Lecture Notes

1 Deviation from the Mean

This week we will look at several theorems to the effect that a random variable is unlikely to deviate very far from its mean.

1.1 Three Examples

Suppose that we roll a fair die. This gives a random variable uniformly 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.

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 1 and Figure 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 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 $\frac{1}{2}$. In fact, we will see that this is a worst-case distribution with respect to deviation from the mean.

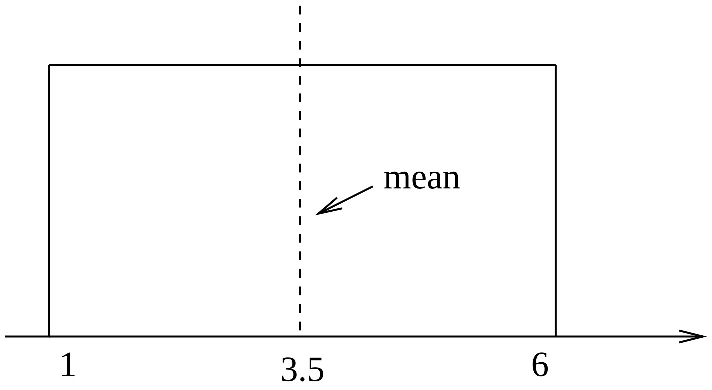

Figure 1: *This is a graph of the uniform distribution arising from rolling a fair die. Outcomes within the range of the distribution are equally likely, regardless of distance from the mean.*

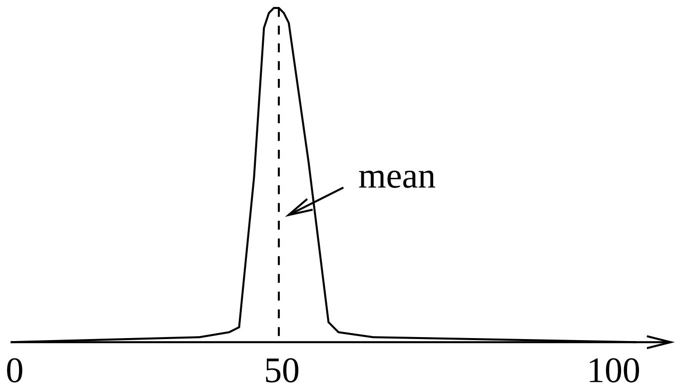

Figure 2: *This is a rough graph of the binomial distribution given by the number of heads that come up when we flip 100 fair, mutually independent coins. Outcomes close to the mean are much more likely than outcomes far from the mean.*

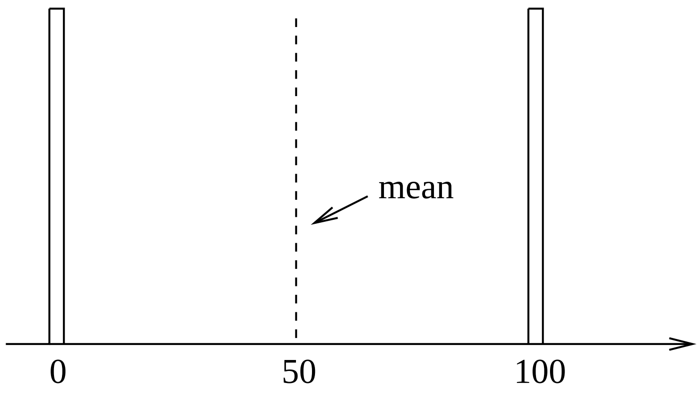

Figure 3: *This is the nasty distribution corresponding to the number of heads that come up when we flip 100 coins that are all glued together. The outcome always differs from the mean by at least 50.*

1.2 Markov's Theorem

Intuitively, Markov's Theorem says that a non-negative random variable can not greatly exceed its mean.

Theorem 1.1 (Markov's Theorem) If R is a non-negative random variable, then for all $x > 0$

$$\Pr(R \geq x) \quad \leq \quad \frac{\mathrm{Ex}(R)}{x}$$

Markov's Theorem is often in expressed in an alternate form, stated below as a corollary.

Corollary 1.2 If R is a non-negative random variable, then for all $c > 0$

$$\Pr(R \geq c \cdot \mathrm{Ex}(R)) \quad \leq \quad \frac{1}{c}$$

Proof. In Markov's Theorem, set $x = c \cdot \mathrm{Ex}(R)$. This gives:

$$Pr(R \geq c \cdot Ex(R)) \leq \frac{Ex(R)}{c \cdot \mathrm{Ex}(R)} = \frac{1}{c}$$

■

Before we prove Markov's Theorem, let's apply it to the three examples in the preceding section. 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:

$$\Pr(R \geq 6) \leq \frac{\mathrm{Ex}(R)}{6} = \frac{3.5}{6} = 0.583\ldots$$

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

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 non-negativity. But as a consequence, Markov's Theorem often leads to weak conclusions like the one above.

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:

$$\Pr(\text{heads} \geq 75) \leq \frac{\mathrm{Ex}(\text{heads})}{75} = \frac{50}{75} = \frac{2}{3}$$

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

These two examples show that Markov's Theorem gives weak results for well-behaved random variables; however, the theorem is actually tight for some nasty examples. Suppose we flip 100 fair coins and use Markov's Theorem to compute the probability of getting all heads:

$$\Pr(\text{heads} \geq 100) \leq \frac{\text{Ex(heads)}}{100} = \frac{50}{100} = \frac{1}{2}$$

If the coins in 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 $\frac{1}{2}$. In this nasty case, Markov's Theorem is actually tight!

1.3 Proof of Markov's Theorem

The next example gives some intuition for why Markov's Theorem is true. Let R be the weight of a person selected randomly and uniformly. Suppose that an average person weighs 100 pounds; that is, $\text{Ex}(R) = 100$. 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 $\frac{1}{2}$. 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:

$$\Pr(R \geq 200) \leq \frac{\text{Ex}(R)}{200} = \frac{100}{200} = \frac{1}{2}$$

Similarly, the probability that a random person weighs at least 300 pounds is at most $\frac{1}{3}$. If more than $\frac{1}{3}$ of the people weighed 300 pounds or more, then the average weight would exceed 100 pounds even if everyone else weighed 0. Again, Markov's Theorem gives the same answer:

$$\Pr(R \geq 300) \leq \frac{\text{Ex}(R)}{300} = \frac{100}{300} = \frac{1}{3}$$

Reasoning similar to that above underlies the proof of Markov's Theorem.

Proof. *(of Markov's Theorem)* We must prove that if R is a non-negative random variable, then for all $x > 0$, we have

$$\Pr(R \geq x) \leq \frac{\text{Ex}(R)}{x}$$

We will show that $\text{Ex}(R) \geq x \cdot \Pr(R \geq x)$. Dividing both sides by x gives the desired result.

$$
\begin{aligned}
\mathrm{Ex}(R) &= \sum^{z \in \mathrm{Range}(R)} z \cdot \Pr(R = z) \\
&\geq \sum^{z \geq x} z \cdot \Pr(R = z) \\
&\geq \sum^{z \geq x} x \cdot \Pr(R = z) \\
&= x \cdot \sum^{z \geq x} \Pr(R = z) \\
&= x \cdot \Pr(R \geq x)
\end{aligned}
$$

In the fist step, we use the definition of expected value to obtain a sum. Since R is a non-negative random variable, z is always non-negative, and so all terms in the sum are also non-negative. In the second step, we restrict the range of the summation to $z \geq x$. This restriction can not increase the value of the summation, since all terms are non-negative. In the third step, we replace z by x in the sum. The inequality follows because $z \geq x$ over the entire range of summation. In the fourth step, we pull x out of the summation, and the final step uses the definition of $\Pr(R \geq x)$. ■

1.4 More Examples of Markov's Theorem

Suppose that N men go to a dinner party and check their hats. At the end of the night, each man is given his own hat back with probability $\frac{1}{N}$. What is the probability that x or more men get the right hat?

We can compute an upper bound with Markov's Theorem. Let the random variable R be the number of men that get the right hat. In a previous lecture, we used linearity of expectation to show that $\mathrm{Ex}(R) = 1$. By Markov's Theorem, the probability that x or more men get the the right hat is:

$$\Pr(R \geq x) \leq \frac{\mathrm{Ex}(R)}{x} = \frac{1}{x}$$

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 very similar. In this case, N people are eating Chinese appetizers arranged on a circular, rotating 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 $\frac{1}{N}$.

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. We showed in a previous lecture that $\mathrm{Ex}(R) = 1$. Applying Markov's Theorem, we find:

$$\begin{aligned}\Pr(R \geq N) &\leq \frac{\text{Ex}(R)}{N} \\ &= \frac{1}{N}\end{aligned}$$

In this case, Markov's Theorem is tight!

1.5 Why R Must be Non-Negative

The proof of Markov's Theorem requires that the random variable R be non-negative. The following example shows that the theorem is false if this restriction is removed. Let R be -10 with probability $\frac{1}{2}$ and 10 with probability $\frac{1}{2}$. Then we have:

$$\text{Ex}(R) = -10 \cdot \frac{1}{2} + 10 \cdot \frac{1}{2} = 0$$

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

$$\begin{aligned}\Pr(R \geq 5) &\leq \frac{\text{Ex}(R)}{5} \\ &= \frac{0}{5} \\ &= 0\end{aligned}$$

This is the wrong answer! Obviously, R is at least 5 with probability $\frac{1}{2}$. Remember that Markov's Theorem applies only to non-negative random variables!

1.6 Deviation Below the Mean

Markov's Theorem says that a random variable is unlikely to greatly exceed the mean. There a similar theorem that says a random variable is unlikely to be very small.

Theorem 1.3 Let L be a real number and let R be a random variable such that $R \leq L$. For all $x < L$, we have:

$$\Pr(R \leq x) \leq \frac{L - \text{Ex}(R)}{L - x}$$

Proof. The event that $R \leq x$ is the same as the event that $L - R \geq L - x$. Therefore:

$$\begin{aligned} \Pr(R \leq x) &= \Pr(L - R \geq L - x) \\ &\leq \frac{\text{Ex}(L - R)}{L - x} \end{aligned}$$

In the second step, we applied Markov's Theorem. This is permissible since $L - R$ is a nonnegative random variable and $L - x > 0$. ■

For example, suppose that the class average on the 6.042 midterm was 75%. What fraction of the class scored below 50%?

There is not enough information here to answer the question exactly, but Theorem 1.3 gives an upper bound. Let R be the score of a random student. Since 100% is the highest possible score, we can set $L = 100$ to meet the condition in the theorem that $R \leq L$. Applying Theorem 1.3, we find:

$$\begin{aligned} \Pr(R \leq 50) &\leq \frac{100 - 75}{100 - 50} \\ &= \frac{1}{2} \end{aligned}$$

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 1.3 often gives weak results. In fact, based on the data given, the entire class could have scored above 50%.

1.7 Using Markov To Analyze Non-Random Events

In the previous examples, 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 $\frac{1}{2}$ the class scored 50% or worse. There is no randomness in this problem, so how can we apply Theorem 1.3 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 1.3 to conclude that $\Pr(R \leq 50\%) \leq \frac{1}{2}$. To see why this means (with certainty) that at most $\frac{1}{2}$ of the students scored 50% or less, we observe that

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

Therefore $\Pr(R \leq 50\%) \leq \frac{1}{2}$ implies that the number of students with score 50% or less is at most $\frac{n}{2}$.

2 Variance

Markov's Theorem uses too little information about a random variable to always give good bounds on deviation from the mean. With more information, we can obtain tighter bounds. One particularly useful piece of information about a random variable is the *variance*.

Definition The variance of a random variable R is denoted $\operatorname{Var}(R)$ and is defined as follows:

$$\operatorname{Var}(R) = \operatorname{Ex}((R - \operatorname{Ex}(R))^2)$$

The expression $\operatorname{Ex}((R - \operatorname{Ex}(R))^2)$ is a bit cryptic; the best approach is to work through it from the inside out. The innermost expression, $R - \operatorname{Ex}(R)$, is precisely the deviation of R from the mean. Squaring this, we obtain, $(R - \operatorname{Ex}(R))^2$. 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. The variance is just the average of this random variable, $\operatorname{Ex}((R - \operatorname{Ex}(R))^2)$. Therefore, intuitively, 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. For this reason, variance is useful in studying the probability that a random variable deviates far from the mean.

2.1 Gambling Games

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

Game A: We win \$2 with probability $\frac{2}{3}$ and lose \$1 with probability $\frac{1}{3}$.

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

Which game is more better financially? We have the same probability ($\frac{2}{3}$) 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 $\frac{2}{3}$ and -1 with probability $\frac{1}{3}$. 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 they are obviously very different! This difference is hidden by expected value, but captured by variance. We can compute the $\mathrm{Var}(A)$ and $\mathrm{Var}(B)$ by working "from the inside out" as follows:

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

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

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 \$20 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!

2.2 Why Variance?

The definition of variance looks rather arbitrary:

$$\mathrm{Var}(R) = \mathrm{Ex}((R - \mathrm{Ex}(R))^2)$$

This says that 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 $\mathrm{Ex}(R - \mathrm{Ex}(R))$?

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

$$\mathrm{Ex}(R - \mathrm{Ex}(R)) = \mathrm{Ex}(R) - \mathrm{Ex}(\mathrm{Ex}(R))$$

Since $\mathrm{Ex}(R)$ is a constant, its expected value is itself.

$$\begin{aligned} &= \mathrm{Ex}(R) - \mathrm{Ex}(R) \\ &= 0 \end{aligned}$$

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

Of course, we could also prevent positive and negative deviations from cancelling by taking an absolute value. That is, we could define variance to be $\mathrm{Ex}(|R - \mathrm{Ex}(R)|)$. There is no great reason not to use this definition. However, the conventional version of variance has some pleasant mathematical properties that the absolute value variant does not. For example, for independent random variables, the variance of a sum is the sum of the variances; that is, $\mathrm{Var}(R_1 + R_2) = \mathrm{Var}(R_1) + \mathrm{Var}(R_2)$. We will prove this fact in the last section.

Nevertheless, due to squaring, the variance of a random variable may be very far from a typical deviation from the mean. For example, in Game B the deviation from the mean is 1000 in one outcome and -2003 in the other. But the variance is a whopping 2,004,002. For this reason, people often describe random variables using standard deviation instead of variance:

Definition The *standard deviation* of a random variable R is denoted σ_R and defined as the square root of the variance:

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

The standard deviation is the square root of the mean of the square of the deviation, or the "root mean square" for short. As an example, the standard deviation of the payoff in Game B is:

$$\begin{aligned} \sigma_B &= \sqrt{\mathrm{Var}(B)} \\ &= \sqrt{2,004,002} \\ &\approx 1416 \end{aligned}$$

The random variable B actually deviates from the mean by either positive 1000 or negative 2003; therefore, the standard deviation of 1416 describes the situations reasonably well.

3 Chebyshev's Theorem

Intuitively, variance and standard deviation describe the tendency of a random variable to deviate from the mean. Chebyshev's Theorem formalizes this intuition.

Theorem 3.1 Let R be a random variable, and let x be a positive real number.

$$\Pr(|R - \mathrm{Ex}(R)| \geq x) \leq \frac{\mathrm{Var}(R)}{x^2}$$

Chebyshev's Theorem is similar to Markov's Theorem, but stronger because it uses the extra information conveyed by the variance. In fact, the proof of Chebyshev's Theorem uses Markov's Theorem.

Proof. Note that the event $|R - \mathrm{Ex}(R)| \geq x$ is the same as the event $(R - \mathrm{Ex}(R))^2 \geq x^2$, since the latter inequality is obtained by squaring both sides of the former and both sides of the former are positive. Therefore, these two events have equal probability:

$$\begin{aligned} \Pr(|R - \mathrm{Ex}(R)| \geq x) &= \Pr((R - \mathrm{Ex}(R))^2 \geq x^2) \\ &\leq \frac{\mathrm{Ex}((R - \mathrm{Ex}(R))^2)}{x^2} \\ &= \frac{\mathrm{Var}(R)}{x^2} \end{aligned}$$

We can apply Markov's Theorem on the second line, since the random variable $(R - \mathrm{Ex}(R))^2$ is non-negative. The third step follows from the definition of variance. ■

There is a nice reformulation of Chebyshev's Theorem in terms of standard deviation.

Corollary 3.2 Let R be a random variable, and let c be a positive real number.

$$\Pr(|R - \mathrm{Ex}(R)| \geq c\sigma_R) \leq \frac{1}{c^2}$$

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

$$\begin{aligned} \Pr(|R - \mathrm{Ex}(R)| \geq c\sigma_R) &\leq \frac{\mathrm{Var}(R)}{(c\sigma_R)^2} \\ &= \frac{1}{c^2} \end{aligned}$$

The last equality holds because variance is the square of standard deviation; in particular, $\mathrm{Var}(R) = \sigma_R^2$. ■

3.1 Example of Chebyshev's Theorem

Suppose that the average I.Q. is 100 and the standard deviation is 10. How rare is an I.Q. of 200 or more?

Let the random variable R be the I.Q. of a random person. We are supposing that $\text{Ex}(R) = 100$ and $\sigma_R = 10$. Furthermore, let's assume that everyone has an I.Q. that is positive. We want to compute $\Pr(R \geq 200)$.

Since R is non-negative, we can get an upper bound using Markov's Theorem as shown below.

$$\begin{aligned}
\Pr(R \geq 200) &\leq \frac{\text{Ex}(R)}{200} \\
&= \frac{100}{200} \\
&= \frac{1}{2}
\end{aligned}$$

Markov's Theorem says that at most half of the people have I.Q. 200 or more; this sounds like a very loose upper bound! Suppose that we now apply Chebyshev's Theorem to the same problem.

$$\begin{aligned}
\Pr(R \geq 200) &= \Pr(|R - 100| \geq 100) \\
&\leq \frac{\text{Var}(R)}{100^2} \\
&= \frac{10^2}{100^2} \\
&= \frac{1}{100}
\end{aligned}$$

The purpose of the first step is to express the desired probability in the form required by Chebyshev's Theorem. (There is equality because R is positive.) We actually apply Chebyshev's Theorem in the second step. The remaining work is substitution and simplification.

Chebyshev's Theorem says that at most one person in a hundred has an I.Q. of 200 or more. Chebyshev's Theorem gives a much tighter bound because it uses additional information, the variance of R.

3.2 A One-Sided Bound

Chebyshev's Theorem gives a "two-sided bound". That is, it bounds the probability that a random variable deviates *above or below* the mean by some amount. What if we want only a one-sided bound? For example, what is the probability that a random variable deviates *above* the mean by some amount?

This question is often answered incorrectly. The erroneous argument runs as follows. By Chebyshev's Theorem, R deviates above or below the mean by some amount with probability p. Therefore,

R deviates above the mean by this amount with probability $\frac{p}{2}$, and R deviates below the mean by this amount with probability $\frac{p}{2}$.

The error in this argument is that a random variable may be more likely to deviate above the mean than below. For example, in the I.Q. question, some people deviate 100 points above the mean; that is, there are people with I.Q. greater than 200. However, by assumption everyone has a positive I.Q.; no one deviates 100 points below the mean. (The argument above is correct, however, for a probability distribution function that is symmetric about the mean.)

The following theorem is analogous to Chebyshev's Theorem, but gives (correct!) one-sided bounds. We will not give a proof.

Theorem 3.3 Let R be a random variable, and let x be a positive real number.

$$\begin{aligned} \Pr(R - \text{Ex}(R) \geq x) &\leq \frac{\text{Var}(R)}{\text{Var}(R) + x^2} \\ \Pr(R - \text{Ex}(R) \leq -x) &\leq \frac{\text{Var}(R)}{\text{Var}(R) + x^2} \end{aligned}$$

This theorem has a corollary analogous to Corollary 3.2.

Corollary 3.4 Let R be a random variable, and let c be a positive real number.

$$\begin{aligned} \Pr(R - \text{Ex}(R) \geq c\sigma_R) &\leq \frac{1}{1 + c^2} \\ \Pr(R - \text{Ex}(R) \leq -c\sigma_R) &\leq \frac{1}{1 + c^2} \end{aligned}$$

Proof. Substitute $x = c\sigma_R$ in the preceding theorem. ■

We can use this theorem to improve our solution to the I.Q. problem just slightly. Recall that R is the I.Q. of a random person and that $\text{Ex}(R) = 100$ and $\sigma_R = 10$.

$$\begin{aligned} \Pr(R \geq 200) &= \Pr(R - 100 \geq 100) \\ &\leq \frac{\text{Var}(R)}{\text{Var}(R) + 100^2} \\ &= \frac{10^2}{10^2 + 100^2} \\ &= \frac{1}{101} \end{aligned}$$

This shows that at most 1 person in 101 has an I.Q. of 200 or more.

3.3 Exam Problems

The following example shows the difference between one-sided and two-sided bounds. On the 6.042 final, what fraction of the class can score two standard deviations or more *above or below* the class average? Surprisingly, we can answer this question without any assumptions about the range or distribution of scores!

In this case, we want a two-sided bound. Let R be the score of a random student. By Corollary 3.2, the probability is at most:

$$\Pr(|R - \mathrm{Ex}(R)| \geq 2\sigma) \leq \frac{1}{2^2} = \frac{1}{4}$$

What fraction of the class can score two standard deviations or more *above* the class average? Now we want a one-sided bound. By Corollary 3.4, the probability is at most:

$$\Pr(R - \mathrm{Ex}(R) \geq 2\sigma) \leq \frac{1}{1+2^2} = \frac{1}{5}$$

This makes sense; the probability of deviating above the mean ($\frac{1}{5}$) should be smaller than the probability of deviating above or below the mean ($\frac{1}{4}$).

3.4 Deviation of an Unbiased Binomial Distribution

Chebyshev's Theorem applies to any random variable. However, for a specific distribution, we can sometimes prove a stronger bound on deviation from the mean. For example, a very strong bound is possible for an unbiased binomial distribution.

Fact 1 For a random variable R with an unbiased binomial distribution

$$\Pr(R - \mathrm{Ex}(R) \geq c\sigma_R) < e^{-c^2/2}$$

This fact will be proven in tutorial. Note that the bound in Chebyshev's Theorem is polynomial ($\frac{1}{c^2}$), but the bound above is exponential ($e^{-c^2/2}$).

For example, suppose that the 6.042 final consists entirely of true/false questions[1] and students guess answers uniformly and mutually independently. A student's score on each question is a uniform Bernoulli variable (0 or 1). A student's exam score is a sum of these Bernoulli variables, and so is a random variable with an unbiased binomial distribution.

Given these assumptions, what is the probability that a student scores two standard deviations above the mean? By Chebyshev's Theorem the probability is at most 1 in 4. However, we can get a better bound using the fact above.

[1] Ha! Fat chance.

$$\begin{aligned}\Pr(R - \mathrm{Ex}(R) \geq 2\sigma_R) &< e^{-2^2/2} \\ &= e^{-2}\end{aligned}$$

The probability that a student scores 2 standard deviations above the mean on this exam is less than 1 in 7.

The difference is more striking if we look at greater deviations. For example, the probability that a student scores 6 standard deviations above the mean is at most $e^{-6^2/2} = e^{-18}$, which is less than one in 65 million. The corresponding bound from Corollary 3.4 is $\frac{1}{37}$.

4 Facts about Variance

The remainder of the lecture is devoted to proving nice and useful facts about variance.

4.1 An Alternate Definition of Variance

There is an equivalent way to define the variance of a random variable that is less intuitive, but often easier to use in both calculations and proofs. This alternate definition is stated below in the form of a theorem.

Theorem 4.1 If R is a random variable, then $\mathrm{Var}(R) = \mathrm{Ex}(R^2) - \mathrm{Ex}(R)^2$.

Remember that $\mathrm{Ex}(R^2)$ is generally not equal to $\mathrm{Ex}(R)^2$. The expected value of a product is the product of the expected values only for independent variables, and R is generally not independent of itself.

Proof.

$$\begin{aligned}\mathrm{Var}(R) &= \mathrm{Ex}((R - \mathrm{Ex}(R))^2) \\ &= \mathrm{Ex}(R^2 - 2R \cdot \mathrm{Ex}(R) + \mathrm{Ex}(R)^2) \\ &= \mathrm{Ex}(R^2) - \mathrm{Ex}(2R \cdot \mathrm{Ex}(R)) + \mathrm{Ex}(\mathrm{Ex}(R)^2) \\ &= \mathrm{Ex}(R^2) - 2\mathrm{Ex}(R) \cdot \mathrm{Ex}(R) + \mathrm{Ex}(R)^2 \\ &= \mathrm{Ex}(R^2) - \mathrm{Ex}(R)^2\end{aligned}$$

The first step uses the definition of variance. In the second step, we multiply out the squared term. The third step uses linearity of expectation. There are two transformations on the fourth line. In the second term, we pull the constant $2\mathrm{Ex}(R)$ out of the expectation. In the third term, we use the fact that the expectation of a constant, namely $\mathrm{Ex}(R)^2$, is that constant. The final step is simplification. ■

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

$$\begin{aligned}
\text{Ex}(R^2) &= \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = 15\frac{1}{6} \\
\text{Ex}(R)^2 &= \left(3\frac{1}{2}\right)^2 = 12\frac{1}{4} \\
\text{Var}(R) &= \text{Ex}(R^2) - \text{Ex}(R)^2 \\
&= 15\frac{1}{6} - 12\frac{1}{4} \\
&= 2\frac{11}{12}
\end{aligned}$$

4.2 Computing Variance on the Fly

Suppose that there is a random variable R defined on a uniform sample space. Now someone begins reading off the value of the random variable at each sample point. Each value is read exactly once. How can we compute the variance?

This problem might come up in practice if we wanted to compute the variance of a huge set of numbers archived on tapes. In effect, we "hear" each number just once, since we want to avoid multiple passes through the data!

With our original definition of variance, this problem is difficult. If the values of the variable are $x_1, x_2, \ldots x_n$, then we must compute:

$$\begin{aligned}
\text{Var}(R) &= \text{Ex}((R - \text{Ex}(R))^2) \\
&= \frac{(x_1 - \text{Ex}(R))^2 + (x_2 - \text{Ex}(R))^2 + \ldots + (x_n - \text{Ex}(R))^2}{n}
\end{aligned}$$

One approach would be to keep a running sum of the values in the numerator. When we are given value x_i, we compute $(x_i - \text{Ex}(R))^2$ and add it to the running sum. At then end, we divide the running sum by n to obtain the variance. The problem with this approach is that we do not know the value of $\text{Ex}(R)$ until after we have seen all of the data! Therefore, we need *two* passes, one to compute $\text{Ex}(R)$ and a second to compute the variance.

A better approach is to use the second definition of variance. This requires that we compute:

$$\begin{aligned}
\text{Var}(R) &= \text{Ex}(R^2) - \text{Ex}(R)^2 \\
&= \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - \left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right)^2
\end{aligned}$$

Now the problem is much easier. We 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, $x_1 + x_2 + \ldots + x_i$ and $x_1^2 + x_2^2 + \ldots + x_i^2$. Then, when we get the $(i+1)$-st number, we add x_{i+1} to the first sum and x_{i+1}^2 to the second sum. After all numbers are heard, we can plug these two summations into the definition above to find the variance. Now finding the variance requires only a single pass through the data!

4.3 A Cute Corollary

The alternate definition of variance given in Theorem 4.1 has a cute implication:

Corollary 4.2 If R is a random variable, then $\text{Ex}(R^2) \geq \text{Ex}(R)^2$.

Proof. We first defined $\text{Var}(R)$ as an average of a squared expression, so $\text{Var}(R)$ is non-negative. Then we proved that $\text{Var}(R) = \text{Ex}(R^2) - \text{Ex}(R)^2$. This implies that $\text{Ex}(R^2) - \text{Ex}(R)^2$ is non-negative. Therefore, $\text{Ex}(R^2) \geq \text{Ex}(R)^2$. ■

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:

$$\begin{aligned}
\text{Ex}(R^2) &= \text{Ex}(R)^2 \\
\Leftrightarrow \text{Ex}(R^2) - \text{Ex}(R)^2 &= 0 \\
\Leftrightarrow \text{Var}(R) &= 0
\end{aligned}$$

4.4 Zero Variance

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

Theorem 4.3 The variance of a random variable R is zero if and only if $R = \text{Ex}(R)$ for all outcomes with positive probability.

The final phrase is a technicality; for an outcome with zero probability, R can take on any value without affecting the variance.

Proof. By the definition of variance, $\text{Var}(R) = 0$ is equivalent to the condition:

$$\text{Ex}((R - \text{Ex}(R))^2) = 0$$

The inner expression, $(R - \text{Ex}(R))^2$, is always non-negative because of the square. As a result, $\text{Ex}((R - \text{Ex}(R))^2) = 0$ if an only if $(R - \text{Ex}(R))^2 = 0$ for all outcomes with positive probability. Now, the conditions $(R - \text{Ex}(R))^2 = 0$ and $R = \text{Ex}(R)$ are also equivalent. Therefore, $\text{Var}(R) = 0$ iff $R = \text{Ex}(R)$ for all outcomes with positive probability. ■

4.5 Dealing with Constants

The following theorem describes how the variance of a random variable changes when it is scaled or shifted by a constant.

Theorem 4.4 Let R be a random variable, and let a and b be constants.

$$\operatorname{Var}(aR+b) = a^2 \operatorname{Var}(R)$$

This theorem makes two points. First, adding a constant b to a random variable does not affect the variance. Second, multiplying a random variable by a constant a changes the variance by an a^2 factor.

Proof. We will transform the left side into the right side. The first step is to expand $\operatorname{Var}(aR+b)$ using the alternate definition of variance.

$$\operatorname{Var}(aR+b) \quad = \quad \operatorname{Ex}((aR+b)^2) - \operatorname{Ex}(aR+b)^2$$

We will work on the first term and then the second term.

$$\begin{aligned}
&= \operatorname{Ex}(a^2R^2 + 2abR + b^2) - \operatorname{Ex}(aR+b)^2 \\
&= (\operatorname{Ex}(a^2R^2) + \operatorname{Ex}(2abR) + \operatorname{Ex}(b^2)) - \operatorname{Ex}(aR+b)^2 \\
&= (a^2\operatorname{Ex}(R^2) + 2ab\operatorname{Ex}(R) + b^2) - \operatorname{Ex}(aR+b)^2
\end{aligned}$$

In the first step, we multiply out the squared expression in the first term. The second step uses linearity of expectation, and in the third step, we pull all constants out of expectations. Now we work on the second term:

$$\begin{aligned}
&= (a^2\operatorname{Ex}(R^2) + 2ab\operatorname{Ex}(R) + b^2) - (a\operatorname{Ex}(R) + b)^2 \\
&= (a^2\operatorname{Ex}(R^2) + 2ab\operatorname{Ex}(R) + b^2) - (a^2\operatorname{Ex}(R)^2 + 2ab\operatorname{Ex}(R) + b^2)
\end{aligned}$$

In the first step, we pull the constants out of the expectation. In the second step, we multiply out the square. Cancelling terms, we find:

$$\begin{aligned}
\operatorname{Var}(aR+b) &= a^2\operatorname{Ex}(R^2) - a^2\operatorname{Ex}(R)^2 \\
&= a^2(\operatorname{Ex}(R^2) - \operatorname{Ex}(R)^2) \\
&= a^2 \operatorname{Var}(R)
\end{aligned}$$

■

A similar rule holds for the standard deviation when a random variable is adjusted by a constant. Recall that standard deviation is the square root of variance. Therefore, adding a constant b to a random variable does not change the standard deviation. Multiplying a random variable by a constant a multiplies the standard deviation by a.

4.6 Variance of a Sum

Earlier, we claimed that for independent random variables, the variance of a sum is the sum of the variances. This is stated formally and proved below.

Theorem 4.5 If R_1 and R_2 are independent random variables, then

$$\mathrm{Var}(R_1 + R_2) = \mathrm{Var}(R_1) + \mathrm{Var}(R_2)$$

Proof. We will transform the left side into the right side. We begin by applying the alternate definition of variance.

$$\mathrm{Var}(R_1 + R_2) = \mathrm{Ex}((R_1 + R_2)^2) - \mathrm{Ex}(R_1 + R_2)^2$$

We will work on the first term and then the second term.

$$\begin{aligned} &= \mathrm{Ex}(R_1^2 + 2R_1R_2 + R_2^2) - \mathrm{Ex}(R_1 + R_2)^2 \\ &= (\mathrm{Ex}(R_1^2) + \mathrm{Ex}(2R_1R_2) + \mathrm{Ex}(R_2^2)) - \mathrm{Ex}(R_1 + R_2)^2 \\ &= (\mathrm{Ex}(R_1^2) + 2\mathrm{Ex}(R_1)\mathrm{Ex}(R_2) + \mathrm{Ex}(R_2^2)) - \mathrm{Ex}(R_1 + R_2)^2 \end{aligned}$$

First, we multiply out the squared expression. The second step uses linearity of expectation. In the last step, we break the expectation of the product R_1R_2 into a product of expectations; this is where we use the fact that R_1 and R_2 are independent. Now we work on the second term.

$$\begin{aligned} &= (\mathrm{Ex}(R_1^2) + 2\mathrm{Ex}(R_1)\mathrm{Ex}(R_2) + \mathrm{Ex}(R_2^2)) - (\mathrm{Ex}(R_1) + \mathrm{Ex}(R_2))^2 \\ &= (\mathrm{Ex}(R_1^2) + 2\mathrm{Ex}(R_1)\mathrm{Ex}(R_2) + \mathrm{Ex}(R_2^2)) - (\mathrm{Ex}(R_1)^2 + 2\mathrm{Ex}(R_1)\mathrm{Ex}(R_2) + \mathrm{Ex}(R_2)^2) \end{aligned}$$

The first step uses linearity of expectation, and in the second step we multiply out the squared expression. Cancelling and rearranging terms, we find:

$$\begin{aligned} \mathrm{Var}(R_1 + R_2) &= (\mathrm{Ex}(R_1^2) - \mathrm{Ex}(R_1)^2) + (\mathrm{Ex}(R_2^2)) - \mathrm{Ex}(R_2)^2) \\ &= \mathrm{Var}(R_1) + \mathrm{Var}(R_2) \end{aligned}$$

■

The independence condition in the theorem is necessary. If we ignored this condition and took R_1 and R_2 to be the same random variable, then we would conclude:

$$\mathrm{Var}(R+R) = \mathrm{Var}(R) + \mathrm{Var}(R)$$

However, by Theorem 4.4, the left side is equal to $4\,\mathrm{Var}(R)$, whereas the right side is $2\,\mathrm{Var}(R)$. Therefore, if we remove the condition that R_1 and R_2 be independent, then the theorem is false.

The preceding theorem extends to a collection of *pairwise independent* random variables. Recall that pairwise independence is actually a much weaker condition than mutual independence.

Theorem 4.6 If $R_1, R_2, \ldots, R_n$ are pairwise independent random variables, then

$$\mathrm{Var}(R_1 + R_2 + \ldots + R_n) = \mathrm{Var}(R_1) + \mathrm{Var}(R_2) + \ldots + \mathrm{Var}(R_n)$$

The proof is by induction and is routine, so we will skip it.

4.7 Variance of a Binomial Distribution

We now have enough tools to find the variance of a binomial distribution. Recall that if a random variable R has a binomial distribution, then

$$\Pr(R = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where n and p are parameters such that $n \geq 1$ and $0 < p < 1$.

We can think of R as the sum of n independent Bernoulli variables. For example, we can regard R as the number of heads that come up when we toss n independent coins, where each coin comes up heads with probability p. Formally, we can write

$$\begin{aligned} R &= R_1 + R_2 + \ldots + R_n \\ R_i &= \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases} \end{aligned}$$

Now we can compute the variance of the binomially distributed variable R.

$$\begin{aligned} \mathrm{Var}(R) &= \mathrm{Var}(R_1) + \mathrm{Var}(R_2) + \ldots + \mathrm{Var}(R_n) \\ &= n\,\mathrm{Var}(R_i) \\ &= n(\mathrm{Ex}(R_i^2) - \mathrm{Ex}(R_i)^2) \\ &= n(\mathrm{Ex}(R_i) - \mathrm{Ex}(R_i)^2) \\ &= n(p - p^2) \end{aligned}$$

The first equation uses Theorem 4.6, and the second uses the fact that all variables R_i are identically distributed and therefore have the same variance. We then expand $\text{Var}(R_i)$ using the alternate definition of variance. In the fourth step, we observe that R_i^2 and R_i are the same random variable. This is always true for a Bernoulli random variable, because $0^2 = 0$ and $1^2 = 1$. The last step uses the fact that the expectation of a Bernoulli variable is the probability that the variable is 1; in this case, R_i is 1 with probability p.

This shows that the binomial distribution has variance $p(1-p)n$ and standard deviation $\sqrt{p(1-p)n}$. In the special case of an unbiased binomial distribution ($p = \frac{1}{2}$), the variance is $\frac{n}{4}$ and the standard deviation is $\frac{\sqrt{n}}{2}$.