19 Deviation from the Mean

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

The most important reason to care about the mean value comes from its connection to estimation by sampling. For example, suppose we want to estimate the average age, income, family size, or other measure of a population. To do this, we determine a random process for selecting people—say, throwing darts at census lists. This process makes the selected person's age, income, and so on into a random variable whose *mean* equals the *actual average* age or income of the population. So, we can select a random sample of people and calculate the average of people in the sample to estimate the true average in the whole population. But when we make an estimate by repeated sampling, we need to know how much confidence we should have that our estimate is OK, and how large a sample is needed to reach a given confidence level. The issue is fundamental to all experimental science. Because of random errors—*noise*—repeated measurements of the same quantity rarely come out exactly the same. Determining how much confidence to put in experimental measurements is a fundamental and universal scientific issue. Technically, judging sampling or measurement accuracy reduces to finding the probability that an estimate *deviates* by a given amount from its expected value.

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

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

19.1 Markov's Theorem

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

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

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

Of course, these are not very strong conclusions. No IQ of over 300 has ever been recorded; and while many IQ's of over 150 have been recorded, the fraction of the population that actually has an IQ that high is very much smaller than 2/3. But though these conclusions are weak, we reached them using just the fact that the average IQ is 100—along with another fact we took for granted, that IQ is never negative. Using only these facts, we can't derive smaller fractions, because there are nonnegative random variables with mean 100 that achieve these fractions. For example, if we choose a random variable equal to 300 with probability 1/3 and 0 with probability 2/3, then its mean is 100, and the probability of a value of 300 or more really is 1/3.

Theorem 19.1.1 (Markov's Theorem)**.** *If R is a nonnegative random variable, then for all $x > 0$*

$$\Pr[R \geq x] \leq \frac{\mathrm{Ex}[R]}{x}. \tag{19.1}$$

Proof. Let y vary over the range of R. Then for any $x > 0$

$$\begin{aligned} \mathrm{Ex}[R] &::= \sum_y y \Pr[R = y] \\ &\geq \sum_{y \geq x} y \Pr[R = y] \geq \sum_{y \geq x} x \Pr[R = y] = x \sum_{y \geq x} \Pr[R = y] \\ &= x \Pr[R \geq x], \end{aligned} \tag{19.2}$$

where the first inequality follows from the fact that $R \geq 0$.

Dividing the first and last expressions in (19.2) 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 19.1.2. *If R is a nonnegative random variable, then for all $c \geq 1$*

$$\Pr[R \geq c \cdot \mathrm{Ex}[R]\,] \leq \frac{1}{c}. \tag{19.3}$$

This Corollary follows immediately from Markov's Theorem(19.1.1) by letting x be $c \cdot \text{Ex}[R]$.

19.1.1 Applying Markov's Theorem

Let's go back to the Hat-Check problem of Section 18.5.2. Now we ask what the probability is that x or more men get the right hat, this is, what the value of $\Pr[G \geq x]$ is.

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

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

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

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

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

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

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

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

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

19.1.2 Markov's Theorem for Bounded Variables

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

Markov's Theorem to the random variable, R, equal to the IQ of a random MIT student to conclude:

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

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

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

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

In fact, we can get better bounds applying Markov's Theorem to $R - b$ instead of R for any lower bound b on R (see Problem 19.3). Similarly, if we have any upper bound, u, on a random variable, S, then $u - S$ will be a nonnegative random variable, and applying Markov's Theorem to $u - S$ will allow us to bound the probability that S is much *less* than its expectation.

19.2 Chebyshev's Theorem

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

Lemma 19.2.1. *For any random variable R and positive real numbers x, z,*

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

Rephrasing (19.2.1) in terms of $|R - \mathrm{Ex}[R]|$, the random variable that measures R's deviation from its mean, we get

$$\Pr[|R - \mathrm{Ex}[R]| \geq x] \leq \frac{\mathrm{Ex}[(R - \mathrm{Ex}[R])^z]}{x^z}. \tag{19.4}$$

The case when $z = 2$ turns out to be so important that the numerator of the right hand side of (19.4) has been given a name:

Definition 19.2.2. The *variance*, $\mathrm{Var}[R]$, of a random variable, R, is:

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

Variance is also known as *mean square deviation*.

The restatement of (19.4) for $z = 2$ is known as *Chebyshev's Theorem*[1]

Theorem 19.2.3 (Chebyshev)**.** *Let R be a random variable and $x \in \mathbb{R}^+$. Then*

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

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

19.2.1 Variance in Two Gambling Games

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

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

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

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

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

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

[1] There are Chebyshev Theorems in several other disciplines, but Theorem 19.2.3 is the only one we'll refer to.

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

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

Similarly, we have for Var[B]:

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

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

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

19.2.2 Standard Deviation

In Game B above, the deviation from the mean is 1001 in one outcome and -2002 in the other. But the variance is a whopping 2,004,002. The happens because the "units" of variance are wrong: if the random variable is in dollars, then the expectation is also in dollars, but the variance is in square dollars. For this reason, people often describe random variables using *standard deviation* instead of variance.

Definition 19.2.4. The *standard deviation*, σ_R, of a random variable, R, is the square root of the variance:

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

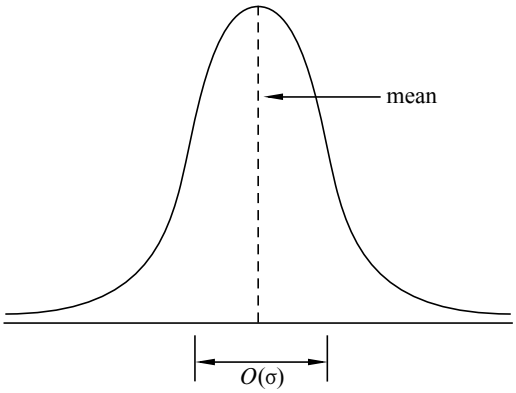

Figure 19.1 The standard deviation of a distribution indicates how wide the "main part" of it is.

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

Example 19.2.5. The standard deviation of the payoff in Game B is:

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

The random variable B actually deviates from the mean by either positive 1001 or negative 2002, so the standard deviation of 1416 describes this situation more closely than the value in the millions of the variance.

For bell-shaped distributions like the one illustrated in Figure 19.1, the standard deviation measures the "width" of the interval in which values are most likely to fall. This can be more clearly explained by rephrasing Chebyshev's Theorem in terms of standard deviation, which we can do by substituting $x = c\sigma_R$ in (19.1):

Corollary 19.2.6. *Let R be a random variable, and let c be a positive real number.*

$$\Pr[|R - \text{Ex}[R]| \geq c\sigma_R] \leq \frac{1}{c^2}. \tag{19.5}$$

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

The IQ Example

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 $\text{Ex}[R] = 100$, $\sigma_R = 10$, and R is nonnegative. We want to compute $\Pr[R \geq 300]$.

We have already seen that Markov's Theorem 19.1.1 gives a coarse bound, namely,

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

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

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

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

19.3 Properties of Variance

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

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

19.3.1 A Formula for Variance

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

Lemma 19.3.1.

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

for any random variable, R.

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

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

$$\begin{aligned}
\text{Var}[R] &= \text{Ex}[(R - \text{Ex}[R])^2] && \text{(Def 19.2.2 of variance)}\\
&= \text{Ex}[(R - \mu)^2] && \text{(def of } \mu\text{)}\\
&= \text{Ex}[R^2 - 2\mu R + \mu^2]\\
&= \text{Ex}[R^2] - 2\mu\,\text{Ex}[R] + \mu^2 && \text{(linearity of expectation)}\\
&= \text{Ex}[R^2] - 2\mu^2 + \mu^2 && \text{(def of } \mu\text{)}\\
&= \text{Ex}[R^2] - \mu^2\\
&= \text{Ex}[R^2] - \text{Ex}^2[R]. && \text{(def of } \mu\text{)}
\end{aligned}$$

■

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

Corollary 19.3.2. *If B is a Bernoulli variable where $p ::= \Pr[B = 1]$, then*

$$\text{Var}[B] = p - p^2 = p(1 - p). \tag{19.6}$$

Proof. By Lemma 18.4.2, $\text{Ex}[B] = p$. But B only takes values 0 and 1, so $B^2 = B$ and equation (19.6) follows immediately from Lemma 19.3.1. ■

19.3.2 Variance of Time to Failure

According to Section 18.4.6, the mean time to failure is $1/p$ for a process that fails during any given hour with probability p. What about the variance?

By Lemma 19.3.1,

$$\text{Var}[C] = \text{Ex}[C^2] - (1/p)^2 \tag{19.7}$$

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

Reasoning about C using conditional expectation worked nicely in Section 18.4.6 to find mean time to failure, and a similar approach works for C^2. Namely, the expected value of C^2 is the probability, p, of failure in the first hour times 1^2, plus the probability, $(1 - p)$, of non-failure in the first hour times the expected value of

$(C+1)^2$. So

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

Combining this with (19.7) proves

Lemma 19.3.3. *If failures occur with probability p independently at each step, and C is the number of steps until the first failure*[2]*, then*

$$\operatorname{Var}[C] = \frac{1-p}{p^2}. \tag{19.8}$$

19.3.3 Dealing with Constants

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

Theorem 19.3.4. *[*Square Multiple Rule *for Variance] Let R be a random variable and a a constant. Then*

$$\operatorname{Var}[aR] = a^2 \operatorname{Var}[R]. \tag{19.9}$$

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

$$\begin{aligned}
\operatorname{Var}[aR] ::= \operatorname{Ex}[(aR - \operatorname{Ex}[aR])^2] \\
&= \operatorname{Ex}[(aR)^2 - 2aR\operatorname{Ex}[aR] + \operatorname{Ex}^2[aR]] \\
&= \operatorname{Ex}[(aR)^2] - \operatorname{Ex}[2aR\operatorname{Ex}[aR]] + \operatorname{Ex}^2[aR] \\
&= a^2\operatorname{Ex}[R^2] - 2\operatorname{Ex}[aR]\operatorname{Ex}[aR] + \operatorname{Ex}^2[aR] \\
&= a^2\operatorname{Ex}[R^2] - a^2\operatorname{Ex}^2[R] \\
&= a^2\left(\operatorname{Ex}[R^2] - \operatorname{Ex}^2[R]\right) \\
&= a^2\operatorname{Var}[R] \qquad \text{(Lemma 19.3.1)}
\end{aligned}$$

[2] That is, C has the geometric distribution with parameter p according to Definition 18.4.6.

■

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

Theorem 19.3.5. *Let R be a random variable, and b a constant. Then*

$$\text{Var}[R + b] = \text{Var}[R]. \tag{19.10}$$

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 19.3.6.

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

19.3.4 Variance of a Sum

In general, the variance of a sum is not equal to the sum of the variances, but variances do add for *independent* variables. In fact, *mutual* independence is not necessary: *pairwise* independence will do. This is useful to know because there are some important situations, such as Birthday Matching in Section 16.4, that involve variables that are pairwise independent but not mutually independent.

Theorem 19.3.7. *If R and S are independent random variables, then*

$$\text{Var}[R + S] = \text{Var}[R] + \text{Var}[S]. \tag{19.11}$$

Proof. We may assume that $\text{Ex}[R] = 0$, since we could always replace R by $R - \text{Ex}[R]$ in equation (19.11); likewise for S. This substitution preserves the independence of the variables, and by Theorem 19.3.5, does not change the variances.

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

$$\text{Ex}[(R + S)^2] = \text{Ex}[R^2] + \text{Ex}[S^2]. \tag{19.12}$$

But (19.12) follows from linearity of expectation and the fact that

$$\text{Ex}[RS] = \text{Ex}[R]\,\text{Ex}[S] \tag{19.13}$$

since R and S are independent:

$$\begin{aligned}
\operatorname{Ex}[(R+S)^2] &= \operatorname{Ex}[R^2 + 2RS + S^2] \\
&= \operatorname{Ex}[R^2] + 2\operatorname{Ex}[RS] + \operatorname{Ex}[S^2] \\
&= \operatorname{Ex}[R^2] + 2\operatorname{Ex}[R]\operatorname{Ex}[S] + \operatorname{Ex}[S^2] \qquad \text{(by (19.13))} \\
&= \operatorname{Ex}[R^2] + 2 \cdot 0 \cdot 0 + \operatorname{Ex}[S^2] \\
&= \operatorname{Ex}[R^2] + \operatorname{Ex}[S^2]
\end{aligned}$$

■

It's easy to see that additivity of variance does not generally hold for variables that are not independent. For example, if $R = S$, then equation (19.11) becomes $\operatorname{Var}[R+R] = \operatorname{Var}[R] + \operatorname{Var}[R]$. By the Square Multiple Rule, Theorem 19.3.4, this holds iff $4\operatorname{Var}[R] = 2\operatorname{Var}[R]$, which implies that $\operatorname{Var}[R] = 0$. So equation (19.11) fails when $R = S$ and R has nonzero variance.

The proof of Theorem 19.3.7 carries over to the sum of any finite number of variables. So we have:

Theorem 19.3.8. *[Pairwise Independent Additivity of Variance] If $R_1, R_2, \ldots, R_n$ are* pairwise *independent random variables, then*

$$\operatorname{Var}[R_1 + R_2 + \cdots + R_n] = \operatorname{Var}[R_1] + \operatorname{Var}[R_2] + \cdots + \operatorname{Var}[R_n]. \tag{19.14}$$

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 Corollary 19.3.2, so by linearity of variance, we have

Lemma 19.3.9 (Variance of the Binomial Distribution)**.** *If J has the (n, p)-binomial distribution, then*

$$\operatorname{Var}[J] = n\operatorname{Var}[I_k] = np(1-p). \tag{19.15}$$

6.042J / 18.062J Mathematics for Computer Science
Spring 2015