Missed Expectations?

In the previous notes, we saw that the average value of a random quantity is captured by the mathematical concept of the expectation of a random variable, and we calculated expectations for several kinds of random variables. Now we will see two things that make expectations so useful. First, they are often very easy to calculate due to the fact that they obey linearity. Second, once you know what the expectation is, you can also get some type of bound on the probability that you are far from the expectation —that is, you can show that really weird things are not that likely to happen. How good a bound you can get depends on what you know about your distribution, but don't worry, even if you know next to nothing, you can still say something relatively interesting.

1 Linearity of Expectation

1.1 Expectation of a Sum

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 1.1. *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 from the definition of $\mathrm{E}[T]$.

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

□

Similarly, we have

Lemma 1.2. *For any random variable R and constant $a \in \mathbb{R}$,*

$$\mathrm{E}[aR] = a\,\mathrm{E}[R].$$

The proof follows easily from the definition of expectation, and we omit it.

Combining Theorem 1.1 and Lemma 1.2, we conclude

Theorem 1.3 (Linearity of Expectation). *For all random variables R_1, R_2 and constants $a_1, a_2 \in \mathbb{R}$,*

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

In other words, expectation is a linear function. The rule and its proof extends directly to cover more than two random variables:

Corollary 1.4. *For any random variables $R_1, \ldots, R_k$ and constants $a_1, \ldots, a_k \in \mathbb{R}$,*

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

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.

1.2 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:

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

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 connected together![1]

Proving that the expected sum is 7 with a tree diagram would be hard; there are 36 cases. And if we did not assume that the dice were independent, the job would be a nightmare!

[1] But each die must remain fair after the connection.

1.3 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?

Without linearity of expectation, this would be a very difficult question to answer. We might try the following. Let the random variable R be the number of men that get their own hat. We want to compute $\mathrm{E}[R]$. By the definition of expectation, we have:

$$\mathrm{E}[R] = \sum_{k=0}^{\infty} k \cdot \Pr\{R = k\}$$

Now we're in trouble, because evaluating $\Pr\{R = k\}$ is a mess and we then need to substitute this mess into a summation. Furthermore, to have any hope, we would need to fix the probability of each permutation of the hats. For example, we might assume that all permutations of hats are equally likely.

Now let's try to use linearity of expectation. As before, let the random variable R be the number of men that get their own hat. The trick is to express R as a sum of indicator variables. In particular, let R_i be an indicator for the event that the ith man gets his own hat. That is, $R_i = 1$ is the event that he gets his own hat, and $R_i = 0$ is the event that he gets the wrong hat. The number of men that get their own hat is the sum of these indicators:

$$R = R_1 + R_2 + \cdots + R_n$$

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!

Let's take the expected value of both sides of the equation above and apply linearity of expectation:

$$\begin{aligned}
\mathrm{E}[R] &= \mathrm{E}[R_1 + R_2 + \cdots + R_n] \\
&= \mathrm{E}[R_1] + \mathrm{E}[R_2] + \cdots + \mathrm{E}[R_n]
\end{aligned}$$

Since the R_i's are indicator variables, $\mathrm{E}[R_i] = \Pr\{R_i\}$ and since every man is as likely to get one hat as another, this is just $1/n$. Putting all this together, we have:

$$\begin{aligned}
\mathrm{E}[R] &= \mathrm{E}[R_1] + \mathrm{E}[R_2] + \cdots + \mathrm{E}[R_n] \\
&= \Pr\{R_1 = 1\} + \Pr\{R_2 = 1\} + \cdots + \Pr\{R_n = 1\} \\
&= n \cdot \frac{1}{n} = 1.
\end{aligned}$$

So we should expect 1 man to get his own hat back on average!

Notice that we did not assume that all permutations of hats are equally likely or even that all permutations are possible. We only needed to know that each man received his own hat with probability $1/n$.

1.4 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 $H_{n,p}$ be the number of heads after the flips. Then $H_{n,p}$ has the binomial distribution with parameters n and p. Now let I_k be the indicator for the kth coin coming up heads. Since I_k is an indicator variable with probability p of being 1, we know that

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

But

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

so by linearity

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

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

2 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 kids 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.

2.1 A Solution Using Linearity of Expectation

Linearity of expectation is somewhat like induction and the pigeonhole principle; it's a simple idea that can be used in all sorts of ingenious ways. For example, we can use linearity of expecatation in a clever way to solve 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 k-th 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 of the X_k, the length of the k-th 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 that we worked out last time. So we have:

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

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

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

The summation on the next-to-last line is the n-th harmonic sum with the terms in reverse order. As you may recall, this sum is denoted H_n and is approximately $\ln n$.

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:

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

3 Conditional Expectation

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

Definition 3.1. We define the *conditional expectation* $\mathrm{E}[R \mid A]$ *of a random variable* R *given event* A:

$$\mathrm{E}[R \mid A] ::= \sum_r r \cdot \Pr\{R = r \mid A\}.$$

In other words, it is the expected value of the variable R once we skew the distribution of R to be conditioned on event A.

Example 3.2. Let D be the outcome of a roll of a fair die. What is $\mathrm{E}[D \mid D \geq 4]$?

$$\sum_{i=1}^{6} i \cdot \Pr\{D = i \mid D \geq 4\} = 1 \cdot 0 + 2 \cdot 0 + 3 \cdot 0 + 4 \cdot \tfrac{1}{3} + 5 \cdot \tfrac{1}{3} + 6 \cdot \tfrac{1}{3} = 5.$$

It is easy to see that the rules for expectation will extend to conditional expectation. For example, conditional expectation will also be linear.

Theorem 3.3. *For any two random variables R_1, R_2, constants $a_1, a_2 \in \mathbb{R}$, and event A,*

$$\mathrm{E}[a_1R_1 + a_2R_2 \mid A] = a_1\,\mathrm{E}[R_1 \mid A] + a_2\,\mathrm{E}[R_2 \mid A].$$

A real benefit of conditional expectation is the way it lets us divide complicated expectation calculations into simpler cases.

Theorem 3.4 (Law of Total Expectation). *If the sample space is the union of the pairwise disjoint events $A_1, A_2, \ldots$, then*

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

Proof.

$$\begin{aligned}
\mathrm{E}[R] &= \sum_r r \cdot \Pr\{R = r\} \\
&= \sum_r r \cdot \sum_i \Pr\{R = r \mid A_i\} \Pr\{A_i\} && \text{(Law of Total Probability)} \\
&= \sum_r \sum_i r \cdot \Pr\{R = r \mid A_i\} \Pr\{A_i\} && \text{(distribute constant } r\text{)} \\
&= \sum_i \sum_r r \cdot \Pr\{R = r \mid A_i\} \Pr\{A_i\} && \text{(exchange order of summation)} \\
&= \sum_i \Pr\{A_i\} \sum_r r \cdot \Pr\{R = r \mid A_i\} && \text{(factor constant } \Pr\{A_i\}\text{)} \\
&= \sum_i \Pr\{A_i\} \mathrm{E}[R \mid A_i]. && \text{(Def. 3.1)}
\end{aligned}$$

□

Example 3.5. Half the people in the world are male, half female. 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?

Let $H(P)$ be the height of the random person P. The events M::="P is male" and F::="P is female" are a partition of the sample space. Then

$$\begin{aligned}
\mathrm{E}[H] &= \mathrm{E}[H \mid M] \Pr\{M\} + \mathrm{E}[H \mid F] \Pr\{F\} \\
&= 5'11'' \cdot \frac{1}{2} + 5'5'' \cdot \frac{1}{2} \\
&= 5'8''.
\end{aligned}$$

We will see in the following sections that the Law of Total Expectation has much more power than one might think.

4 The Expected Value of a Product

4.1 The Product of Independent Expectations

We have determined that the expectation of a sum is the sum of the expectations. The same is not always true for products: in general, the expectation of a product need *not* equal the product of the expectations. But it is true in an important special case, namely, when the random variables are *independent*.

Lemma 4.1. *If R_1 and R_2 are independent random variables, then*

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

The Lemma follows immediately from Definition 3.1 of conditional expectation and the fact that $\Pr\{R_1 = r\} = \Pr\{R_1 = r \mid R_2 = a\}$.

Theorem 4.2. *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. We apply the Law of Total Expectation by conditioning on the value of R_1.

$$\begin{aligned}
\mathrm{E}[R_1 \cdot R_2] &= \sum_{r \in \mathrm{range}(R_1)} \mathrm{E}[R_1 \cdot R_2 \mid R_1 = r] \cdot \Pr\{R_1 = r\} && \text{(Thm 3.4)} \\
&= \sum_r \mathrm{E}[r \cdot R_2 \mid R_1 = r] \cdot \Pr\{R_1 = r\} \\
&= \sum_r r \cdot \mathrm{E}[R_2 \mid R_1 = r] \cdot \Pr\{R_1 = r\} && \text{(Lemma 1.2)} \\
&= \sum_r r \cdot \mathrm{E}[R_2] \cdot \Pr\{R_1 = r\} && \text{Lemma 4.1} \\
&= \mathrm{E}[R_2] \sum_r r \cdot \Pr\{R_1 = r\} && \text{(factor out constant } \mathrm{E}[R_2]) \\
&= \mathrm{E}[R_2] \cdot \mathrm{E}[R_1].
\end{aligned}$$

□

Theorem 4.2 extends routinely to a collection of mutually independent variables.

Corollary 4.3. *If random variables $R_1, R_2, \ldots, R_k$ are mutually independent, then*

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

4.2 The Product of Two Dice

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:

$$\mathrm{E}[R_1 \cdot R_2] = \mathrm{E}[R_1] \cdot \mathrm{E}[R_2] = 3.5 \cdot 3.5 = 12.25. \tag{1}$$

Here the first equality holds by Theorem 4.2 because the dice are independent.

Now suppose that the two dice are *not* independent; in fact, assume that the second die is always the same as the first. In this case, the product of expectations will not equal the expectation of the product.

To verify this, 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 without Theorem 4.2 as follows:

$$\begin{aligned}
\mathrm{E}[R_1 \cdot R_2] &= \mathrm{E}\left[R_1^2\right] && (R_2 = R_1)\\
&= \sum_{i=1}^{6} i^2 \cdot \Pr\left\{R_1^2 = i^2\right\}\\
&= \sum_{i=1}^{6} i^2 \cdot \Pr\{R_1 = i\}\\
&= \frac{1^2}{6} + \frac{2^2}{6} + \frac{3^2}{6} + \frac{4^2}{6} + \frac{5^2}{6} + \frac{6^2}{6}\\
&= \frac{91}{6}\\
&\neq 12.25\\
&= \mathrm{E}[R_1] \cdot \mathrm{E}[R_2]. && \text{from (1)}
\end{aligned}$$

5 Expect the Mean

We have seen several examples of random variables that never take a value equal to their mean. But experience suggests that we can *expect* the values of a variable to be *near* its mean – usually – which is why the mean is also called the "expectation." In other words, the values of a random variable *probably* won't deviate *very much* from the mean. We will describe some basic results about this central topic of *deviation from the mean*, and we will indicate how these results apply for testing hypotheses and estimating by sampling.

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

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

6 Markov's Theorem

Markov's theorem 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. IQ was devised so that the average IQ measurement would be 100. Now from this fact alone we can conclude that at most 1/2 the population can have an IQ of 200 or more, because if more than half had an IQ of 200, then the average would have to be more than $(1/2)200 = 100$, contradicting the fact that the average is 100. So the probability that a randomly chosen person has an IQ of 200 or more is at most 1/2. Of course this is not a very strong conclusion; in fact no IQ of over 200 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 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 200 with probability 1/2, and 0 with probability 1/2, then its mean is 100, and the probability of a value of 200 or more is really 1/2. So we can't hope to get a upper better bound on the probability of 200 than 1/2.

Theorem 6.1 (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 $R(w) \geq x$ then $R(w) \geq x = x \cdot 1 = xI_x(w)$, and
- if $R(w) < x$ then $R(w) \geq 0 = x \cdot 0 = xI_x(w)$.

Therefore,

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

□

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

Corollary 6.2. *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 $x = c \cdot \mathrm{E}[R]$. □

6.1 Examples of Markov's Theorem

Suppose that n men go to a dinner party and check their hats. At the end of the night, the hats are randomly permuted and returned, so each man gets his own hat back with probability $1/n$. 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 previous notes, we used linearity of expectation to show that $\mathrm{E}[R] = 1$. By Markov's Theorem, the probability that x or more men get the right hat is:

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

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 $1/n$.

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. You can show that $\mathrm{E}[R] = 1$ (right?). 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. But in reality, the probability of this event is $1/(n!)$. So for the hat check problem, Markov's Theorem case gives probability bounds that are way off.

6.2 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 by 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.

7 Chebyshev's Theorem

We have versions of Markov's Theorem for the probability of deviation *above* 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 7.1. *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

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

where the inequality (2) follows by applying Lemma 7.1 to the nonnegative random variable, $(R - \mathrm{E}[R])^2$. Assuming that the quantity $\mathrm{E}[(R - \mathrm{E}[R])^2]$ 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 7.2. 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 (2) as

Theorem 7.3 (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 $\mathrm{E}[(R - \mathrm{E}[R])^2]$ for variance is a bit cryptic; the best approach is to work through it from the inside out. The innermost expression, $R - \mathrm{E}[R]$, 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.

7.1 Variance in Two Gambling Games

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

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

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

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

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

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 $\text{Var}[A]$ by working "from the inside out" as follows:

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

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

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

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!

7.2 Standard Deviation

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

Definition 7.4. The *standard deviation*, σ_R, of a random variable, R, is the square root of the variance:

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

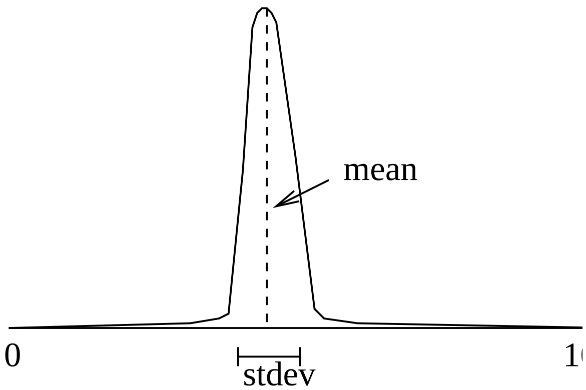

Figure 1: The standard deviation of a distribution indicates how wide the "main part" of it is.

So the standard deviation is the square root of the mean 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 7.5. The standard deviation of the payoff in Game B is:

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

The random variable B actually deviates from the mean by either positive 1001 or negative 2002; 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 7.6. *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:

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

□

7.3 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 200 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 $\Pr\{R \geq 200\}$.

We have already seen that Markov's Theorem 6.1 gives a coarse bound, namely,

$$\Pr\{R \geq 200\} \leq \frac{1}{2}.$$

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

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

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 a hundred has an IQ of 200 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.

8 Properties of Variance

8.1 Why Variance?

The definition of variance of R as $\mathrm{E}[(R - \mathrm{E}[R])^2]$ may seem rather arbitrary. The variance is the average *of the square* of the deviation from the mean. For this reason, variance is sometimes called the "mean squared deviation." But why bother squaring? Why not simply compute the average deviation from the mean? That is, why not define variance to be $\mathrm{E}[R - \mathrm{E}[R]]$?

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

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

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

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

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 canceling by taking an absolute value. That is, we could define variance to be $\mathrm{E}[|R - \mathrm{E}[R]|]$. There is no logical reason not to use this definition. However, the conventional version of variance has some valuable mathematical properties which the absolute value version does not. We describe these properties in the following sections and use them to determine the variance of some important probability distributions.

8.2 An Alternative Definition of Variance

There is an equivalent way to define the variance of a random variable that is less intuitive, but is often easier to use in calculations and proofs:

Theorem 8.1.
$$\mathrm{Var}[R] = \mathrm{E}\left[R^2\right] - \mathrm{E}^2[R],$$
for any random variable, R.

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

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

Proof. Let $\mu = \mathrm{E}[R]$. Then
$$\begin{aligned}
\mathrm{Var}[R] &= \mathrm{E}\left[(R - \mathrm{E}[R])^2\right] && \text{(Def. 7.2 of variance)}\\
&= \mathrm{E}\left[(R-\mu)^2\right] && \text{(def. of } \mu)\\
&= \mathrm{E}\left[R^2 - 2\mu R + \mu^2\right] \\
&= \mathrm{E}\left[R^2\right] - 2\mu\,\mathrm{E}[R] + \mu^2 && \text{(linearity of expectation)}\\
&= \mathrm{E}\left[R^2\right] - 2\mu^2 + \mu^2 && \text{(def. of } \mu)\\
&= \mathrm{E}\left[R^2\right] - \mu^2 \\
&= \mathrm{E}\left[R^2\right] - \mathrm{E}^2[R]. && \text{(def. of } \mu)
\end{aligned}$$
□

For example, if B is a Bernoulli variable where $p ::= \Pr\{B = 1\}$, then
$$\mathrm{Var}[B] = p - p^2 = p(1-p). \tag{3}$$

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 $\mathrm{E}[B^2] = p$, so (3) follows immediately from (8.1). □

8.2.1 Zero Variance

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

Lemma 8.2. *The variance of a random variable, R, is zero if and only if* $\Pr\{R = \mathrm{E}[R]\} = 1$.

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

Proof. By the definition of variance,

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

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

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

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

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

□

8.2.2 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 8.3. *Let R be a random variable, and let a and b be constants. Then*

$$\mathrm{Var}[aR + b] = a^2 \,\mathrm{Var}[R]. \tag{4}$$

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 changes the variance by a *square factor*.

Proof. We will transform the left side of (4) into the right side. The first step is to expand $\mathrm{Var}[aR + b]$ using the alternate definition of variance.

$$\mathrm{Var}[aR + b] = \mathrm{E}\left[(aR + b)^2\right] - \mathrm{E}^2[aR + b].$$

We will work on the first term and then the second term. For the first term, note that by linearity of expectation,

$$\mathrm{E}\left[(aR+b)^2\right] = \mathrm{E}\left[a^2R^2 + 2abR + b^2\right] = a^2\,\mathrm{E}\left[R^2\right] + 2ab\,\mathrm{E}\left[R\right] + b^2. \tag{5}$$

Similarly for the second term:

$$\mathrm{E}^2\left[aR+b\right] = (a\,\mathrm{E}\left[R\right] + b)^2 = a^2\mathrm{E}^2\left[R\right] + 2ab\,\mathrm{E}\left[R\right] + b^2. \tag{6}$$

Finally, we subtract the expanded second term from the first.

$$\begin{aligned}
\mathrm{Var}\left[aR+b\right] &= \mathrm{E}\left[(aR+b)^2\right] - \mathrm{E}^2\left[aR+b\right] && \text{(Theorem 8.1)}\\
&= a^2\,\mathrm{E}\left[R^2\right] + 2ab\,\mathrm{E}\left[R\right] + b^2 - \\
&\quad (a^2\mathrm{E}^2\left[R\right] + 2ab\,\mathrm{E}\left[R\right] + b^2) && \text{(by (5) and (6))}\\
&= a^2\,\mathrm{E}\left[R^2\right] - a^2\mathrm{E}^2\left[R\right]\\
&= a^2(\mathrm{E}\left[R^2\right] - \mathrm{E}^2\left[R\right])\\
&= a^2\,\mathrm{Var}\left[R\right] && \text{(Theorem 8.1)}
\end{aligned}$$

□

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. So we have

Corollary 8.4. *The standard deviation of $aR+b$ equals a times the standard deviation of R.*

8.3 Variance of a Sum

Earlier, we claimed that for independent random variables, the variance of a sum is the sum of the variances.

An independence condition is necessary. If we ignored independence, then we would conclude that $\mathrm{Var}\left[R+R\right] = \mathrm{Var}\left[R\right] + \mathrm{Var}\left[R\right]$. However, by Theorem 8.3, the left side is equal to $4\,\mathrm{Var}\left[R\right]$, whereas the right side is $2\,\mathrm{Var}\left[R\right]$. This implies that $\mathrm{Var}\left[R\right] = 0$, which, by Lemma 8.2, essentially only holds if R is constant.

However, *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 8.5. *[Pairwise Independent Additivity of Variance] If $R_1, R_2, \ldots, R_n$ are* pairwise *independent random variables, then*

$$\mathrm{Var}\left[R_1 + R_2 + \cdots + R_n\right] = \mathrm{Var}\left[R_1\right] + \mathrm{Var}\left[R_2\right] + \cdots + \mathrm{Var}\left[R_n\right].$$

Proof. By linearity of expectation, we have

$$\begin{aligned}
\mathrm{E}\left[\left(\sum_{i=1}^{n} R_i\right)^2\right] &= \mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} R_i R_j\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}\left[R_i R_j\right] && \text{(linearity)} \\
&= \sum_{1 \le i \ne j \le n} \mathrm{E}\left[R_i\right]\mathrm{E}\left[R_j\right] + \sum_{i=1}^{n} \mathrm{E}\left[R_i^2\right]. && \text{(pairwise independence)} \qquad (7)
\end{aligned}$$

In (7), we use the fact that the expectation of the product of two independent variables is the product of their expectations.

Also,

$$\begin{aligned}
\mathrm{E}^2\left[\sum_{i=1}^{n} R_i\right] &= \left(\mathrm{E}\left[\sum_{i=1}^{n} R_i\right]\right)^2 \\
&= \left(\sum_{i=1}^{n} \mathrm{E}\left[R_i\right]\right)^2 && \text{(linearity)} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}\left[R_i\right]\mathrm{E}\left[R_j\right] \\
&= \sum_{1 \le i \ne j \le n} \mathrm{E}\left[R_i\right]\mathrm{E}\left[R_j\right] + \sum_{i=1}^{n} \mathrm{E}^2\left[R_i\right]. && (8)
\end{aligned}$$

So,

$$\begin{aligned}
\mathrm{Var}\left[\left(\sum_{i=1}^{n} R_i\right)\right] &= \mathrm{E}\left[\left(\sum_{i=1}^{n} R_i\right)^2\right] - \mathrm{E}^2\left[\sum_{i=1}^{n} R_i\right] && \text{(Theorem 8.1)} \\
&= \sum_{1 \le i \ne j \le n} \mathrm{E}\left[R_i\right]\mathrm{E}\left[R_j\right] + \sum_{i=1}^{n} \mathrm{E}\left[R_i^2\right] - \\
&\quad \left(\sum_{1 \le i \ne j \le n} \mathrm{E}\left[R_i\right]\mathrm{E}\left[R_j\right] + \sum_{i=1}^{n} \mathrm{E}^2\left[R_i\right]\right) && \text{(by (7) and (8))} \\
&= \sum_{i=1}^{n} \mathrm{E}\left[R_i^2\right] - \sum_{i=1}^{n} \mathrm{E}^2\left[R_i\right] \\
&= \sum_{i=1}^{n} (\mathrm{E}\left[R_i^2\right] - \mathrm{E}^2\left[R_i\right]) && \text{(reordering the sums)} \\
&= \sum_{i=1}^{n} \mathrm{Var}\left[R_i\right]. && \text{(Theorem 8.1)}
\end{aligned}$$

□

Now we have a simple way of computing the expectation of a variable $H_{n,p}$ which has a binomial distribution with parameters n and p. We know that $H_{n,p} = \sum_{k=1}^{n} I_k$ 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 (3), so by linearity of variance, we have

Lemma (Variance of the Binomial Distribution).

$$\text{Var}\left[H_{n,p}\right] = n\,\text{Var}\left[I_k\right] = np(1-p). \tag{9}$$

9 Estimation by Random Sampling

9.1 Estimating Voting Preferences using Chebyshev's Theorem

In Notes 13, we used bounds on the binomial distribution to determine confidence levels for a poll of voter preferences of Clinton vs. Giulliani. 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 13: 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 (9) we have

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

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:

$$\begin{aligned}
\text{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \text{Var}\left[S_n\right] && \text{(by (4))} \\
&\leq \left(\frac{1}{n}\right)^2 n\frac{1}{4} && \text{(by (9.1))} \\
&= \frac{1}{4n}. && (10)
\end{aligned}$$

Now from Chebyshev and (10) we have:

$$\Pr\left\{\left|\frac{S_n}{n} - p\right| \geq 0.04\right\} \leq \frac{\text{Var}\left[S_n/n\right]}{(0.04)^2} = \frac{1}{4n(0.04)^2} = \frac{156.25}{n}. \tag{11}$$

To make our our estimate with with 95% confidence, we want the righthand side of (11) to be at most $1/20$. So we choose n so that

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

that is,

$$n \geq 3,125.$$

You may remember that in Notes 13 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 13, 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.

9.2 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[2] 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 M_n be the number of matching pairs of birthdays among the n choices, that is,

$$M_n ::= \sum_{1 \le i < j \le n} E_{i,j}. \tag{12}$$

So by linearity of expectation

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

Also, by Theorem 8.5, the variances of pairwise independent variables are additive, so

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

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

$$\Pr\{|M_{100} - 14| \ge x\} < \frac{14}{x^2}.$$

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.

10 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*

$$S_n ::= \sum_{i=1}^{n} G_i. \tag{13}$$

[2]In the U.S., Fall birthdays are more common than Winter birthdays, so $\Pr\{B_i = B_j\}$ is actually a bit larger than $1/365$.

Then

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

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

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

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

$$\begin{aligned}
\mathrm{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \mathrm{Var}\left[S_n\right] && \text{(by (4))} \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^n G_i\right] && \text{(def. of } S_n\text{)} \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\left[G_i\right] && \text{(pairwise independent additivity)} \\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. && (14)
\end{aligned}$$

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

$$\begin{aligned}
\Pr\left\{\left|\frac{S_n}{n} - \mu\right| \geq x\right\} &\leq \frac{\mathrm{Var}\left[S_n/n\right]}{x^2}. && \text{(Chebyshev's bound)} \\
&= \frac{\sigma^2/n}{x^2} && \text{(by (14))} \\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

□

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 shows that by choosing a large enough sample size, n, we can get arbitrarily accurate estimates of the mean with confidence arbitrarily close to 100%.