Expectation & Variance

1 Expectation

1.1 Average & Expected Value

The ***expectation*** of a random variable is its average value, where each value is weighted according to the probability that it comes up. The expectation is also called the ***expected value*** or the ***mean*** of the random variable.

For example, suppose we select a student uniformly at random from the class, and let R be the student's quiz score. Then $\mathrm{E}[R]$ is just the class average —the first thing everyone wants to know after getting their test back! Similarly, the first thing you usually want to know about a random variable is its expected value.

Definition 1.1.

$$\begin{aligned}\mathrm{E}[R] ::= & \sum_{x \in \text{range}(R)} x \cdot \Pr\{R = x\} \\ = & \sum_{x \in \text{range}(R)} x \cdot \mathrm{PDF}_R(x).\end{aligned} \tag{1}$$

Let's work through an example. Let R be the number that comes up on a fair, six-sided die. Then by (1), the expected value of R is:

$$\begin{aligned}\mathrm{E}[R] &= \sum_{k=1}^{6} k \cdot \frac{1}{6} \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= \frac{7}{2}\end{aligned}$$

This calculation shows that the name "expected value" is a little misleading; the random variable might *never* actually take on that value. You can't roll a $3\frac{1}{2}$ on an ordinary die!

There is an even simpler formula for expectation:

Theorem 1.2. *If R is a random variable defined on a sample space, $\mathcal{S}$, then*

$$\mathrm{E}[R] = \sum_{\omega \in \mathcal{S}} R(\omega) \Pr\{\omega\} \tag{2}$$

The proof of Theorem 1.2, like many of the elementary proofs about expectation in these notes, follows by judicious regrouping of terms in the defining sum (1):

Proof.

$$\begin{aligned}
\mathrm{E}[R] ::= & \sum_{x \in \text{range}(R)} x \cdot \Pr\{R = x\} && \text{(Def 1.1 of expectation)} \\
= & \sum_{x \in \text{range}(R)} x \left(\sum_{\omega \in [R=x]} \Pr\{\omega\} \right) && \text{(def of } \Pr\{R = x\}) \\
= & \sum_{x \in \text{range}(R)} \sum_{\omega \in [R=x]} x \Pr\{\omega\} && \text{(distributing } x \text{ over the inner sum)} \\
= & \sum_{x \in \text{range}(R)} \sum_{\omega \in [R=x]} R(\omega) \Pr\{\omega\} && \text{(def of the event } [R = x]) \\
= & \sum_{\omega \in \mathcal{S}} R(\omega) \Pr\{\omega\}
\end{aligned}$$

The last equality follows because the events $[R = x]$ for $x \in \text{range}(R)$ partition the sample space, $\mathcal{S}$, so summing over the outcomes in $[R = x]$ for $x \in \text{range}(R)$ is the same as summing over $\mathcal{S}$. □

In general, the defining sum (1) is better for calculating expected values and has the advantage that it does not depend on the sample space, but only on the density function of the random variable. On the other hand, the simpler sum over all outcomes given in Theorem 1.2 is sometimes easier to use in proofs about expectation.

1.2 Expected Value of an Indicator Variable

The expected value of an indicator random variable for an event is just the probability of that event. (Remember that a random variable I_A is the indicator random variable for event A, if $I_A = 1$ when A occurs and $I_A = 0$ otherwise.)

Lemma 1.3. *If I_A is the indicator random variable for event A, then*

$$\mathrm{E}[I_A] = \Pr\{A\}.$$

Proof.

$$\begin{aligned}
\mathrm{E}[I_A] &= 1 \cdot \Pr\{I_A = 1\} + 0 \cdot \Pr\{I_A = 0\} \\
&= \Pr\{I_A = 1\} \\
&= \Pr\{A\}. && \text{(def of } I_A)
\end{aligned}$$

□

For example, if A is the event that a coin with bias p comes up heads, $\mathrm{E}[I_A] = \Pr\{I_A = 1\} = p$.

1.3 Mean Time to Failure

A computer program crashes at the end of each hour of use with probability p, if it has not crashed already. What is the expected time until the program crashes?

If we let C be the number of hours until the crash, then the answer to our problem is $\mathrm{E}[C]$. Now the probability that, for $i > 0$, the first crash occurs in the ith hour is the probability that it does not crash in each of the first $i-1$ hours and it does crash in the ith hour, which is $(1-p)^{i-1}p$. So from formula (1) for expectation, we have

$$\begin{aligned}
\mathrm{E}[C] &= \sum_{i\in\mathbb{N}} i \cdot \Pr\{R = i\} \\
&= \sum_{i\in\mathbb{N}^+} i(1-p)^{i-1}p \\
&= p\sum_{i\in\mathbb{N}^+} i(1-p)^{i-1} \\
&= p\frac{1}{(1-(1-p))^2} && \text{(by (3))} \\
&= \frac{1}{p}
\end{aligned}$$

As an alternative to applying the formula

$$\sum_{i\in\mathbb{N}^+} ix^{i-1} = \frac{1}{(1-x)^2} \tag{3}$$

from Notes 11 (which you remembered, right?), there is a useful trick for calculating expectations of nonegative integer valued variables:

Lemma 1.4. *If R is a nonegative integer-valued random variable, then:*

$$\mathrm{E}[R] = \sum_{i\in\mathbb{N}} \Pr\{R > i\} \tag{4}$$

Proof. Consider the sum:

$$\begin{array}{ccccccc}
\Pr\{R=1\} & + & \Pr\{R=2\} & + & \Pr\{R=3\} & + & \cdots \\
 & + & \Pr\{R=2\} & + & \Pr\{R=3\} & + & \cdots \\
 & & & + & \Pr\{R=3\} & + & \cdots \\
 & & & & & + & \cdots
\end{array}$$

The successive columns sum to $1\cdot\Pr\{R=1\}$, $2\cdot\Pr\{R=2\}$, $3\cdot\Pr\{R=3\}$, ... Thus, the whole sum is equal to:

$$\sum_{i\in\mathbb{N}} i\cdot\Pr\{R=i\}$$

which equals $\mathrm{E}[R]$ by (1). On the other hand, the successive rows sum to $\Pr\{R>0\}$, $\Pr\{R>1\}$, $\Pr\{R>2\}$, ... Thus, the whole sum is also equal to:

$$\sum_{i\in\mathbb{N}} \Pr\{R>i\},$$

which therefore must equal $\mathrm{E}[R]$ as well. □

Now $\Pr\{C > i\}$ is easy to evaluate: a crash happens later than the ith hour iff the system did not crash during the first i hours, which happens with probability $(1-p)^i$. Plugging this into (4) gives:

$$\begin{aligned}
\mathrm{E}[C] &= \sum_{i\in\mathbb{N}} (1-p)^i \\
&= \frac{1}{1-(1-p)} && \text{(sum of geometric series)} \\
&= \frac{1}{p}
\end{aligned}$$

So, for example, if there is a 1% chance that the program crashes at the end of each hour, then the expected time until the program crashes is $1/0.01 = 100$ hours. The general principle here is well-worth remembering: if a system fails at each time step with probability p, then the expected number of steps up to the first failure is $1/p$.

As a further example, suppose a couple really wants to have a baby girl. For simplicity assume there is a 50% chance that each child they have is a girl, and the genders of their children are mutually independent. If the couple insists on having children until they get a girl, then how many baby boys should they expect first?

This is really a variant of the previous problem. The question, "How many hours until the program crashes?" is mathematically the same as the question, "How many children must the couple have until they get a girl?" In this case, a crash corresponds to having a girl, so we should set $p = 1/2$. By the preceding analysis, the couple should expect a baby girl after having $1/p = 2$ children. Since the last of these will be the girl, they should expect just one boy.

Something to think about: If every couple follows the strategy of having children until they get a girl, what will eventually happen to the fraction of girls born in this world?

1.4 Linearity of Expectation

Expected values obey a simple, very helpful rule called *Linearity of Expectation*. Its simplest form says that the expected value of a sum of random variables is the sum of the expected values of the variables.

Theorem 1.5. *For any random variables R_1 and R_2,*

$$\mathrm{E}[R_1 + R_2] = \mathrm{E}[R_1] + \mathrm{E}[R_2].$$

Proof. Let $T ::= R_1 + R_2$. The proof follows straightforwardly by rearranging terms in the sum (2)

$$\begin{aligned}
\mathrm{E}[T] &= \sum_{\omega\in\mathcal{S}} T(\omega)\cdot \Pr\{\omega\} && \text{(Theorem 1.2)} \\
&= \sum_{\omega\in\mathcal{S}} (R_1(\omega) + R_2(\omega))\cdot \Pr\{\omega\} && \text{(def of } T\text{)} \\
&= \sum_{\omega\in\mathcal{S}} R_1(\omega)\Pr\{\omega\} + \sum_{\omega\in\mathcal{S}} R_2(\omega)\Pr\{\omega\} && \text{(rearranging terms)} \\
&= \mathrm{E}[R_1] + \mathrm{E}[R_2]. && \text{(Theorem 1.2)}
\end{aligned}$$

□

A small extension of this proof, which we leave to the reader, implies

Theorem 1.6 (Linearity of Expectation)**.** *For random variables R_1, R_2 and constants $a_1, a_2 \in \mathbb{R}$,*

$$\mathrm{E}\left[a_1 R_1 + a_2 R_2\right] = a_1 \mathrm{E}\left[R_1\right] + a_2 \mathrm{E}\left[R_2\right].$$

In other words, expectation is a linear function. A routine induction extends the result to more than two variables:

Corollary 1.7. *For any random variables $R_1, \ldots, R_k$ and constants $a_1, \ldots, a_k \in \mathbb{R}$,*

$$\mathrm{E}\left[\sum_{i=1}^{k} a_i R_i\right] = \sum_{i=1}^{k} a_i \mathrm{E}\left[R_i\right].$$

The great thing about linearity of expectation is that *no independence is required*. This is really useful, because dealing with independence is a pain, and we often need to work with random variables that are not independent.

1.4.1 Expected Value of Two Dice

What is the expected value of the sum of two fair dice?

Let the random variable R_1 be the number on the first die, and let R_2 be the number on the second die. We observed earlier that the expected value of one die is 3.5. We can find the expected value of the sum using linearity of expectation:

$$\mathrm{E}\left[R_1 + R_2\right] = \mathrm{E}\left[R_1\right] + \mathrm{E}\left[R_2\right] = 3.5 + 3.5 = 7.$$

Notice that we did *not* have to assume that the two dice were independent. The expected sum of two dice is 7, even if they are glued together (provided the dice remain fair after the gluing). Proving that this expected sum is 7 with a tree diagram would be a bother: there are 36 cases. And if we did not assume that the dice were independent, the job would be really tough!

1.4.2 The Hat-Check Problem

There is a dinner party where n men check their hats. The hats are mixed up during dinner, so that afterward each man receives a random hat. In particular, each man gets his own hat with probability $1/n$. What is the expected number of men who get their own hat?

Letting G be the number of men that get their own hat, we want to find the expectation of G. But all we know about G is that the probability that a man gets his own hat back is $1/n$. There are many different probability distributions of hat permutations with this property, so we don't know enough about the distribution of G to calculate its expectation directly. But linearity of expectation makes the problem really easy.

The trick is to express G as a sum of indicator variables. In particular, let G_i be an indicator for the event that the ith man gets his own hat. That is, $G_i = 1$ if he gets his own hat, and $G_i = 0$ otherwise. The number of men that get their own hat is the sum of these indicators:

$$G = G_1 + G_2 + \cdots + G_n. \tag{5}$$

These indicator variables are *not* mutually independent. For example, if $n-1$ men all get their own hats, then the last man is certain to receive his own hat. But, since we plan to use linearity of expectation, we don't have worry about independence!

Now since G_i is an indicator, we know $1/n = \Pr\{G_i = 1\} = \mathrm{E}[G_i]$ by Lemma 1.3. Now we can take the expected value of both sides of equation (5) and apply linearity of expectation:

$$\begin{aligned}\mathrm{E}[G] &= \mathrm{E}[G_1 + G_2 + \cdots + G_n] \\ &= \mathrm{E}[G_1] + \mathrm{E}[G_2] + \cdots + \mathrm{E}[G_n] \\ &= \frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n} = n\left(\frac{1}{n}\right) = 1.\end{aligned}$$

So even though we don't know much about how hats are scrambled, we've figured out that on average, just one man gets his own hat back!

1.4.3 Expectation of a Binomial Distribution

Suppose that we independently flip n biased coins, each with probability p of coming up heads. What is the expected number that come up heads?

Let J be the number of heads after the flips, so J has the (n,p)-binomial distribution. Now let I_k be the indicator for the kth coin coming up heads. By Lemma 1.3, we have

$$\mathrm{E}[I_k] = p.$$

But

$$J = \sum_{k=1}^{n} I_k,$$

so by linearity

$$\mathrm{E}[J] = \mathrm{E}\left[\sum_{k=1}^{n} I_k\right] = \sum_{k=1}^{n} \mathrm{E}[I_k] = \sum_{k=1}^{n} p = pn.$$

In short, the expectation of an (n,p)-binomially distributed variable is pn.

1.4.4 The Coupon Collector Problem

Every time I purchase a kid's meal at Taco Bell, I am graciously presented with a miniature "Racin' Rocket" car together with a launching device which enables me to project my new vehicle across any tabletop or smooth floor at high velocity. Truly, my delight knows no bounds.

There are n different types of Racin' Rocket car (blue, green, red, gray, etc.). The type of car awarded to me each day by the kind woman at the Taco Bell register appears to be selected uniformly and independently at random. What is the expected number of kid's meals that I must purchase in order to acquire at least one of each type of Racin' Rocket car?

The same mathematical question shows up in many guises: for example, what is the expected number of people you must poll in order to find at least one person with each possible birthday? Here, instead of collecting Racin' Rocket cars, you're collecting birthdays. The general question is commonly called the ***coupon collector problem*** after yet another interpretation.

A clever application of linearity of expectation leads to a simple solution to the coupon collector problem. Suppose there are five different types of Racin' Rocket, and I receive this sequence:

blue green green red blue orange blue orange gray

Let's partition the sequence into 5 segments:

$$\underbrace{\text{blue}}_{X_0}\quad\underbrace{\text{green}}_{X_1}\quad\underbrace{\text{green}\quad\text{red}}_{X_2}\quad\underbrace{\text{blue}\quad\text{orange}}_{X_3}\quad\underbrace{\text{blue}\quad\text{orange}\quad\text{gray}}_{X_4}$$

The rule is that a segment ends whenever I get a new kind of car. For example, the middle segment ends when I get a red car for the first time. In this way, we can break the problem of collecting every type of car into stages. Then we can analyze each stage individually and assemble the results using linearity of expectation.

Let's return to the general case where I'm collecting n Racin' Rockets. Let X_k be the length of the kth segment. The total number of kid's meals I must purchase to get all n Racin' Rockets is the sum of the lengths of all these segments:

$$T = X_0 + X_1 + \cdots + X_{n-1}$$

Now let's focus our attention on X_k, the length of the kth segment. At the beginning of segment k, I have k different types of car, and the segment ends when I acquire a new type. When I own k types, each kid's meal contains a type that I already have with probability k/n. Therefore, each meal contains a new type of car with probability $1 - k/n = (n-k)/n$. Thus, the expected number of meals until I get a new kind of car is $n/(n-k)$ by the "mean time to failure" formula. So we have:

$$\mathrm{E}\left[X_k\right] = \frac{n}{n-k}$$

Linearity of expectation, together with this observation, solves the coupon collector problem:

$$\begin{aligned}
\mathrm{E}\left[T\right] &= \mathrm{E}\left[X_0 + X_1 + \cdots + X_{n-1}\right] \\
&= \mathrm{E}\left[X_0\right] + \mathrm{E}\left[X_1\right] + \cdots + \mathrm{E}\left[X_{n-1}\right] \\
&= \frac{n}{n-0} + \frac{n}{n-1} + \cdots + \frac{n}{3} + \frac{n}{2} + \frac{n}{1} \\
&= n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{3} + \frac{1}{2} + \frac{1}{1}\right) \\
&\quad n\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1} + \frac{1}{n}\right) \\
&= nH_n \sim n\ln n.
\end{aligned}$$

Let's use this general solution to answer some concrete questions. For example, the expected number of die rolls required to see every number from 1 to 6 is:

$$6H_6 = 14.7\ldots$$

And the expected number of people you must poll to find at least one person with each possible birthday is:

$$365H_{365} = 2364.6\ldots$$

1.4.5 The Number-Picking Game

Here is a game that you and I could play that reveals a strange property of expectation.

First, you think of a probability density function on the natural numbers. Your distribution can be absolutely anything you like. For example, you might choose a uniform distribution on $1, 2, \ldots, 6$, like the outcome of a fair die roll. Or you might choose a binomial distribution on $0, 1, \ldots, n$. You can even give every natural number a non-zero probability, provided that the sum of all probabilities is 1.

Next, I pick a random number z according to your distribution. Then, you pick a random number y_1 according to the same distribution. If your number is bigger than mine ($y_1 > z$), then the game ends. Otherwise, if our numbers are equal or mine is bigger ($z \geq y_1$), then you pick a new number y_2 with the same distribution, and keep picking values y_3, y_4, etc. until you get a value that is strictly bigger than my number, z. What is the expected number of picks that you must make?

Certainly, you always need at least one pick, so the expected number is greater than one. An answer like 2 or 3 sounds reasonable, though one might suspect that the answer depends on the distribution. Let's find out whether or not this intuition is correct.

The number of picks you must make is a natural-valued random variable, so from formula (4) we have:

$$\mathrm{E}\left[\#\text{ picks by you}\right] = \sum_{k \in \mathbb{N}} \Pr\left\{(\#\text{ picks by you}) > k\right\} \tag{6}$$

Suppose that I've picked my number z, and you have picked k numbers $y_1, y_2, \ldots, y_k$. There are two possibilities:

- If there is a unique largest number among our picks, then my number is as likely to be it as any one of yours. So with probability $1/(k+1)$ my number is larger than all of yours, and you must pick again.
- Otherwise, there are several numbers tied for largest. My number is as likely to be one of these as any of your numbers, so with probability greater than $1/(k+1)$ you must pick again.

In both cases, with probability at least $1/(k+1)$, you need more than k picks to beat me. In other words:

$$\Pr\left\{(\#\text{ picks by you}) > k\right\} \geq \frac{1}{k+1} \tag{7}$$

This suggests that in order to minimize your rolls, you should choose a distribution such that ties are very rare. For example, you might choose the uniform distribution on $\left\{1, 2, \ldots, 10^{100}\right\}$. In this case, the probability that you need more than k picks to beat me is very close to $1/(k+1)$ for moderate values of k. For example, the probability that you need more than 99 picks is almost exactly 1%. This sounds very promising for you; intuitively, you might expect to win within a reasonable number of picks on average!

Unfortunately for intuition, there is a simple proof that the expected number of picks that you need in order to beat me is *infinite*, regardless of the distribution! Let's plug (7) into (6):

$$\begin{aligned}\mathrm{E}\left[\#\text{ picks by you}\right] &= \sum_{i \in \mathbb{N}} \frac{1}{k+1} \\ &= \infty\end{aligned}$$

This phenomenon can cause all sorts of confusion! For example, suppose you have a communication network where each packet of data has a $1/k$ chance of being delayed by k or more steps. This sounds good; there is only a 1% chance of being delayed by 100 or more steps. But the *expected* delay for the packet is actually infinite!

There is a larger point here as well: not every random variable has a well-defined expectation. This idea may be disturbing at first, but remember that an expected value is just a weighted average. And there are many sets of numbers that have no conventional average either, such as:

$$\{1, -2, 3, -4, 5, -6, \dots\}$$

Strictly speaking, we should qualify virtually all theorems involving expectation with phrases such as "...provided all expectations exist." But we're going to leave that assumption implicit.

Random variables with infinite or ill-defined expectations are more the exception than the rule, but they do creep in occasionally.

1.5 The Expected Value of a Product

While the expectation of a sum is the sum of the expectations, the same is usually not true for products. But it is true in an important special case, namely, when the random variables are *independent*.

For example, suppose we throw two *independent*, fair dice and multiply the numbers that come up. What is the expected value of this product?

Let random variables R_1 and R_2 be the numbers shown on the two dice. We can compute the expected value of the product as follows:

$$\mathrm{E}[R_1 \cdot R_2] = \mathrm{E}[R_1] \cdot \mathrm{E}[R_2] = 3.5 \cdot 3.5 = 12.25. \tag{8}$$

Here the first equality holds because the dice are independent.

At the other extreme, suppose the second die is always the same as the first. Now $R_1 = R_2$, and we can compute the expectation, $\mathrm{E}\left[R_1^2\right]$, of the product of the dice explicitly, confirming that it is not equal to the product of the expectations.

$$\begin{aligned}
\mathrm{E}[R_1 \cdot R_2] &= \mathrm{E}\left[R_1^2\right] \\
&= \sum_{i=1}^{6} i^2 \cdot \Pr\left\{R_1^2 = i^2\right\} \\
&= \sum_{i=1}^{6} i^2 \cdot \Pr\{R_1 = i\} \\
&= \frac{1^2}{6} + \frac{2^2}{6} + \frac{3^2}{6} + \frac{4^2}{6} + \frac{5^2}{6} + \frac{6^2}{6} \\
&= 15\ 1/6 \\
&\neq 12\ 1/4 \\
&= \mathrm{E}[R_1] \cdot \mathrm{E}[R_2].
\end{aligned}$$

Theorem 1.8. *For any two* independent *random variables R_1, R_2,*

$$\mathrm{E}[R_1 \cdot R_2] = \mathrm{E}[R_1] \cdot \mathrm{E}[R_2].$$

Proof. The event $[R_1 \cdot R_2 = r]$ can be split up into events of the form $[R_1 = r_1 \text{ and } R_2 = r_2]$ where $r_1 \cdot r_2 = r$. So

$$\begin{aligned}
&\mathrm{E}[R_1 \cdot R_2] \\
&::= \sum_{r \in \text{range}(R_1 \cdot R_2)} r \cdot \Pr\{R_1 \cdot R_2 = r\} \\
&= \sum_{r_i \in \text{range}(R_i)} r_1 r_2 \cdot \Pr\{R_1 = r_1 \text{ and } R_2 = r_2\} \\
&= \sum_{r_1 \in \text{range}(R_1)} \sum_{r_2 \in \text{range}(R_2)} r_1 r_2 \cdot \Pr\{R_1 = r_1 \text{ and } R_2 = r_2\} && \text{(ordering terms in the sum)} \\
&= \sum_{r_1 \in \text{range}(R_1)} \sum_{r_2 \in \text{range}(R_2)} r_1 r_2 \cdot \Pr\{R_1 = r_1\} \cdot \Pr\{R_2 = r_2\} && \text{(indep. of } R_1, R_2) \\
&= \sum_{r_1 \in \text{range}(R_1)} \left(r_1 \Pr\{R_1 = r_1\} \cdot \sum_{r_2 \in \text{range}(R_2)} r_2 \Pr\{R_2 = r_2\} \right) && \text{(factoring out } r_1 \Pr\{R_1 = r_1\}) \\
&= \sum_{r_1 \in \text{range}(R_1)} r_1 \Pr\{R_1 = r_1\} \cdot \mathrm{E}[R_2] && \text{(def of } \mathrm{E}[R_2]) \\
&= \mathrm{E}[R_2] \cdot \sum_{r_1 \in \text{range}(R_1)} r_1 \Pr\{R_1 = r_1\} && \text{(factoring out } \mathrm{E}[R_2]) \\
&= \mathrm{E}[R_2] \cdot \mathrm{E}[R_1]. && \text{(def of } \mathrm{E}[R_1])
\end{aligned}$$

□

Theorem 1.8 extends routinely to a collection of mutually independent variables.

Corollary 1.9. *If random variables $R_1, R_2, \ldots, R_k$ are mutually independent, then*

$$\mathrm{E}\left[\prod_{i=1}^{k} R_i\right] = \prod_{i=1}^{k} \mathrm{E}[R_i].$$

1.6 Conditional Expectation

Just like event probabilities, expectations can be conditioned on some event.

Definition 1.10. The *conditional expectation*, $\mathrm{E}[R \mid A]$, of a random variable, R, given event, A, is:

$$\mathrm{E}[R \mid A] ::= \sum_{r \in \text{range}(R)} r \cdot \Pr\{R = r \mid A\}. \tag{9}$$

In other words, it is the average value of the variable R when values are weighted by their conditional probabilities given A.

For example, we can compute the expected value of a roll of a fair die, *given*, for example, that the number rolled is at least 4. We do this by letting R be the outcome of a roll of the die. Then by equation (9),

$$\mathrm{E}\left[R \mid R \geq 4\right] = \sum_{i=1}^{6} i \cdot \Pr\left\{R = i \mid R \geq 4\right\} = 1 \cdot 0 + 2 \cdot 0 + 3 \cdot 0 + 4 \cdot \tfrac{1}{3} + 5 \cdot \tfrac{1}{3} + 6 \cdot \tfrac{1}{3} = 5.$$

The power of conditional expectation is that it lets us divide complicated expectation calculations into simpler cases. We can find the desired expectation by calculating the conditional expectation in each simple case and averaging them, weighing each case by its probability.

For example, suppose that 49.8% of the people in the world are male and the rest female —which is more or less true. Also suppose the expected height of a randomly chosen male is $5'\,11''$, while the expected height of a randomly chosen female is $5'\,5''$. What is the expected height of a randomly chosen individual? We can calculate this by averaging the heights of men and women. Namely, let H be the height (in feet) of a randomly chosen person, and let M be the event that the person is male and F the event that the person is female. We have

$$\begin{aligned}\mathrm{E}\left[H\right] &= \mathrm{E}\left[H \mid M\right]\Pr\left\{M\right\} + \mathrm{E}\left[H \mid F\right]\Pr\left\{F\right\} \\ &= (5 + 11/12) \cdot 0.498 + (5 + 5/12) \cdot 0.502 \\ &= 5.665\end{aligned}$$

which is a little less that 5′ 8″.

The Law of Total Expectation justifies this method.

Theorem 1.11 (Law of Total Expectation)**.** *Let $A_1, A_2, \ldots$ be a partition of the sample space. Then*

$$\mathrm{E}\left[R\right] = \sum_i \mathrm{E}\left[R \mid A_i\right]\Pr\left\{A_i\right\}.$$

Proof.

$$\begin{aligned}
\mathrm{E}\left[R\right] &::= \sum_{r \in \text{range}(R)} r \cdot \Pr\left\{R = r\right\} && \text{(Def 1.1 of expectation)} \\
&= \sum_r r \cdot \sum_i \Pr\left\{R = r \mid A_i\right\}\Pr\left\{A_i\right\} && \text{(Law of Total Probability)} \\
&= \sum_r \sum_i r \cdot \Pr\left\{R = r \mid A_i\right\}\Pr\left\{A_i\right\} && \text{(distribute constant } r\text{)} \\
&= \sum_i \sum_r r \cdot \Pr\left\{R = r \mid A_i\right\}\Pr\left\{A_i\right\} && \text{(exchange order of summation)} \\
&= \sum_i \Pr\left\{A_i\right\} \sum_r r \cdot \Pr\left\{R = r \mid A_i\right\} && \text{(factor constant } \Pr\left\{A_i\right\}\text{)} \\
&= \sum_i \Pr\left\{A_i\right\}\mathrm{E}\left[R \mid A_i\right]. && \text{(Def 1.10 of cond. expectation)}
\end{aligned}$$

□

1.6.1 Properties of Conditional Expectation

Many rules for conditional expectation correspond directly to rules for ordinary expectation.

For example, linearity of conditional expectation carries over with the same proof:

Theorem 1.12. *For any two random variables R_1, R_2, constants $a_1, a_2 \in \mathbb{R}$, and event A,*

$$\mathrm{E}\left[a_1 R_1 + a_2 R_2 \mid A\right] = a_1 \mathrm{E}\left[R_1 \mid A\right] + a_2 \mathrm{E}\left[R_2 \mid A\right].$$

Likewise,

Theorem 1.13. *For any two* independent *random variables R_1, R_2, and event, A,*

$$\mathrm{E}\left[R_1 \cdot R_2 \mid A\right] = \mathrm{E}\left[R_1 \mid A\right] \cdot \mathrm{E}\left[R_2 \mid A\right].$$

2 Expect the Mean

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.

More generally, if we independently sample a random variable many times and compute the average of the sample values, then we really can expect this average to be close to the expectation most of the time. In this section we work out a fundamental theorem about how repeated samples of a random variable *deviate from the mean*. This theorem provides an explanation of exactly how sampling can be used to test hypotheses and estimate unknown quantities.

2.1 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 *intelligence quotient*,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 has ever been recorded. But 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, a much smaller fraction than $2/3$ of the population actually has an IQ that high.

But although these conclusions about IQ are weak, they are actually the *strongest possible* general conclusions that can be reached about a nonnegative random variable using *only* the fact that 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 than $1/3$ on the probability of a value ≥ 300.

Theorem 2.1 (Markov's Theorem)**.** *If R is a nonnegative random variable, then for all $x > 0$*

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

Proof. We will show that $\mathrm{E}[R] \geq x \Pr\{R \geq x\}$. Dividing both sides by x gives the desired result.

So let I_x be the indicator variable for the event $[R \geq x]$, and consider the random variable xI_x. Note that

$$R \geq xI_x,$$

because at any sample point, w,

- if $R(\omega) \geq x$ then $R(\omega) \geq x = x \cdot 1 = xI_x(\omega)$, and
- if $R(\omega) < x$ then $R(\omega) \geq 0 = x \cdot 0 = xI_x(\omega)$.

Therefore,

$$\begin{aligned}
\mathrm{E}[R] &\geq \mathrm{E}[xI_x] && (\text{since } R \geq xI_x)\\
&= x\,\mathrm{E}[I_x] && (\text{linearity of } \mathrm{E}[\cdot])\\
&= x \Pr\{I_x = 1\} && (I_x \text{ is an indicator})\\
&= x \Pr\{R \geq x\}. && (\text{def of } I_x)
\end{aligned}$$

□

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

Corollary 2.2. *If R is a nonnegative random variable, then for all $c \geq 1$*

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

Proof. In Markov's Theorem, set $x = c \cdot \mathrm{E}[R]$. □

2.1.1 Applying Markov's Theorem

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 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 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.

2.1.2 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:

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

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 $\mathrm{E}[T] = 50$, 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.

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.

2.2 Chebyshev's Theorem

We have separate versions of Markov's Theorem for the probability of deviation *above* the mean and *below* the mean, but often we want bounds that apply to *distance* from the mean in either direction, that is, bounds on the probability that $|R - \mathrm{E}[R]|$ is large.

It is a bit messy to apply Markov's Theorem directly to this problem, because it's generally not easy to compute $\mathrm{E}[\,|R - \mathrm{E}[R]|\,]$. However, since $|R|$ and hence $|R|^k$ are nonnegative variables for any R, Markov's inequality also applies to the event $[|R|^k \geq x^k]$. But this event is equivalent to the event $[|R| \geq x]$, so we have:

Lemma 2.3. *For any random variable R, any positive integer k, and any $x > 0$,*

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

The special case of this Lemma for $k = 2$ can be applied to bound the random variable, $|R - \mathrm{E}[R]|$, that measures R's deviation from its mean. Namely

$$\Pr\{|R - \mathrm{E}[R]| \geq x\} = \Pr\left\{(R - \mathrm{E}[R])^2 \geq x^2\right\} \leq \frac{\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]}{x^2}, \tag{10}$$

where the inequality (10) follows by applying Lemma 2.3 to the nonnegative random variable, $(R - \mathrm{E}[R])^2$. Assuming that the quantity $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]$ above is finite, we can conclude that the probability that R deviates from its mean by more than x is $O(1/x^2)$.

Definition 2.4. The *variance*, $\mathrm{Var}[R]$, of a random variable, R, is:

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

So we can restate (10) as

Theorem 2.5 (Chebyshev)**.** *Let R be a random variable, and let x be a positive real number. Then*

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

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 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.

2.2.1 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:

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

The expected payoff is the same for both games, but they are obviously very different! This difference is not apparent in their expected value, but is captured by variance. We can compute the $\mathrm{Var}[A]$ by working "from the inside out" as follows:

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

Similarly, we have for $\mathrm{Var}[B]$:

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

The variance of Game A is 2 and the variance of Game B is more than two million! Intuitively, this means that the payoff 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!

2.3 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

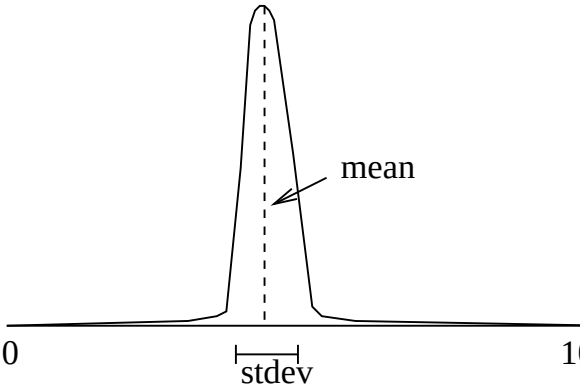

Figure 1: The standard deviation of a distribution indicates how wide the "main part" of it is.

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 random variables using standard deviation instead of variance.

Definition 2.6. The *standard deviation*, σ_R, of a random variable, R, is the square root of the variance:

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

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 "expected (average) deviation from the mean," since we can think of the square root on the outside as canceling the square on the inside.

Example 2.7. 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; therefore, the standard deviation of 1416 describes this situation reasonably well.

Intuitively, the standard deviation measures the "width" of the "main part" of the distribution graph, as illustrated in Figure 1.

There is a useful, simple reformulation of Chebyshev's Theorem in terms of standard deviation.

Corollary 2.8. *Let R be a random variable, and let c be a positive real number.*

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

Here we see explicitly how the "likely" values of R are clustered in an $O(\sigma_R)$-sized region around $\text{E}\,[R]$, 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:

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

□

2.3.1 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 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 2.1 gives a coarse bound, namely,

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

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

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

The purpose of the first step is to express the desired probability in the form required by Chebyshev's Theorem; the equality holds because R is nonnegative. Chebyshev's Theorem then yields the inequality.

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.

2.4 Properties of Variance

The definition of variance of R as $\mathrm{E}\left[(R - \mathrm{E}[R])^2\right]$ may seem rather arbitrary. A direct measure of average deviation would be $\mathrm{E}[\,|R - \mathrm{E}[R]|\,]$. But variance has some valuable mathematical properties which the direct measure does not, as we explain below.

2.4.1 A Formula for Variance

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

Theorem 2.9.

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

for any random variable, R.

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

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

$$\begin{aligned}
\mathrm{Var}[R] &= \mathrm{E}\left[(R - \mathrm{E}[R])^2\right] && \text{(Def 2.4 of variance)} \\
&= \mathrm{E}\left[(R-\mu)^2\right] && \text{(def of } \mu\text{)} \\
&= \mathrm{E}\left[R^2 - 2\mu R + \mu^2\right] \\
&= \mathrm{E}\left[R^2\right] - 2\mu\,\mathrm{E}[R] + \mu^2 && \text{(linearity of expectation)} \\
&= \mathrm{E}\left[R^2\right] - 2\mu^2 + \mu^2 && \text{(def of } \mu\text{)} \\
&= \mathrm{E}\left[R^2\right] - \mu^2 \\
&= \mathrm{E}\left[R^2\right] - \mathrm{E}^2[R]\,. && \text{(def of } \mu\text{)}
\end{aligned}$$

□

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

$$\mathrm{Var}[B] = p - p^2 = p(1-p). \tag{11}$$

Proof. Since B only takes values 0 and 1, we have $\mathrm{E}[B] = p \cdot 1 + (1-p)\cdot 0 = p$. Since $B = B^2$, we also have $\mathrm{E}\left[B^2\right] = p$, so (11) follows immediately from Theorem 2.9. □

2.4.2 Dealing with Constants

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

Theorem 2.10. *Let R be a random variable, and a a constant. Then*

$$\mathrm{Var}[aR] = a^2\,\mathrm{Var}[R]\,. \tag{12}$$

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

$$\begin{aligned}
\mathrm{Var}[aR] ::= \mathrm{E}\left[(aR - \mathrm{E}[aR])^2\right] \\
&= \mathrm{E}\left[(aR)^2 - 2aR\,\mathrm{E}[aR] + \mathrm{E}^2[aR]\right] \\
&= \mathrm{E}\left[(aR)^2\right] - \mathrm{E}\left[2aR\,\mathrm{E}[aR]\right] + \mathrm{E}^2[aR] \\
&= a^2\,\mathrm{E}\left[R^2\right] - 2\,\mathrm{E}[aR]\,\mathrm{E}[aR] + \mathrm{E}^2[aR] \\
&= a^2\,\mathrm{E}\left[R^2\right] - a^2\mathrm{E}^2[R] \\
&= a^2\left(\mathrm{E}\left[R^2\right] - \mathrm{E}^2[R]\right) \\
&= a^2\,\mathrm{Var}[R] && \text{(by Theorem 2.9)}
\end{aligned}$$

□

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

Theorem 2.11. *Let R be a random variable, and b a constant. Then*

$$\mathrm{Var}[R + b] = \mathrm{Var}[R]\,. \tag{13}$$

Recalling that the standard deviation is the square root of variance, we immediately get:

Corollary 2.12. *The standard deviation of $aR + b$ equals a times the standard deviation of R:*

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

2.4.3 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 involving variables that are pairwise independent but not mutually independent. Matching birthdays is an example of this kind, as we shall see below.

Theorem 2.13. *[Pairwise Independent Additivity of Variance] If $R_1, R_2, \ldots, R_n$ are* pairwise *independent random variables, then*

$$\text{Var}[R_1 + R_2 + \cdots + R_n] = \text{Var}[R_1] + \text{Var}[R_2] + \cdots + \text{Var}[R_n]. \tag{14}$$

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 (14). This substitution preserves the independence of the variables, and by Theorem 2.11, does not change the variances.

Now by Theorem 2.9, $\text{Var}[R_i] = \text{E}\left[R_i^2\right]$ and $\text{Var}[R_1 + R_2 + \cdots + R_n] = \text{E}\left[(R_1 + R_2 + \cdots + R_n)^2\right]$, so we need only prove

$$\text{E}\left[(R_1 + R_2 + \cdots + R_n)^2\right] = \text{E}\left[R_1^2\right] + \text{E}\left[R_2^2\right] + \cdots + \text{E}\left[R_n^2\right] \tag{15}$$

But (15) follows from linearity of expectation and the fact that

$$\text{E}[R_i R_j] = \text{E}[R_i]\,\text{E}[R_j] = 0 \cdot 0 = 0 \tag{16}$$

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

$$\begin{aligned}
\text{E}\left[(R_1 + R_2 + \cdots + R_n)^2\right] &= \text{E}\left[\sum_{1\le i,j\le n} R_i R_j\right] \\
&= \sum_{1\le i,j\le n} \text{E}[R_i R_j] \\
&= \sum_{1\le i\le n} \text{E}\left[R_i^2\right] + \sum_{1\le i\ne j\le n} \text{E}[R_i R_j] \\
&= \sum_{1\le i\le n} \text{E}\left[R_i^2\right] + \sum_{1\le i\ne j\le n} 0 \qquad \text{(by (16))} \\
&= \text{E}\left[R_1^2\right] + \text{E}\left[R_2^2\right] + \cdots + \text{E}\left[R_n^2\right].
\end{aligned}$$

□

Now we have a simple way of computing the expectation of a variable J which has an (n,p)-binomial distribution. We know that $J = \sum_{k=1}^{n} I_k$ where the I_k are mutually independent 0-1-valued variables with $\Pr\{I_k = 1\} = p$. The variance of each I_k is $p(1-p)$ by (11), so by linearity of variance, we have

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

$$\operatorname{Var}[J] = n \operatorname{Var}[I_k] = np(1-p). \tag{17}$$

2.5 Estimation by Random Sampling

2.5.1 Polling again

In Notes 12, we used bounds on the binomial distribution to determine confidence levels for a poll of voter preferences of Clinton vs. Giuliani. 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 Notes 12: we will poll n randomly chosen voters and let S_n be the total number in our sample who preferred Clinton. We use S_n/n as our estimate of the actual fraction, p, of all voters who prefer Clinton. We want to choose n so that our estimate will be within 0.04 of p at least 95% of the time.

Now S_n is binomially distributed, so from (17) we have

$$\operatorname{Var}[S_n] = n(p(1-p)) \le n \cdot \frac{1}{4} = \frac{n}{4}$$

The bound of $1/4$ follows from the easily verified fact that $p(1-p)$ is maximized when $p = 1-p$, that is, when $p = 1/2$.

Next, we bound the variance of S_n/n:

$$\begin{aligned}
\operatorname{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \operatorname{Var}[S_n] && \text{(by (12))} \\
&\le \left(\frac{1}{n}\right)^2 \frac{n}{4} && \text{(by (2.5.1))} \\
&= \frac{1}{4n}
\end{aligned} \tag{18}$$

Now from Chebyshev and (18) we have:

$$\Pr\left\{\left|\frac{S_n}{n} - p\right| \ge 0.04\right\} \le \frac{\operatorname{Var}[S_n/n]}{(0.04)^2} = \frac{1}{4n(0.04)^2} = \frac{156.25}{n} \tag{19}$$

To make our our estimate with 95% confidence, we want the righthand side of (19) to be at most $1/20$. So we choose n so that

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

that is,

$$n \ge 3,125.$$

You may remember that in Notes 12 we calculated that it was actually sufficient to poll only 664 voters —many fewer than the 3,125 voters we derived using Chebyshev's Theorem. So the bound from Chebyshev's Theorem is not nearly as good as the bound we got earlier. This should not be surprising. In applying the Chebyshev Theorem, we used only a bound on the variance of S_n. In Notes 12, on the other hand, we used the fact that the random variable S_n was binomial (with known parameter, n, and unknown parameter, p). It makes sense that more detailed information about a distribution leads to better bounds. But even though the bound was not as good, this example nicely illustrates an approach to estimation using Chebyshev's Theorem that is more widely applicable than binomial estimations.

2.5.2 Birthdays again

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've already seen that in a class of one hundred or more, there is a very high probability that some pair of students have birthdays on the same day of the month. We can also easily calculate the expected number of pairs of students with matching birthdays. But is it likely the number of matching pairs in a typical class will actually be close to the expected number? We can take the same approach to answering this question as we did in estimating voter preferences.

But notice that having matching birthdays for different pairs of students are not mutually independent events. For example, knowing that Alice and Bob have matching birthdays, and also that Ted and Alice have matching birthdays obviously implies that Bob and Ted have matching birthdays. On the other hand, knowing that Alice and Bob have matching birthdays tells us nothing about whether Alice and Carol have matching birthdays, namely, these two events really are independent. So even though the events that various pairs of students have matching birthdays are not mutually independent, indeed not even three-way independent, they are *pairwise* independent.

This allows us to apply the same reasoning to Birthday Matching as we did for voter preference. Namely, let $B_1, B_2, \ldots, B_n$ be the birthdays of n independently chosen people, and let $E_{i,j}$ be the indicator variable for the event that the ith and jth people chosen have the same birthdays, that is, the event $[B_i = B_j]$. For simplicity, we'll assume that for $i \neq j$, the probability that $B_i = B_j$ is $1/365$. So the B_i's are mutually independent variables, and hence the $E_{i,j}$'s are *pairwise* independent variables, which is all we will need.

Let D be the number of matching pairs of birthdays among the n choices, that is,

$$D ::= \sum_{1 \leq i < j \leq n} E_{i,j}. \tag{20}$$

So by linearity of expectation

$$\mathrm{E}[D] = \mathrm{E}\left[\sum_{1 \leq i < j \leq n} E_{i,j}\right] = \sum_{1 \leq i < j \leq n} \mathrm{E}[E_{i,j}] = \binom{n}{2} \cdot \frac{1}{365}.$$

Also, by Theorem 2.13, the variances of pairwise independent variables are additive, so

$$\text{Var}[D] = \text{Var}\left[\sum_{1\le i<j\le n} E_{i,j}\right] = \sum_{1\le i<j\le n} \text{Var}[E_{i,j}] = \binom{n}{2}\cdot\frac{1}{365}\left(1-\frac{1}{365}\right).$$

Now for a class of $n = 100$ students, we have $\text{E}[D] \approx 14$ and $\text{Var}[D] < 14(1 - 1/365) < 14$. So by Chebyshev's Theorem

$$\Pr\{|D - 14| \ge x\} < \frac{14}{x^2}.$$

Letting $x = 6$, we conclude that there is a better than 50% chance that in a class of 100 students, the number of pairs of students with the same birthday will be between 8 and 20.

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

$$S_n ::= \sum_{i=1}^{n} G_i. \tag{21}$$

Then

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

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

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

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

$$\begin{aligned}
\text{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \text{Var}\left[S_n\right] && \text{(by (12))} \\
&= \frac{1}{n^2}\,\text{Var}\left[\sum_{i=1}^{n} G_i\right] && \text{(def of } S_n\text{)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left[G_i\right] && \text{(pairwise independent additivity)} \\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. && (22)
\end{aligned}$$

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

$$\begin{aligned}
\Pr\left\{\left|\frac{S_n}{n}-\mu\right| \geq x\right\} &\leq \frac{\text{Var}\left[S_n/n\right]}{x^2}. && \text{(Chebyshev's bound)} \\
&= \frac{\sigma^2/n}{x^2} && \text{(by (22))} \\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

□

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[1] of Large Numbers: by choosing a large enough sample size, n, we can get arbitrarily accurate estimates of the mean with confidence arbitrarily close to 100%.

[1]This is the *Weak* Law of Large Numbers. As you might suppose, there is also a Strong Law, but it's outside the scope of 6.042.