_2 are ~~not~~ dependent since then it would generally not be true that

$$\text{Ex}[R_1 R_2] = \text{Ex}[R_1] \cdot \text{Ex}[R_2] \qquad \text{(eqn J1)}$$

in the last step of the proof. For example, suppose that $R_1 = R_2 = R$. Then Equation J1 holds only if R is essentially constant.

— INSERT AK goes here —
(text on p 1161)

18.1.8 Binomial Distributions

Unfortunately, there is no product rule for computing ~~the~~ variances ~~of~~, even if the random variables are mutually independent. However, we can use Theorem 17.4.8 to quickly compute the variance of a random variable with a general binomial distribution.

— INSERT AL goes here ———
(text on p 1163)

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

$$\Pr[I_p = 1] = p.$$

From Lemma 17.4.2, we know that

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

By Theorem 17.4.8, this means that

$$\text{Var}[J] = n \,\text{Var}[I_p] = np(1-p),$$

as claimed. □

For example, suppose we flip n mutually independent[1] fair coins. Let R be the number of heads, then ~~We already know~~ ~~linearity~~ Theorem 17.4.8 tells us that ~~the variance of the number~~

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

Hence,

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

This value is small compared with ~~Ex[R~~

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

which should not be surprising since we already knew from Section 16.5 (check #) that R is unlikely to stray very far from its mean.

[1] Actually, ~~the~~ we only need to assume pairwise independence for this to be true ~~from~~ using Theorem 17.4.8

18.2 Markov's Theorem

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

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

&

18.2.1 A Motivating Example

~~< INSERT AM goes here >~~

~~(this is 4.~~

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

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

~~The first technical result about deviation will be Markov's Theorem, which gives a simple, but typically coarse, upper bound on the probability that the value of a random variable is more than a certain multiple of its mean. Markov's result holds if we know nothing about a random variable except what its mean is and that its values are non-negative. Accordingly, Markov's Theorem is very general, but also is much weaker than results which take into account more information about the distribution of the variable.~~

~~In many situations, we not only know the mean, but also another numerical quantity called the *variance* of the random variable. The second basic result is Chebyshev's Theorem, which combines Markov's Theorem and information about the variance to give more refined bounds.~~

~~The final result we obtain about deviation is Chernoff's bound. Chernoff's bound applies to a random variable that is a sum of bounded independent random variables. Its bound is exponentially tighter than the other two.~~

~~**EDITING NOTE**: A random variable may never take a value anywhere near its expected value, so why is its expected value important? The reason is suggested by a property of gambling games that most people recognize intuitively. Suppose your gamble hinges on the roll of two dice, where you win if the sum of the dice is seven. If the dice are fair, the probabilty you win is 1/6, which is also your expected number of wins in one roll. Of course there's no such thing as 1/6 of a win in one roll, since either you win or you don't. But if you play *many times*, you would expect that the *fraction* of times you win would be close to 1/6. In fact, if you played a lot of times and found that your fraction of wins wasn't pretty close to 1/6, you would become pretty sure that the dice weren't fair.~~

~~17.2 Markov's Theorem~~

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

The idea behind Markov's Theorem can be explained with a simple example ~~of~~ involving *intelligence quotient*s, or IQs. ~~, IQ.~~ This quantity was devised so that the average IQ measurement would be 100. ~~Now~~ From this fact alone we can conclude that at most 1/3 the population can have an IQ of 300 or more, because if more than a third had an IQ of 300, then the average would have to be *more* than (1/3)300 = 100, contradicting the fact that the average is 100. So the probability that a randomly chosen person has an IQ of 300 or more is at most 1/3. Of course this is not a very strong conclusion; in fact no IQ ~~of over 300~~ over 200 ~~3 or more~~ has ever been recorded. ~~But by~~ # By the same logic, we can also conclude that at most 2/3 of the population can have an IQ of 150 or more. IQ's ~~of~~ over 150 have certainly been recorded, ~~though again,~~ although a much smaller fraction than 2/3 of the population actually has an IQ that high.

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

~~Markov~~

~~**EDITING NOTE**:~~

~~Note that very different distributions can still have the same mean.~~

~~*Example* 17.2.1. Suppose that we roll a fair die. This gives a random variable uniformly~~

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

18.2.2 The Theorem

distributed on $1, 2, \ldots, 6$. The mean, or expected value, is 3.5. Of course, this random variable never takes on exactly the expected value; in fact, the outcome deviates from the mean by at least 0.5 with probability 1. Furthermore, there is a $\frac{2}{3}$ probability that the outcome deviates from the mean by at least 1.5 (roll 1, 2, 5, or 6), a $\frac{1}{3}$ probability that the outcome deviates by at least 2.5 (roll 1 or 6), and zero probability that the outcome deviates by more than 2.5.

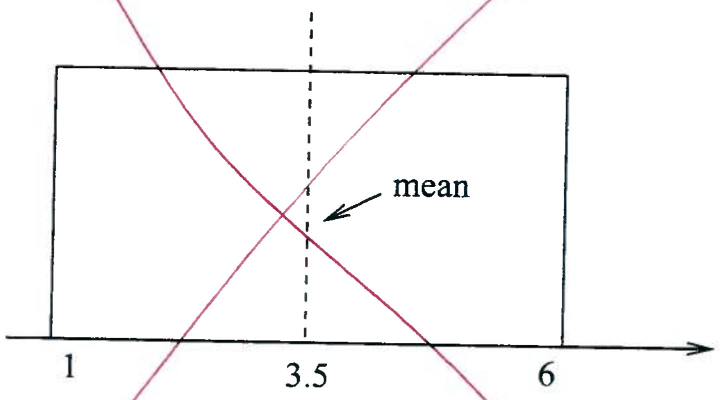
Example 17.2.2. A random variable with the binomial distribution is much less likely to deviate far from the mean. For example, suppose we flip 100 fair, mutually independent coins and count the number of heads. The expected number of heads is 50. There is an 8% chance that the outcome is exactly the mean, and the probability of flipping more than 75 heads or fewer than 25 is less than 1 in a billion.

The probability distribution functions for the two preceding examples are graphed in

~~Figure 17.1 and Figure 17.2. There is a big difference! For the uniform distribution, the graph is flat; that is, outcomes far from the mean are as likely as outcomes close to the mean. However, the binomial distribution has a peak centered on the expected value and the tails fall off rapidly. This shape implies that outcomes close to the expected value are vastly more likely than outcomes far from the expected value. In other words, a random variable with the binomial distribution rarely deviates far from the mean.~~

~~On the other hand, we can define a random variable that always deviates substantially from its expected value. Suppose that we glue 100 coins together, so that with probability 1/2 all are heads and with probability 1/2 all are tails. The graph of the probability distribution function for the number of heads is shown in Figure 17.3. While the expected value of this random variable is 50, the actual value is always 0 or 100.~~

~~Even in this last example, however, the random variable is twice the mean with probability only 1/2. In fact, we will see that this is a worst-case distribution with respect to~~



~~Figure 17.1: This is a graph of the uniform distribution arising from rolling a fair die. Outcomes within the range of the distribution are equally likely, regardless of distance from the mean.~~

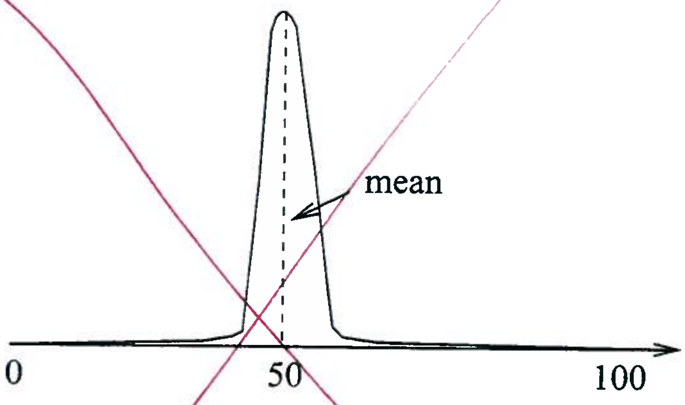
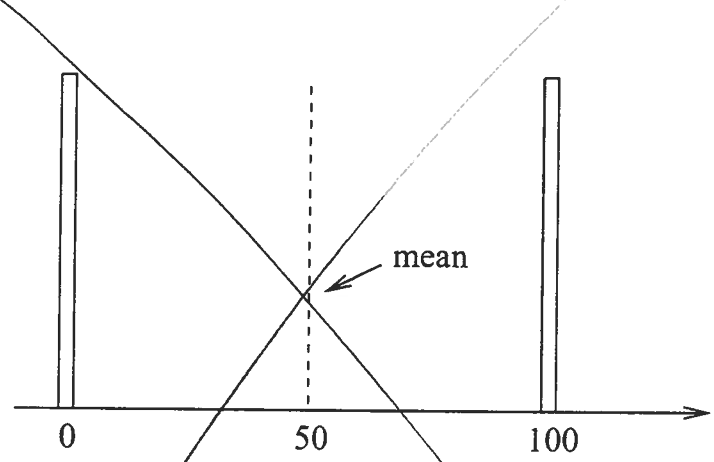


Figure 17.2: This is a rough graph of the binomial distribution given by the number of heads that come up when we flip 100 fair, mutually independent coins. Outcomes close to the mean are much more likely than outcomes far from the mean.



~~Figure 17.3: This is the nasty distribution corresponding to the number of heads that come up when we flip 100 coins that are all glued together. The outcome always differs from the mean by at least 50.~~

~~deviation from the mean.~~

~~Theorem Statement and Some Applications~~

Theorem 17.2.3 (Markov's Theorem). *If R is a nonnegative random variable, then for all* $x > 0$

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

~~EDITING NOTE:~~

~~Before we prove Markov's Theorem, let's apply it to the three examples in the preceding subsection. First, let the random variable R be the number that comes up when we roll a fair die. By Markov's Theorem, the probability of rolling a 6 is at most:~~

$$\Pr\{R \geq 6\} \leq \frac{\mathrm{E}[R]}{6} = \frac{3.5}{6} = 0.583\ldots$$

This conclusion is true, but weak. The actual probability of rolling a 6 is $1/6 = 0.166\ldots$.

This is typical of Markov's Theorem. The theorem is easy to apply because it requires so little information about a random variable, only the expected value and nonnegativity. But as a consequence, Markov's Theorem often leads to weak conclusions like the ~~one above.~~

As an example, suppose ~~Suppose~~ that we flip 100 mutually independent, fair coins. Markov's Theorem says that the probability of throwing 75 or more heads is at most:

$$\Pr\{\text{heads} \geq 75\} \leq \frac{\mathrm{E}[\text{heads}]}{75} = \frac{50}{75} = \frac{2}{3}.$$

Markov's Theorem says that the probability of 75 or more heads is at most $2/3$, but the actual probability is less than 1 in a billion!

This ~~These two~~ examples show that Markov's Theorem gives weak results for some well-behaved random variables; however, the theorem is actually tight for other ~~some nasty~~ examples. ~~Sup-~~ For instance, supthis is insert BB and goes to p 1126

As an example, suppose that

~~pose~~ we flip 100 fair coins and use Markov's Theorem to compute the probability of getting all heads:

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

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

~~**Proof of Markov's Theorem**~~

~~Let R be the weight of a person selected randomly and uniformly. Suppose that an average person weighs 100 pounds; that is, $\mathrm{E}[R] = 100$. What is the probability that a random person weighs at least 200 pounds?~~

~~There is insufficient information for an exact answer. However, we can safely say that the probability that $R \geq 200$ is most $1/2$. If more than half of the people weigh 200 pounds or more, then the average weight would exceed 100 pounds, even if everyone else weighed zero! Markov's Theorem gives the same result:~~

$$\Pr\{R \geq 200\} \leq \frac{\mathrm{E}[R]}{200} = \frac{100}{200} = \frac{1}{2}.$$

~~Reasoning similar to that above underlies the proof of Markov's Theorem. Since expectation is a weighted average of all the outcomes of the random variable, that is, a sum over all the variables the random variable can assume, we can give a lower bound on the expectation by removing some of the terms from the sum defining the expectation; this new sum can then be modified into an expression involving the probability of an event in the tail $[R \geq x]$.~~

■

Proof. For any $x > 0$

$$
\begin{aligned}
\overset{\text{Ex}}{\mathrm{E}}[R] &::= \sum_{y \in \text{range}(R)} y \Pr\{R = y\} \\
&\geq \sum_{\substack{y \geq x, \\ y \in \text{range}(R)}} y \Pr\{R = y\} && \text{(because } R \geq 0\text{)} \\
&\geq \sum_{\substack{y \geq x, \\ y \in \text{range}(R)}} x \Pr\{R = y\} \\
&= x \sum_{\substack{y \geq x, \\ y \in \text{range}(R)}} \Pr\{R = y\} \\
&= x \Pr\{R \geq x\}. && \cancel{(17.1)}
\end{aligned}
$$

Hence, $\Pr[R \geq x] \leq \dfrac{\text{Ex}[R]}{x}$, as claimed. □

~~Dividing the first and last expression (17.1) by x gives the desired result.~~ ■

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

Corollary 17.2.4. *If R is a nonnegative random variable, then for all $c \geq 1$*

Ex

$$\Pr\{R \geq c \cdot \mathrm{E}[R]\} \leq \frac{1}{c}. \qquad (17.2)$$

Proof: Set $x = c \,\mathrm{Ex}[R]$ in Theorem 17.2.3. ☐

~~This Corollary follows immediately from Markov's Theorem(17.2.3) by letting x be $c \cdot \mathrm{E}[R]$.~~

— INSERT BB goes here —
(it is text on p 1123)

— INSERT B goes here —

17.2.1 Applying ~~Markov's Theorem~~

Hat Check, Revisited ← sub~~sub~~section

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

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

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

~~For example, there is no better than a 20% chance that 5 men get the right hat, regardless~~

INSERT B

The Chinese Appetizer Problem

Suppose that n people are seated at a circular table and that each person has an ~~out~~ ~~a unique~~ appetizer in front of them on a rotating Chinese banquet tray. Just as everyone is about to dig in, some joker spins the tray so that each person receives a random appetizer. ~~If~~ ~~the n appetizers are all different, what is the~~ ~~each per~~ ~~the probabi~~ ~~probability~~

we are interested in the number of people R that get the~~ir~~ same appetizer as before, assuming that the n appetizers are all different. ~~This similar to the Hat-check problem that we studied in Section 16 17.3.2 (check #). In fact, if the hats were placed on the banquet tray, the problems would be identical.~~

Each person ~~xx~~ gets their original appetizer with probability $\frac{1}{n}$. Hence, by Linearity of Expectation,

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

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

$$\begin{aligned} \Pr[R = n] &= \Pr[R \geq n] \\ &\leq \frac{\text{Ex}[R]}{n} \\ &= \frac{1}{n}. \end{aligned}$$

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

The Chinese Appetizer problem is similar to the Hat Check problem that we studied in Section 17.3.2 (check #), except that no distribution was specified in the Hat Check problem — we were told only that ~~each~~ each person gets their correct hat back with probability 1/n. If the hats are scrambled according to a random permutation, then the probability that everyone gets the right hat back is 1/n!, which is much less than the 1/n upper bound given by Markov's Theorem. So in this case, ~~Markov's~~ the bound given by Markov's Theorem is not ~~so~~ close to the actual probability.

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

$$\Pr[R \geq 2] \leq \frac{\mathrm{Ex}[R]}{2}$$
$$= \frac{1}{2}.$$

~~For the Chinese A~~

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

— INSERT BC goes here —
(it is text on pp 1130-1131)

[1] Proving this requires some effort.

The Chines Appetizer Problem ← sabsubsection

Suppose that there are n people seated at a Chinese restaurant and that each person has an appetizer. ~~at their seat~~ Also suppose th

of the number of people at the ~~dinner~~ party.

~~The Chinese Appetizer problem is similar to the Hat-Check problem. In this case,~~

~~then and that the ap appetizers are~~

n people are eating appetizers arranged on a circular, rotating Chinese banquet tray. Someone then spins the tray so that each person receives ~~a random~~ appetizer. ~~What is~~ ~~the~~ probability that everyone gets the same appetizer as before?

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

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

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

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

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

18.2.3 ~~17.2.2~~ Markov's Theorem for Bounded Variables

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

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

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

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

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

More generally, we can get better bounds applying Markov's Theorem to $R - l$ instead of R for any lower bound l ~~≥ 0~~ on R, even when l is negative.

~~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.~~

~~EDITING NOTE:~~

— INSERT C goes here —

INSERT C

Theorem C1: Let R be a random variable for which $R \geq \ell$ for some $\ell \in \mathbb{R}$. Then for all $x > \ell$,

$$\Pr(R \geq x) \leq \frac{\text{Ex}[R] - \ell}{x - \ell}.$$

Proof: Define

$$T ::= R - \ell.$$

Then T is a non-negative random variable with mean

$$\begin{aligned}\text{Ex}[T] &= \text{Ex}[R - \ell] \\ &= \text{Ex}[R] - \ell.\end{aligned}$$

Hence, Markov's theorem implies that

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

The result then follows from the fact that

$$\begin{aligned}\Pr[R \geq x] &= \Pr[R - \ell \geq x - \ell] \\ &= \Pr[T \geq x - \ell]. \qquad \square\end{aligned}$$

This is insert BC and goes to p. B-3

Why R Must be Nonnegative

subsubsection

Remember that Markov's Theorem applies only to nonnegative random variables! ~~The~~ Indeed, ~~following example shows that~~ the theorem is false if this restriction is removed. For example, let ~~Let~~ R be -10 with probability $1/2$ and 10 with probability $1/2$. Then ~~we have:~~

Ex

$$\mathrm{E}[R] = -10 \cdot \frac{1}{2} + 10 \cdot \frac{1}{2} = 0$$

Suppose that we now tried to compute $\Pr\{R \ge 5\}$ using Markov's Theorem:

$$\Pr\{R \ge 5\} \le \frac{\mathrm{E}[R]}{5} = \frac{0}{5} = 0.$$

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

On the other hand, we can still apply Markov's Theorem indirectly to derive a bound on the probability that an arbitrary variable like R is 5 more. Namely, given any random variable, R with expectation 0 and values ≥ -10, we can conclude that $\Pr\{R \ge 5\} \le 2/3$. To prove this fact, we define

~~end~~ end of BC

~~Proof.~~ ~~Let~~ $T ::= R + 10$. Then, ~~Now~~ T is a nonnegative random variable with expectation Ex $\mathrm{E}[R + 10] =$ Ex $\mathrm{E}[R] + 10 = 10$, so Markov's Theorem applies and tells us that $\Pr\{T \geq 15\} \leq 10/15 = 2/3$. But $T \geq 15$ iff $R \geq 5$, so $\Pr\{R \geq 5\} \leq 2/3$, as claimed. ■

18.2.4 Deviations Below the Mean ← subsection

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

Theorem 17.2.5. *Let* $u \in \mathbb{R}$ ~~*b be a real number*~~ *and let R be a random variable such that* $R \leq$ ~~b~~ u. Then for ~~For~~ *all* $x <$ ~~b~~ u, ~~*we have*~~

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

— INSERT D goes here —

INSERT D

Proof: The proof is similar to that of Theorem C1.

Define

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

Then S is a non-negative random variable with mean

$$\begin{aligned} \text{Ex}[S] &= \text{Ex}[u - R] \\ &= u - \text{Ex}[R]. \end{aligned}$$

Hence, Markov's Theorem implies that

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

The result then follows from the fact that

$$\begin{aligned} \Pr[R \leq x] &= \Pr[u - S \leq x] \\ &= \Pr[S \geq u - x]. \quad \square \end{aligned}$$

Proof. The event that ~~$R \le x$ is the same as the event that $\overset{u}{\cancel{l}} - R \ge \overset{u}{\cancel{l}} - x$. Therefore:~~

$$\begin{aligned} \Pr\{R \le x\} &= \Pr\{l - R \ge l - x\} \\ &\le \frac{\mathrm{E}[l - R]}{l - x}. \qquad \text{(by Markov' Theorem)} \qquad (17.3) \end{aligned}$$

Applying Markov's Theorem in line (17.3) is permissible since ~~$l - R$ is a nonnegative random variable and $l - x > 0$.~~ ■

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

There is not enough information here to answer the question exactly, but Theorem 17.2.5 gives an upper bound. Let R be the score of a random student. Since 100 is the highest possible score, we can set $\overset{u}{\cancel{l}} = 100$ to meet the condition in the theorem that ~~$R \le l$~~ $R \le u$.

Applying Theorem 17.2.5, we find:

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

That is, at most half of the class scored 50 or worse. This makes sense; if more than half of the class scored 50 or worse, then the class average could not be 75, even if everyone else scored 100. As with Markov's Theorem, Theorem 17.2.5 often gives weak results. In fact, based on the data given, the entire class could have scored above 50.

~~■~~

~~EDITING NOTE:~~

18.2.5 *Using ~~Markov~~ Markov's Theorem To Analyze Non-Random Events* ← subsection

In the previous example~~s~~, we used a theorem about a random variable to conclude facts about non-random data. For example, we concluded that if the average score on a test is 75, then at most 1/2 the class scored 50 or worse. There is no randomness in this problem, so how can we apply Theorem 17.2.5 to reach this conclusion?

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

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

We then use Theorem 17.2.5 to conclude that $\Pr\{R \leq 50\} \leq 1/2$. To see why this means (with certainty) that at most $1/2$ of the students scored 50 or less, we observe that

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

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

18.3

~~17.3~~ Chebyshev's Theorem

— INSERT E goes here —

~~There's a really good trick for getting more mileage out of Markov's Theorem: instead of applying it to the variable, R, apply it to some function of R. One useful choice of functions to use turns out to be taking a power of $|R|$.~~

~~In particular, since $|R|^\alpha$ is nonnegative, Markov's inequality also applies to the event $[|R|^\alpha \geq x^\alpha]$. But this event is equivalent to the event $[|R| \geq x]$, so we have:~~

Lemma 17.3.1. *For any random variable R, $\alpha \in \mathbb{R}^+$, and $x > 0$,*

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

~~Rephrasing (17.3.1) in terms of the random variable, $|R - \mathrm{E}[R]|$, that measures R's deviation from its mean, we get~~

— INSERT F goes here —

INSERT E

As we have just seen, Markov's Theorem can be extended by applying it to fuctions of a random variable R such as $R - l$ and $u - R$. Even ~~more mileage~~ stronger results can be obtained by applying Markov's Theorem to powers of R. ~~[illegible]~~ ~~For example:~~

INSERT F

Proof: The event $|R| \geq x$ is the same as the event $|R|^\alpha \geq x^\alpha$. Since $|R|^\alpha$ is non-negative, the result immediately follows from Markov's theorem. □

Similarly,

$$\Pr\{|R - \overset{\text{Ex}}{\mathrm{E}}[R]| \geq x\} \leq \frac{\overset{\text{Ex}}{\mathrm{E}}[(R - \overset{\text{Ex}}{\mathrm{E}}[R])^{\alpha}]}{x^{\alpha}}. \tag{17.4}$$

~~The case when $\alpha = 2$ is turns out to be so important that numerator of the right hand side of (17.4) has been given a name:~~

this is INSERT 8B and goes to p A-3

Definition 17.3.2. The *variance*, Var $[R]$, of a random variable, R, is:

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

The restatement of ~~(17.4)~~ Equation 17.4 for $\alpha = 2$ is known as *Chebyshev's Theorem*.

Theorem 17.3.3 (Chebyshev). *Let R be a random variable and $x \in \mathbb{R}^+$. Then*

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

~~The expression $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]$ for variance is a bit cryptic; the best approach is to work through it from the inside out. The innermost expression, $R - \mathrm{E}[R]$, is precisely the~~

— INSERT 6 goes here —

INSERT 6

Proof: ~~The~~ Define

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

Then,

$$\begin{aligned}
\Pr[|R - \text{Ex}[R]| \ge x] &= \Pr[|T| \ge x] \\
&= \Pr[T^2 \ge x^2] \\
&\le \frac{\text{Ex}[T^2]}{x^2} && \text{(by Markov's Theorem)} \\
&= \frac{\text{Ex}[(R - \text{Ex}[R])^2]}{x^2} \\
&= \frac{\text{Var}[R]}{x^2}, && \text{(by Definition 17.3.2)}
\end{aligned}$$

as claimed. □

~~deviation of R above its mean. Squaring this, we obtain, $(R - \mathrm{E}[R])^2$. 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.~~

17.3.1 Variance in ~~Two Gambling Games~~

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

This is INSERT 2A — it goes to p. A-2

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

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

which game would you rather play? Which game is better financially? We have the same probability, $2/3$, of winning each game, but that does not tell the whole story. What about the expected return for each game? Let random variables A and B be the payoffs for the two games. For example,

A is 2 with probability 2/3 and -1 with probability 1/3. We can compute the expected payoff for each game as follows:

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

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

The expected payoff is the same for both games, but they are obviously very different!

~~This difference is not apparent in their expected value, but is captured by variance. We can compute the Var [A] by working "from the inside out" as follows:~~

This is insert 2C & goes top A-4

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

(eqn A7)

ZC-cont

For Game B, we have:

~~Similarly, we have for Var [B]:~~

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

The variance of Game A is 2 and the variance of Game B is more than two million! Intuitively, this means that the ~~payoff~~ result in Game A is usually close to the expected value of $1, but the ~~payoff~~ result 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~~ OK as was 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!

end of insert ZC

This is insert ZE and ~~ZC cont~~
~~2D & goes~~ goes to p A-4

18.1.2

~~17.3.2~~ Standard Deviation

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

Definition 17.3.4. The *standard deviation*, σ_R, of a random variable, R, is the square root of the variance:

Ex Ex

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

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

For example, the standard deviations for A and B are:

~~*Example* 17.3.5. The standard deviation of the payoff in Game B is:~~

$$\sigma_A = \sqrt{Var[A]} = \sqrt{2} \approx 1.41,$$

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

no #

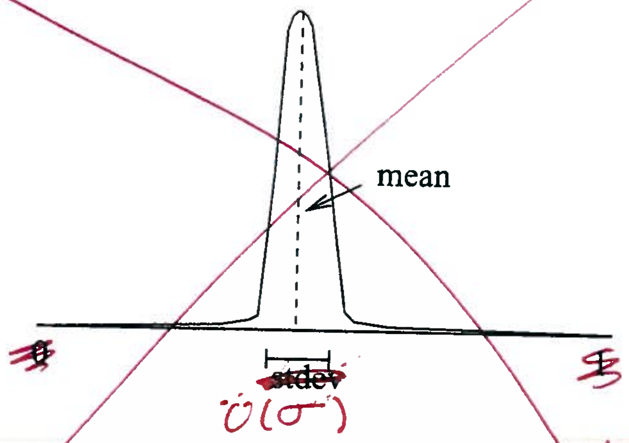
The random variable B actually deviates from the mean by either positive 1001 or negative 2002; therefore, the standard deviation of 1416 describes this situation reasonably well.

end of Insert Z5

~~Intuitively, the standard deviation measures the "width" of the "main part" of the distribution graph, as illustrated in Figure 17.4.~~

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

David: Keep this Figure, but make the indicated changes. It goes on p. I-1



OK? 17.4

~~with a bell-curve-shaped pdf~~

Figure ~~17.4~~: ~~The standard deviation of a distribution indicates~~ how wide the "~~main~~ part" ~~of it~~is.

Figure 17.4. If the pdf of a random variable is "bell-shaped," then the width of the bell is O(σ).

~~Figure I.1.~~

Corollary 17.3.6. *Let R be a random variable, and let c be a positive real number.*

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

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

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

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

■

~~**The IQ Example**~~

~~For~~ As an example, suppose

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

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

We have already seen that Markov's Theorem 17.2.3 gives a coarse bound, namely,

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

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

$$\Pr\{R \geq 300\} = \Pr\{|R - 100| \geq 200\} \leq \frac{\mathrm{Var}[R]}{200^2} = \frac{10^2}{200^2} = \frac{1}{400}. \qquad \text{(eqn I3)}$$

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

— INSERT I goes here

~~17.4 Properties of Variance~~

~~The definition of variance of R as $\mathrm{E}[(R - \mathrm{E}[R])^2]$ may seem rather arbitrary.~~

INSERT I

More generally, Corollary 17.3.6 tells us that a random variable is never likely to stray by more than a few standard deviations from its mean. For example, plugging $c=3$ into Corollary 17.3.6, we find that the probability that a random variable strays from the mean by more than 3σ is at most $1/9$.

This fact has a nice pictorial characterization for pdf's with a "bell-curve" shape; namely, the width of the bell is $O(\sigma)$, as shown in Figure 17.4.

18.3.1 ~~#~~ Bounds on ~~tt~~ ~~Sided~~ one-sided Errors

~~In the IQ~~

Corollary 17.3.6 ~~shows~~ gives bounds on the probability of deviating from the mean in either direction. If you only care about deviations in one direction, as was the case in the IQ example, then slightly better bounds can be obtained.

Theorem I2: For any random variable R and any $c > 0$,

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

and

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

The proof of Theorem I2 is trickier than the proof of Chebyshev's Theorem and we will not give the details here. ~~In general~~ In fact, the bounds in Theorem I2 are the best bounds that you ~~a~~ can obtain if you know only the mean and ~~variance~~ standard deviation of the random variable R.

~~As an example, Theorem~~

Referring to the IQ example, Theorem I2 tells us that

$$\Pr[R \geq 300] \leq \Pr[R - 100 \geq 20\sigma_R]$$

$$\leq \frac{1}{401},$$

which is a very slight improvement over Equation I3.

As another example, suppose we give an exam. What fraction of the class can score more than

~~what~~
2 standard deviations from the average?
If R is ~~a~~ the score of a random student, then
~~the answer is~~

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

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

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

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

This is insert 2D and goes to p A-4

~~EDITING NOTE~~

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

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

$$\overset{\text{Ex}}{\mathrm{E}}[R - \overset{\text{Ex}}{\mathrm{E}}[R]] = \overset{\text{Ex}}{\mathrm{E}}[R] - \overset{\text{Ex}}{\mathrm{E}}[\overset{\text{Ex}}{\mathrm{E}}[R]].$$

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

$$\overset{\text{Ex}}{\mathrm{E}}[R - \overset{\text{Ex}}{\mathrm{E}}[R]] = \overset{\text{Ex}}{\mathrm{E}}[R] - \overset{\text{Ex}}{\mathrm{E}}[R] = 0.$$

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

Of course, we could also prevent positive and negative deviations from canceling by taking an absolute value. In other words, we could compute ~~A direct measure of average deviation would be~~ $\overset{Ex}{\mathrm{E}}\left[\,|R - \overset{Ex}{\mathrm{E}}[R]|\,\right]$. But ~~the direct~~ this measure doesn't have the many useful properties that variance has, ~~which is what this section is about.~~ and so mathematicians went with squaring.

end of insert 3D

18.1.3 ~~17.4.1~~ ~~A Formula for Variance~~ An Alternative Formulation

This is insert ~~TAB~~ 2E & goes to PA-4

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

Lemma 17.4.1. For any random variable R,

$$\mathrm{Var}[R] = \overset{Ex}{\mathrm{E}}\left[R^2\right] - \overset{Ex^2}{\mathrm{E}^2}[R].$$

ZF cont

~~*for any random variable, R.*~~

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

~~EDITING NOTE~~: Remember that $\overset{\text{Ex}}{\mathrm{E}}[R^2]$ is generally not equal to $\overset{\text{Ex}^2}{\mathrm{E}^2}[R]$. We know the expected value of a product is the product of the expected values for independent variables, but not in general. And R is not independent of itself unless it is constant.

of Lemma [illegible] 17.4.1

Proof. Let $\mu = \overset{Ex}{\mathrm{E}}[R]$. Then

write out def...

$$
\begin{aligned}
\mathrm{Var}[R] &= \overset{Ex}{\mathrm{E}}\left[(R - \overset{Ex}{\mathrm{E}}[R])^2\right] && \text{(}\overset{\text{Definition}}{\cancel{\text{Def}}}\text{ 17.3.2 of variance)} \\
&= \overset{Ex}{\mathrm{E}}\left[(R-\mu)^2\right] && \text{(def of } \mu\text{)} \\
&= \overset{Ex}{\mathrm{E}}\left[R^2 - 2\mu R + \mu^2\right] \\
&= \overset{Ex}{\mathrm{E}}\left[R^2\right] - 2\mu\,\overset{Ex}{\mathrm{E}}[R] + \mu^2 && \text{(linearity of expectation)} \\
&= \overset{Ex}{\mathrm{E}}\left[R^2\right] - 2\mu^2 + \mu^2 && \text{(def of } \mu\text{)} \\
&= \overset{Ex}{\mathrm{E}}\left[R^2\right] - \mu^2 \\
&= \overset{Ex}{\mathrm{E}}\left[R^2\right] - \overset{Ex^2}{\mathrm{E}^2}[R]. && \text{(def of } \mu\text{)}
\end{aligned}
$$

end of insert ZF ■

~~For example, if B is a Bernoulli variable where $p ::= \Pr\{B = 1\}$, then~~

this is insert AH and goes to p A-6

Lemma 17.4.2. Let B be an indicator random variable for which Pr[B=1] = p. Then

$$\mathrm{Var}[B] = p - p^2 = p(1-p). \tag{17.5}$$

Proof. By Lemma 16.3.3, Ex E$[B] = p$. But since B only takes values 0 and 1, $B^2 = B$. So

$$\mathrm{Var}[B] = \mathrm{Ex}(B^2] - \mathrm{Ex}^2[B]$$

~~Lemma 17.4.2 follows immediately from Lemma 17.4.1.~~ ■

$$= p - p^2,$$

as claimed. □

~~17.4.2 Variance of Time to Failure~~

~~According to section 16.3.3, the mean time to failure is $1/p$ for a process that fails during any given hour with probability p. What about the variance? That is, let C be the hour of the first failure, so $\Pr\{C = i\} = (1-p)^{i-1}p$. We'd like to find a formula for Var $[C]$.~~

~~By Lemma 17.4.1,~~

$$\mathrm{Var}[C] = \mathrm{E}[C^2] - (1/p)^2 \tag{17.6}$$

so all we need is a formula for $\mathrm{E}\left[C^2\right]$:

$$\begin{aligned}\mathrm{E}\left[C^2\right] &::= \sum_{i\geq 1} i^2(1-p)^{i-1}p \\ &= p\sum_{i\geq 1} i^2 x^{i-1} \qquad \text{(where } x = 1-p\text{)}. \qquad (17.7)\end{aligned}$$

But (13.2) gives the generating function $x(1+x)/(1-x)^3$ for the nonnegative integer squares, and this implies that the generating function for the sum in (17.7) is $(1+x)/(1-x)^3$. So,

$$\begin{aligned}\mathrm{E}\left[C^2\right] &= p\,\frac{(1+x)}{(1-x)^3} \qquad \text{(where } x = 1-p\text{)} \\ &= p\,\frac{2+p}{p^3} \\ &= \frac{1-p}{p^2} + \frac{1}{p^2}, \qquad (17.8)\end{aligned}$$

Combining (17.6) and (17.8) gives a simple answer:

$$\mathrm{Var}\left[C\right] = \frac{1-p}{p^2}. \qquad (17.9)$$

~~It's great to be able to apply generating function expertise to knock off equation (17.9) mechanically just from the definition of variance, but there's a more elementary, and memorable, alternative. In section 16.3.3 we used conditional expectation to find the mean time to failure, and a similar approach works for the variance. Namely, the expected value of C^2 is the probability, p, of failure in the first hour times 1^2, plus $(1-p)$ times the expected value of $(C+1)^2$. So~~

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

~~which directly simplifies to (17.8).~~

~~**EDITING NOTE:**~~

~~Lemma 17.4.1 also gives a convenient way to compute the variance of a random variable: find the expected value of the square and subtract the square of the expected value. For~~

This is INSERT A-I
& goes top A-8

e

example, we can compute the variance of the outcome of a fair die as follows:

$$\overset{Ex}{\mathrm{E}}\left[R^2\right] = \frac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2) = \frac{91}{6},$$

$$\overset{Ex^2}{\mathrm{E}^2}\left[R\right] = \left(3\frac{1}{2}\right)^2 = \frac{49}{4},$$

$$\mathrm{Var}\left[R\right] = \overset{Ex}{\mathrm{E}}\left[R^2\right] - \overset{Ex^2}{\mathrm{E}^2}\left[R\right] = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.$$

~~This result is particularly useful when we want to estimate the variance of a random variable from a sequence $x_1, x_2, \ldots, x_n$, of sample values of the variable.~~

~~**Definition.** For any sequence of real numbers $x_1, x_2, \ldots, x_n$, define the *sample mean*, μ_n, and the *sample variance*, v_n, of the sequence to be:~~

$$\mu_n ::= \frac{\sum_{i=1}^n x_i}{n}.$$

$$v_n ::= \frac{\sum_{i=1}^n (x_i - \mu_n)^2}{n}.$$

~~Notice that if we define a random variable, R, which is equally likely to take each of the values in the sequence, that is $\Pr\{R = x_i\} = 1/n$ for $i = 1, \ldots, n$, then $\mu_n = \mathrm{E}[R]$~~

~~and $v_n = \text{Var}[R]$. So Lemma 17.4.1 applies to R and lets us conclude that~~

$$v_n = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2. \tag{17.10}$$

~~This leads to a simple procedure for computing the sample mean and variance while reading the sequence $x_1, \ldots, x_n$ from left to right. Namely, maintain a sum of all numbers seen and also maintain a sum of the squares of all numbers seen. That is, we store two values, starting with the values x_1 and x_1^2. Then, as we get to the next number, x_i, we add it to the first sum and add its square, x_i^2, to the second sum. After a single pass through the sequence $x_1, \ldots, x_n$, we wind up with the values of the two sums $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$. Then we just plug these two values into (17.10) to find the sample variance.~~

■

~~**EDITING NOTE:**~~

this is insert A6 and goes to p A-5

~~Expectation Squared~~

The alternate ~~definition~~ formulation of variance given in Lemma 17.4.1 has a cute implication:

Corollary 17.4.3. *If R is a random variable, then* $\overset{\text{Ex}}{\mathrm{E}}[R^2] \geq \overset{\text{Ex}^2}{\mathrm{E}^2}[R]$.

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

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

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

This happens ~~when~~ precisely when ~~E~~

$$\Pr[R = \mathrm{Ex}[R]] = 1,$$

~~EDITING NOTE:~~

namely when R is a constant.[1]

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

Zero Variance ← sub subsection

When does a random variable, R, have zero variance? ... When the random variable *never* deviates from the mean!

Lemma. *The variance of a random variable, R, is zero if and only if* $\Pr\{R = \overset{Ex}{\mathrm{E}}[R]\} = 1$.

So saying that $\mathrm{Var}[R] = 0$ is almost the same as saying that R is constant. Namely, it takes the constant value equal to its expectation on all sample points with nonzero probability. (It can take on any finite values on sample points with zero probability without affecting the variance.)

of Lemma #

Proof. By the definition of variance,

$$\mathrm{Var}[R] = 0 \quad \text{iff} \quad \overset{Ex}{\mathrm{E}}\left[(R - \overset{Ex}{\mathrm{E}}[R])^2\right] = 0.$$

The inner expression on the right, $(R - \overset{Ex}{\mathrm{E}}[R])^2$, is always nonnegative because of the square. As a result, $\overset{Ex}{\mathrm{E}}[(R - \overset{Ex}{\mathrm{E}}[R])^2] = 0$ if and only if $\Pr\{(R - \overset{Ex}{\mathrm{E}}[R])^2 \neq 0\}$ is zero,

Ex

~~which is the same as saying that $\Pr\{(R - \mathrm{E}[R])^2 = 0\}$ is one. That is,~~

$$\mathrm{Var}[R] = 0 \text{ IFF } \Pr\{(R - \mathrm{E}[R])^2 = 0\} = 1.$$

~~But the $(R - \mathrm{E}[R])^2 = 0$ and $R = \mathrm{E}[R]$ are different descriptions of the same event.~~

~~Therefore,~~

$$\mathrm{Var}[R] = 0 \quad \text{iff} \quad \Pr\{R = \mathrm{E}[R]\} = 1.$$

This is INSERT AJ and goes to page A-10

18.1.6 ~~17.4.3~~ Dealing with Constants

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

Theorem 17.4.4. *Let R be a random variable, and* a and b be constants. ~~*a a constant.*~~ *Then*

$$\mathrm{Var}[aR+b] = a^2\,\mathrm{Var}[R]\,. \tag{17.11}$$

A J cont

Proof. Beginning with ~~the definition of variance~~ Lemma 17.4.1 and repeatedly applying linearity of expectation, we have:

$$\mathrm{Var}[aR+b] = \mathrm{Ex}[(aR+b)^2] - \mathrm{Ex}^2[aR+b]$$

~~$\mathrm{Var}[aR] ::= \mathrm{E}[(aR - \mathrm{E}[aR])^2]$~~

$$= \mathrm{Ex}[a^2R^2 + \cancel{a}2abR + b^2] - (a\,\mathrm{Ex}[R]+b)^2$$

~~$= \mathrm{E}[(aR)^2 - 2aR\,\mathrm{E}[aR] + \mathrm{E}^2[aR]]$~~

$$= \cancel{\mathcal{E}}\, a^2\,\mathrm{Ex}[R^2] + 2ab\,\mathrm{Ex}[R] + b^2 - a^2\mathrm{Ex}^2[R]^{\varepsilon} - 2ab\,\mathrm{Ex}[R] - b^2$$

~~$= \mathrm{E}[(aR)^2] - \mathrm{E}[2aR\,\mathrm{E}[aR]] + \mathrm{E}^2[aR]$~~

~~$= a^2\,\mathrm{E}[R^2] - 2\,\mathrm{E}[aR]\,\mathrm{E}[aR] + \mathrm{E}^2[aR]$~~

$$= a^2\,\overset{\mathrm{Ex}}{\mathrm{E}}[R^2] - a^2\overset{\mathrm{Ex}^2}{\mathrm{E}^2}[R]$$

$$= a^2(\overset{\mathrm{Ex}}{\mathrm{E}}[R^2] - \overset{\mathrm{Ex}^2}{\mathrm{E}^2}[R])$$

$$= a^2\,\mathrm{Var}[R] \quad \text{(by Lemma 17.4.1)}$$

■

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

AJ cont

David: change all refs to 17.4.5 to point to 17.4.4

~~**Theorem 17.4.5.** *Let R be a random variable, and b a constant. Then*~~

~~$\text{Var}[R+b] = \text{Var}[R]$. (17.12)~~

~~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 17.4.6.

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

~~17.4.4~~ 18.1.7 Variance of a Sum

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

end of Insert AJ

Theorem 17.4.7. *If R_1 and R_2 are independent random variables, then*

$$\text{Var}[R_1 + R_2] = \text{Var}[R_1] + \text{Var}[R_2]. \tag{17.13}$$

x

~~*Proof.* We may assume that $\text{E}[R_i] = 0$ for $i = 1, 2$, since we could always replace R_i by $R_i - \text{E}[R_i]$ in equation (17.13). This substitution preserves the independence of the variables, and by Theorem~~ 17.4.4 ~~17.4.5, does not change the variances.~~

~~Now by Lemma 17.4.1, $\text{Var}[R_i] = \text{E}[R_i^2]$ and $\text{Var}[R_1 + R_2] = \text{E}[(R_1 + R_2)^2]$, so we need only prove~~

$$\cancel{\text{E}[(R_1 + R_2)^2] = \text{E}[R_1^2] + \text{E}[R_2^2].} \tag{17.14}$$

~~But (17.17) follows from linearity of expectation and the fact that~~

$$\cancel{\text{E}[R_1 R_2] = \text{E}[R_1]\,\text{E}[R_2]} \tag{17.15}$$

since R_1 and R_2 are independent:

$$
\begin{aligned}
\mathrm{E}\left[(R_1+R_2)^2\right] &= \mathrm{E}\left[R_1^2+2R_1R_2+R_2^2\right] \\
&= \mathrm{E}\left[R_1^2\right]+2\,\mathrm{E}\left[R_1R_2\right]+\mathrm{E}\left[R_2^2\right] \\
&= \mathrm{E}\left[R_1^2\right]+2\,\mathrm{E}\left[R_1\right]\mathrm{E}\left[R_2\right]+\mathrm{E}\left[R_2^2\right] && \text{(by (17.18))} \\
&= \mathrm{E}\left[R_1^2\right]+2\cdot 0\cdot 0+\mathrm{E}\left[R_2^2\right] \\
&= \mathrm{E}\left[R_1^2\right]+\mathrm{E}\left[R_2^2\right]
\end{aligned}
$$

This is insert AK and goes to pA-11.

Theorem 17.4.7 does not necessarily hold if R_1 and R_2 ~~are~~

An independence condition is necessary. If we ignored independence, then we would conclude that Var $[R+R]$ = Var $[R]$ + Var $[R]$. However, by Theorem 17.4.4, the left side is equal to 4 Var $[R]$, whereas the right side is 2 Var $[R]$. This implies that Var $[R]$ = 0, which, by the Lemma above, essentially only holds if R is constant.

this is insert AK and goes to p A-11

The proof of Theorem 17.4.7 carries over straightforwardly to the sum of any finite number of variables. ~~So we have:~~

Theorem 17.4.8. *[Pairwise Independent Additivity of Variance] If $R_1, R_2, \ldots, R_n$ are* pairwise *independent random variables, then*

$$\text{Var}[R_1 + R_2 + \cdots + R_n] = \text{Var}[R_1] + \text{Var}[R_2] + \cdots + \text{Var}[R_n]. \tag{17.16}$$

~~EDITING NOTE:~~

~~*Proof.* We may assume that $\text{E}[R_i] = 0$ for $i = 1, \ldots, n$, since we could always replace R_i by $(R_i - \text{E}[R_i])$ in equation (17.16). This substitution preserves the independence of the variables, and by Theorem 17.4.5, does not change the variances.~~

~~Now by Lemma 17.4.1, $\text{Var}[R_i] = \text{E}[R_i^2]$ and~~

$$\text{Var}[R_1 + R_2 + \cdots + R_n] = \text{E}[(R_1 + R_2 + \cdots + R_n)^2].$$

so we need only prove

$$\mathrm{E}\left[(R_1 + R_2 + \cdots + R_n)^2\right] = \mathrm{E}\left[R_1^2\right] + \mathrm{E}\left[R_2^2\right] + \cdots + \mathrm{E}\left[R_n^2\right] \tag{17.17}$$

But (17.17) follows from linearity of expectation and the fact that

$$\mathrm{E}\left[R_i R_j\right] = \mathrm{E}\left[R_i\right] \mathrm{E}\left[R_j\right] = 0 \cdot 0 = 0 \tag{17.18}$$

for $i \neq j$, since R_i and R_j are independent. Namely,

$$\begin{aligned}
\mathrm{E}\left[(R_1 + R_2 + \cdots + R_n)^2\right] &= \mathrm{E}\left[\sum_{1 \le i,j \le n} R_i R_j\right] \\
&= \sum_{1 \le i,j \le n} \mathrm{E}\left[R_i R_j\right] && \text{linearity of E []} \\
&= \sum_{1 \le i \le n} \mathrm{E}\left[R_i^2\right] + \sum_{1 \le i \neq j \le n} \mathrm{E}\left[R_i R_j\right] && \text{(rearranging the sum)} \\
&= \sum_{1 \le i \le n} \mathrm{E}\left[R_i^2\right] + \sum_{1 \le i \neq j \le n} 0 && \text{(by (17.18))} \\
&= \mathrm{E}\left[R_1^2\right] + \mathrm{E}\left[R_2^2\right] + \cdots + \mathrm{E}\left[R_n^2\right].
\end{aligned}$$

■

■

~~Now we have a simple way of computing the variance of a variable, J, that has an (n, p)-binomial distribution. We know that $J = \sum_{k=1}^{n} I_k$ where the I_k are mutually independent indicator variables with $\Pr\{I_k = 1\} = p$. The variance of each I_k is $p(1-p)$ by Lemma 17.4.2, so by linearity of variance, we have~~

this is insert AL and goes to p. A-12

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

$$\mathrm{Var}[J] = \cancel{n \,\mathrm{Var}[I_k] =} np(1-p). \tag{17.19}$$

~~17.5 Estimation by Random Sampling~~

~~***Polling again***~~

~~EDITING NOTE:~~

David: put all of 17.5 & 17.6 into a holding bin for possible use in ch 16. This means we remove from here to p 1180 from ch 18.

This paragraph reflects an alternative exposition where polling estimation and confidence were based only on binomial distribution properties, even before expectation was introduced.

In Chapter [none], we used bounds on the binomial distribution to determine confidence levels for a poll of voter preferences of Franken vs. Coleman. Now that we know the variance of the binomial distribution, we can use Chebyshev's Theorem as an alternative approach to calculate poll size.

The setup is the same as in Chapter [none]

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

Let p be this unknown fraction, and let's suppose we have some random process — say throwing darts at voter registration lists —which will select each voter with equal probability. We can define a Bernoulli variable, K, by the rule that $K = 1$ if the random voter most prefers Brown, and $K = 0$ otherwise.

Now to estimate p, we take a large number, n, of random choices of voters[1] and count the fraction who favor Brown. That is, we define variables $K_1, K_2, \ldots,$ where K_i is interpreted to be the indicator variable for the event that the ith chosen voter prefers Brown. Since our choices are made independently, the K_i's are independent. So formally, we model our estimation process by simply assuming we have mutually independent Bernoulli variables $K_1, K_2. \ldots,$ each with the same probability, p, of being

[1] We're choosing a random voter n times *with replacement*. That is, we don't remove a chosen voter from the set of voters eligible to be chosen later; so we might choose the same voter more than once in n tries! We would get a slightly better estimate if we required n *different* people to be chosen, but doing so complicates both the selection process and its analysis, with little gain in accuracy.

equal to 1. Now let S_n be their sum, that is,

$$S_n ::= \sum_{i=1}^{n} K_i. \tag{17.20}$$

So S_n has the binomial distribution with parameter n, which we can choose, and unknown parameter p.

The variable S_n/n describes the fraction of voters we will sample who favor Scott Brown. Most people intuitively expect this sample fraction to give a useful approximation to the unknown fraction, p —and they would be right. So we will use the sample value, S_n/n, as our *statistical estimate* of p and use the Pairwise Independent Sampling Theorem 17.5.1 to work out how good an estinate this is.

17.5.1 Sampling

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

$$\Pr\left\{\left|\frac{S_n}{n} - p\right| \le 0.04\right\} \ge 0.95\,. \tag{17.21}$$

So we better determine the number, n, of times we must poll voters so that inequality (17.21) will hold.

EDITING NOTE: the value, S_n/n, of our estimate will, with probability at least $1-\delta$, be within ϵ of the actual fraction in the nation favoring Brown.

We let ϵ be the margin of error we can tolerate, and let δ be the probability that our result lies outside this margin, so in this case we'd have $\epsilon = 0.04$ and $\delta \le 0.05$.

We want to determine the number, n, of times we must poll voters so that the value, S_n/n, of our estimate will, with probability at least $1-\delta$, be within ϵ of the actual fraction in the nation favoring Brown. ■

Now S_n is binomially distributed, so from (17.19) we have

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

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

Next, we bound the variance of S_n/n:

$$\begin{aligned}
\text{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \text{Var}\left[S_n\right] && \text{(by (17.11))} \\
&\leq \left(\frac{1}{n}\right)^2 \frac{n}{4} && \text{(by (17.5.1))} \\
&= \frac{1}{4n} && (17.22)
\end{aligned}$$

Now from Chebyshev and (17.22) we have:

$$\Pr\left\{\left|\frac{S_n}{n} - p\right| \geq 0.04\right\} \leq \frac{\text{Var}\left[S_n/n\right]}{(0.04)^2} = \frac{1}{4n(0.04)^2} = \frac{156.25}{n} \qquad (17.23)$$

To make our our estimate with 95% confidence, we want the righthand side of (17.23) to be at most 1/20. So we choose n so that

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

that is,

$$n \geq 3,125.$$

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

17.5.2 Matching Birthdays

There are important cases where the relevant distributions are not binomial because the mutual independence properties of the voter preference example do not hold. In these cases, estimation methods based on the Chebyshev bound may be the best approach. Birthday Matching is an example. We already saw in Section 14.5 that in a class of 85 students it is virtually certain that two or more students will have the same birthday. This suggests that quite a few pairs of students are likely to have the same birthday. How many?

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

Unlike the situation with voter preferences, having matching birthdays for different pairs of students are not mutually independent events, but the matchings are *pairwise independent*, as explained in Section 14.5. as we did for voter preference. Namely, let $B_1, B_2, \dots, B_n$ be the birthdays of n independently chosen people, and let $E_{i,j}$ be the indicator variable for the event that the ith and jth people chosen have the same birthdays, that is, the event $[B_i = B_j]$. So our probability model, the B_i's are mutually independent variables, the $E_{i.j}$'s are pairwise independent. Also, the expectations of $E_{i.j}$ for $i \neq j$ equals the probability that $B_i = B_j$, namely, $1/d$.

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

$$D ::= \sum_{1 \le i < j \le n} E_{i,j}. \tag{17.24}$$

So by linearity of expectation

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

Similarly,

$$\begin{aligned}
\mathrm{Var}[D] &= \mathrm{Var}\left[\sum_{1 \le i < j \le n} E_{i,j}\right] \\
&= \sum_{1 \le i < j \le n} \mathrm{Var}[E_{i,j}] && \text{(by Theorem 17.4.8)} \\
&= \binom{n}{2} \cdot \frac{1}{d}\left(1 - \frac{1}{d}\right). && (\textit{byLemma } 17.4.2)
\end{aligned}$$

In particular, for a class of $n = 85$ students with $d = 365$ possible birthdays, we have $\mathrm{E}[D] \approx 9.7$ and $\mathrm{Var}[D] < 9.7(1 - 1/365) < 9.7$. So by Chebyshev's Theorem

$$\Pr\{|D - 9.7| \ge x\} < \frac{9.7}{x^2}.$$

Letting $x = 5$, we conclude that there is a better than 50% chance that in a class of 85 students, the number of pairs of students with the same birthday will be between 5 and 14.

17.5.3 Pairwise Independent Sampling

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

Theorem 17.5.1 (Pairwise Independent Sampling). *Let $G_1, \ldots, G_n$ be pairwise indepen-*

dent variables with the same mean, μ, and deviation, σ. Define

$$S_n ::= \sum_{i=1}^{n} G_i. \tag{17.25}$$

Then

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

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

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

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

$$\begin{aligned}
\text{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \text{Var}\left[S_n\right] && \text{(by (17.11))}\\
&= \frac{1}{n^2}\,\text{Var}\left[\sum_{i=1}^{n} G_i\right] && \text{(def of } S_n\text{)}\\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\left[G_i\right] && \text{(pairwise independent additivity)}\\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. && (17.26)
\end{aligned}$$

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

$$\begin{aligned}
\Pr\left\{\left|\frac{S_n}{n} - \mu\right| \geq x\right\} &\leq \frac{\text{Var}\left[S_n/n\right]}{x^2}. && \text{(Chebyshev's bound)}\\
&= \frac{\sigma^2/n}{x^2} && \text{(by (17.26))}\\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

■

The Pairwise Independent Sampling Theorem provides a precise general statement about how the average of independent samples of a random variable approaches the mean. In particular, it proves what is known as the Law of Large Numbers[2] : by choosing a large enough sample size, we can get arbitrarily accurate estimates of the mean with confidence arbitrarily close to 100%.

Corollary 17.5.2. *[Weak Law of Large Numbers] Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean, μ, and the same finite deviation, and let*

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

Then for every $\epsilon > 0$,

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

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

17.6 Confidence versus Probability

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

EDITING NOTE: Estimates of the binomial distribution show that a sample size around 664 would do.

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

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

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

What's objectionable about this statement is that it talks about the probability or "chance" that a real world fact is true, namely that the actual fraction, p, of voters favoring Brown is more than $1/3$. But p is what it is, and it simply makes no sense to talk about the probability that it is something else. For example, suppose p is actually 0.3; then it's nonsense to ask about the probability that it is within 0.04 of 1250/3125 —it simply isn't.

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 $1250/3125 \pm 0.04$.

end of material going into holding bin for CH16

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

EDITING NOTE: Maybe include example from CP_drug_confidence here.

— INSERT J goes here —

~~17.7 The Chernoff Bound~~

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

Like all major web services, Fussbook has a load balancing problem. Specifically, Fussbook receives 24,000 forum posts every 10 minutes. Each post is assigned to one of m computers for processing, and each computer works sequentially through its assigned tasks. Processing an average post takes a computer $1/4$ second. Some posts, such

18.4 ~~Better~~ Bounds for Sums of Random Variables

If all you know about a random variable are its mean and variance, then ~~Chebyshev bounds~~ Chebyshev's theorem is the best you can do when it comes to bounding the probability that the random variable ~~stays~~ deviates from its mean. In some cases, however, we know more ~~and~~ — for example, that the random variable has a binomial distribution — and then it is possible to prove much stronger bounds. ~~On the probability~~ Instead of polynomially small bounds such as $1/c^2$, we can ~~attain~~ sometimes even attain exponentially small bounds such as ~~$1/e^c$~~ ~~[illegible]~~ ~~[illegible]~~ ~~Such~~ $1/e^c$. ~~Such~~ As we will soon discover, this is the case whenever ~~when~~ the random variable T is the sum of n mutually independent random variables $T_1, T_2, \ldots, T_n$ where $0 \leq T_i \leq 1$. ~~We A special case~~ ~~The binomial~~ A random variable with a binomial distribution is just ~~the example of~~

~~18.4.1 A Motivat~~

~~such a distribution~~

~~Such a T.~~

just one of many examples of such a T. Here is another.

18.4.1 A Motivating Example

as pointless grammar critiques and snide witticisms, are easier. But the most protracted harangues require 1 full second.

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

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

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

~~real-w~~

known in advance, which is typical for workload problems in ~~[illegible]~~ the real world.

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

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

Some arithmetic shows that Fussbook's traffic is sufficient to keep $m = 10$ computers running at 100% capacity with perfect load balancing. Surely, more than 10 servers are needed to cope with random fluctuations in task length and imperfect load balance. But how many is enough? 11? 15? 20? 100? We'll answer that question with a new mathematical tool.

18.4.2 ~~17.7.1~~ The Chernoff Bound

The Chernoff[1] bound is a hammer that you can use to nail a great many problems. Roughly, the Chernoff bound says that certain random variables are very unlikely to significantly exceed their expectation. For example, if the expected load on a computer is just a bit below its capacity, then that computer is unlikely to be overloaded, provided the conditions of the Chernoff bound are satisfied.

More precisely, the Chernoff Bound says that *the sum of lots of little, independent random variables is unlikely to significantly exceed the mean*. The Markov and Chebychev bounds lead to the same kind of conclusion but typically provide much weaker conclusions.

~~EDITING NOTE:~~ In particular, the Markov and Chebychev bounds are polynomial, while the Chernoff bound is exponential.

Here is the theorem. The proof ~~is at the end of the chapter.~~ will come later in Section 18.4.4. ~~will come later~~ ~~is at the end of the section.~~

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

Theorem 17.7.1 (Chernoff Bound). *Let $T_1, \ldots T_n$ be mutually independent random variables such that $0 \le T_i \le 1$ for all i. Let $T = T_1 + \cdots + T_n$. Then for all $c \ge 1$,*

$$\Pr\{T \ge c\,\mathrm{E}[T]\} \le e^{-k\,\mathrm{E}[T]} \tag{17.27}$$

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

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

18.4.3 More Examples

~~A Simple Example~~

The Chernoff bound is pretty easy to apply, though the details can be daunting at first. Let's walk through a simple example to get the hang of it.

What ~~are the odds~~ is the probability that the number of heads that come up in 1000 independent tosses of a fair coin exceeds the expectation by 20% or more? Let T_i be an indicator variable for the event that the i-th coin is heads. Then the total number of heads is $T = T_1 + \cdots + T_{1000}$. The Chernoff bound requires that the random variables T_i be mututally independent and take on values in the range $[0, 1]$. Both conditions hold here. In fact, this example is similar to many applications of the Chernoff bound in that every T_i is *either* 0 or 1, since they're indicators.

center this

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

k as defined in the theorem:

$$k = c \ln c - c + 1 = 0.0187\ldots$$

Plugging this value into the Chernoff bound gives:

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

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

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

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

which is pretty darn small.

Alternatively, the bound also becomes stronger for larger deviations. For example, suppose we're interested in the odds of getting 30% or more extra heads in 1000 tosses, rather than 20%. In that case, $c = 1.3$ instead of 1.2. Consequently, the parameter k rises from 0.0187 to about 0.0410, which may seem insignificant. But because k appears in the exponent of the upper bound, the final probability decreases from around 1 in 10,000 to about 1 in a billion!

Pick-4

Pick-4 is a lottery game where you pick a 4-digit number between 0000 and 9999. If your number comes up in a random drawing, then you win $5,000. Your chance of winning is 1 in 10,000. And if 10 million people play, then the expected number of winners is 1000. The lottery operator's nightmare is that the number of winners is much greater; say, 2000 or more. ~~What are the odds of that?~~ is the probability that will happen?

Let T_i be an indicator for the event that the i-th player wins. Then $T = T_1 + \cdots + T_n$ is the total number of winners. If we assume[1] 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.

~~EDITING NOTE: Add comment about how unrealistic these assumptions are because people frequently play a few favorite numbers.~~

~~The assumptions would be plausible for a version where people buy tickets with randomly assigned numbers, so they can't pick their own number.~~

~~Now~~ since 2000 winners would be twice the expected number, ~~so~~ we choose $c = 2$, com-

[1] ~~Actually~~ As we noted in chapter 17, ~~individual picks~~ human choices are often not ~~ver~~ uniform and they can be ~~inadvertently~~ highly dependent. For example, lots of people will pick an important date. So ~~state~~ the lottery folks should not ~~ta place~~ get too much comfort from the analysis that follows, unless they are sure to assign ~~the~~ random 4-digit numbers to each player.

pute $k = c \ln c - c + 1 = 0.386\ldots$, and plug these values into the Chernoff bound:

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

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

$$\begin{aligned}
\Pr\{T \geq 1.1\,\mathrm{E}[T]\} &\leq e^{-k\,\mathrm{E}[T]} \\
&= e^{-}0.00484 \cdots * 1000 \\
&< 0.01
\end{aligned}$$

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

~~17.7.2~~ Randomized Load Balancing ← sub sub section

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

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

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

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

$$\overset{Ex}{\mathrm{E}}[T] = \frac{24,000 \text{ tasks} \cdot 1/4 \text{ second per task}}{m \text{ machines}}$$

$$= 6000/m \text{ seconds}$$

For example, if there are $m = 10$ machines, then the expected load on the first server is 600 seconds, which is 100% of its capacity.

Now we can use the Chernoff bound to upper bound the probability that the first server is overloaded:

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

Equality holds on the first line when $c = m/10$, since $c\,\mathrm{E}[T] = (m/10)\cdot(6000/m) = 600$.

The probability that *some* server is overloaded is at most m times the probability that the first server is overloaded: by the union bounds so

$\leq \sum_{i=1}^{m} \Pr[$ ~~the~~ server i is overloaded$]$

$$\Pr\{\text{some server is overloaded}\} \leq m e^{-(c\ln c - c + 1)\cdot 6000/m}$$

$= m \Pr[$the first server is overloaded$]$

$\leq$

Some values of this upper bound are tabulated below:

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

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

18.4.4 ~~17.7.3~~ Proof of the Chernoff Bound

The proof of the Chernoff bound is somewhat involved. Heck, even *Chernoff* didn't come up with it! His friend, Herman Rubin, showed him the argument. Thinking the bound not very significant, Chernoff did not credit Rubin in print. He felt pretty bad when it became famous![1]

~~EDITING NOTE:~~ 1 See ~~References~~: "A Conversation with Herman Chernoff", Statistical Science 1996, Vol 11, No 4, pp 335-350.

Here is the theorem again, for reference:

Theorem 17.7.2 (Chernoff Bound). *Let $T_1, \ldots T_n$ be mutually independent random variables such that $0 \le T_i \le 1$ for all i. Let $T = T_1 + \cdots + T_n$. Then for all $c \ge 1$,*

$$\Pr\{T \ge c\,\mathrm{E}[T]\} \le e^{-k\,\mathrm{E}[T]}$$

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

1

For clarity, we'll go through the proof "top down"; that is, we'll use facts that are proved immediately afterward.

Proof. The key step is to exponentiate both sides of the inequality $T \underline{>} c\,\overset{Ex}{\mathrm{E}}[T]$ and then apply the Markov bound.

$$
\begin{aligned}
\Pr\{T \geq c\,\overset{Ex}{\mathrm{E}}[T]\} &= \Pr\left\{c^T \geq c^{c\,\overset{Ex}{\mathrm{E}}[T]}\right\} \\
&\leq \frac{\overset{Ex}{\mathrm{E}}\left[c^T\right]}{c^{c\,\mathrm{E}[T]}} && \text{(by Markov)} \\
&\leq \frac{e^{(c-1)\,\mathrm{E}[T]}}{c^{c\,\mathrm{E}[T]}} \\
&= e^{-(c\ln c - c + 1)\,\mathrm{E}[T]}
\end{aligned}
$$

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

$$\mathrm{E}\left[c^T\right] \leq e^{(c-1)\,\mathrm{E}[T]}$$

which is proved below in Lemma 17.7.3. The final step is simplification, using the fact ~~(Recall~~ that c^c is equal to $e^{c \ln c}$.~~)~~ ■

Algebra aside, there is a brilliant idea in this proof: in this context, exponentiating somehow supercharges the Markov bound. This is not true in general! One unfortunate side-effect is that we have to bound some nasty expectations involving exponentials in order to complete the proof. This is done in the two lemmas below, where variables take on values as in Theorem 17.7.1.

Lemma 17.7.3.

$$\overset{Ex}{\mathrm{E}}\left[c^T\right] \le e^{(c-1)\,\overset{Ex}{\mathrm{E}}[T]}$$

Proof.

$$
\begin{aligned}
\overset{\text{Ex}}{\mathrm{E}}\left[c^{T}\right] &= \overset{\text{Ex}}{\mathrm{E}}\left[c^{T_1+\cdots+T_n}\right] \\
&= \overset{\text{Ex}}{\mathrm{E}}\left[c^{T_1}\cdots c^{T_n}\right] \\
&= \overset{\text{Ex}}{\mathrm{E}}\left[c^{T_1}\right]\cdots \overset{\text{Ex}}{\mathrm{E}}\left[c^{T_n}\right] \\
&\le e^{(c-1)\overset{\text{Ex}}{\mathrm{E}}[T_1]}\cdots e^{(c-1)\overset{\text{Ex}}{\mathrm{E}}[T_n]} \\
&= e^{(c-1)(\overset{\text{Ex}}{\mathrm{E}}[T_1]+\cdots+\overset{\text{Ex}}{\mathrm{E}}[T_n])} \\
&= e^{(c-1)\overset{\text{Ex}}{\mathrm{E}}[T_1+\cdots+T_n]} \\
&= e^{(c-1)\overset{\text{Ex}}{\mathrm{E}}[T]}
\end{aligned}
$$

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

$$\overset{E*}{\mathrm{E}}\left[c^{T_i}\right] \le e^{(c-1)\,\overset{E*}{\mathrm{E}}[T_i]}$$

which is proved in Lemma 17.7.4. The last steps are simplifications using algebra and linearity of expectation. ■

Lemma 17.7.4.

$$\overset{E*}{\mathrm{E}}\left[c^{T_i}\right] \le e^{(c-1)\,\mathrm{E}[T_i]}$$

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

$$
\begin{aligned}
\overset{Ex}{\mathrm{E}}\left[c^{T_i}\right] &= \sum_{v} c^{v} \Pr\{T_i = v\} \\
&\le \sum_{v} (1 + (c-1)v) \Pr\{T_i = v\} \\
&= \sum_{v} \Pr\{T_i = v\} + (c-1)v \Pr\{T_i = v\} \\
&= \sum_{v} \Pr\{T_i = v\} + \sum_{v} (c-1)v \Pr\{T_i = v\} \\
&= 1 + (c-1) \sum_{v} v \Pr\{T_i = v\} \\
&= 1 + (c-1)\, \overset{Ex}{\mathrm{E}}[T_i] \\
&\le e^{(c-1)\,\mathrm{E}[T_i]} \qquad .
\end{aligned}
$$

The first step uses the definition of expectation. The second step relies on the inequality $c^v \le 1 + (c-1)v$, which holds for all $\forall\, v \in$ $[0,1]$ and $c \ge 1$. This follows from the general principle that a convex function, namely c^v, is less than the linear function, $1 + (c-1)v$, between their points of intersection, namely $v = 0$ and 1. This inequality is why the variables T_i are restricted to the interval $[0, 1]$. We then multiply out inside the summation and split into two sums. The first sum adds the probabilities of all possible outcomes, so it is equal to 1. After pulling the constant $c-1$ out of the second sum, we're left with the definition of $\overset{\text{Ex}}{\mathrm{E}}[T_i]$. The final step uses the standard inequality $1+z \le e^z$, which holds for all ~~real z~~ $z \ge 0$. ■

~~EDITING NOTE: Add problems~~

— INSERT M goes here —

~~18.5 Problems~~

INSERT M

18.~~4.6~~ 4.5 Mutually Independent Events ~~Murphy's Law~~

~~Given a collection of mutually independent events $A_1, A_2, \ldots, A_n$, the Chernoff Bound~~

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 R~~ ~~By Linearity of Expectation,~~

~~Let T be the number of the events that occur. Then, by Linearity of Expectation, we know that~~

~~$Ex[T] = Pr[A_1] + Pr[A_2] + \cdots + Pr[A_n]$.~~

~~In fact, this is true even if the events are not independent.~~ ~~Moreover~~

~~Markov's Theorem tells us that~~

~~$Pr[T \geq c] \leq \frac{Ex[T]}{c}$~~

~~If th~~

~~for any~~

~~As usual, let T_i be den~~

~~Let p_i denote the probability that A_i occurs and~~

~~Let~~ Let T_i be the indicator random variable for A_i and define $p_i = \Pr[T_i = 1] = \Pr[A_i]$ for $1 \le i \le n$. Define $T = T_1 + T_2 + \cdots + T_n$ to be the number of events that occur.

~~We~~ 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.

p/ ~~From~~ By Theorem ___, we also know that

$$\mathrm{Var}[T] = \mathrm{Var}[T_1] + \mathrm{Var}[T_2] + \cdots + \mathrm{Var}[T_n]$$

~~$= p_1(1-p_1) + p_2(1-p_2) + \cdots$~~

$$= \sum_{i=1}^{n} p_i(1-p_i),$$

~~and thus that~~

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

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

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

~~$\Pr[T \geq c] \leq \frac{Ex[T]}{c}$~~

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

Chebyshev's Theorem gives us the stronger result that

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

If Ex[T] ≤ 1, this means that ~~Hence,~~ the probability that at least one event occurs is ~~at most~~

$$\Pr[T \geq 1] \leq Ex[T]$$

$$= \sum_{i=1}^{n} p_i .$$

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

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

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

By a substitution of variables, we can also use the Chernoff Bound to prove that ~~T is ex~~ the probability that T is much lower than $Ex[T]$ is also exponentially small. We conclude this chapter with a special case of this result, which we call Murphy's Law[1]:

Theorem M2 (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

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

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

Proof:

$$\Pr[T=0] = \Pr[\bar{A}_1 \wedge \bar{A}_2 \wedge \cdots \wedge \bar{A}_n]$$

~~$= \Pr[\bar{A}_1]\cdot\Pr[\bar{A}_2]\cdots\Pr[\bar{A}_n]$~~

~~$= (1-\Pr[A_1])(1-\Pr[A_2])\cdots(1-\Pr[A_n])$~~

$$= \prod_{i=1}^{n} \Pr[\bar{A}_i] \qquad \text{(by independence of } A_i\text{)}$$

$$= \prod_{i=1}^{n} (1-\Pr[A_i])$$

$$\le \prod_{i=1}^{n} e^{-\Pr[A_i]} \qquad (\text{since } \forall x,\ 1-x \le e^{-x})$$

$$= e^{-\sum_{i=1}^{n} \Pr[A_i]}$$

$$= e^{-\sum_{i=1}^{n} Ex[T_i]}$$ ~~$e^{-Ex[T]}$~~ (since T_i is an indicator ~~random~~ for A_i)

~~$= e^{-\sum_{i=1}^{n} T_i}$~~

$$= e^{-Ex[T]} \qquad \text{(linearity of expectation)} \quad \square$$

~~This means~~ For example, ~~if~~ given any set of mutually independent events, if you expect 10 ~~events~~ 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 alot of unrelated things that can go wrong and their probabilities sum to a number much greater than 1, then Theorem ML ~~s~~ proves that some of them surely will go wrong.

This result ~~helps~~ ~~explain~~ can help to explain "coincidences" or "acts of God" or ~~oth~~ 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 ~~s~~ example, someone does win the lottery. # In fact, if there are 100,000 random tickets in Pick-4, ~~the~~ Theorem ML says that the probability that there is no winner is less than ~~[illegible]~~ ~~1/22000~~ $e^{-10} < 1/22000$.

~~Class Problems~~

~~Practice Problems~~

18. ~~6~~ 5 Problems

~~Class Problems~~

~~Homework Problems~~

~~Practice Problems~~

~~Class Problems~~

~~Exam Problems~~