3.3 Some properties of the expected value

If X is a discrete random variable with Range$(X) = \{x_1, x_2, x_3, ...\}$, then its expected value is the weighted average:

$$E(X) = \sum_{i=1}^{\infty} x_i P(X = x_i).$$

In this section we itemize some of the important properties of the expected value

3.3.1 The mean may not exist or be finite

To begin, we make the important point that the expected value may not be finite or even exist. To see why not, consider the case of a discrete random variable X with Range$(X) = \mathbb{N} = \{0, 1, 2, ...\}$. In this case, the generic formula for the expected value is

$$E(X) = \sum_{n=0}^{\infty} nP(X = n).$$

As $n \to \infty$, then n of course blows up. We know that $\sum_{n=0}^{\infty} P(X = n) = 1$, so we must have that $P(X = n) \to 0$ as $n \to \infty$. So the $P(X = n)$ part decays to zero at the same time as the n part blows up. If the probabilities $P(X = n)$ do not go to zero *fast enough*, then they won't reign in the growth of the n part, and the entire thing will blow up. Observe:

Example 3.6. Let X be a random variable with Range$(X) = \{1, 2, 3, ...\}$ and

$$P(X = n) = \frac{6}{\pi^2 n^2}, \quad n = 1, 2, ...$$

We see that this is a valid pmf:

$$\begin{aligned}\sum_{n=1}^{\infty} P(X = n) &= \sum_{n=1}^{\infty} \frac{6}{\pi^2 n^2} \\ &= \frac{6}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} && \text{(well known p-series)} \\ &= \frac{6}{\pi^2} \frac{\pi^2}{6} \\ &= 1.\end{aligned}$$

So the probabilities go to zero as n grows, but not fast enough:

$$\begin{aligned}E(X) &= \sum_{n=1}^{\infty} nP(X = n) \\ &= \sum_{n=1}^{\infty} \cancel{n} \frac{6}{\pi^2 n^{\cancel{2}}} \\ &= \frac{6}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n} && \text{(harmonic series)} \\ &= \infty.\end{aligned}$$

So the expected value is infinite. Gross!

Example 3.7. Let X be a random variable with $\text{Range}(X) = \{1, 2, 3, ...\}$ and

$$P(X = n) = \frac{1}{n(n+1)}, \quad n = 1, 2, ...$$

You can verify that this is a valid pmf, but then

$$\begin{aligned} E(X) &= \sum_{n=1}^{\infty} nP(X = n) \\ &= \sum_{n=1}^{\infty} \frac{n}{n(n+1)} \\ &= \sum_{n=1}^{\infty} \frac{1}{n+1} \\ &= \infty. \end{aligned}$$

Example 3.8. Now we consider an example where, instead of the expected value existing and being infinite, we say that it simply does not exist. Let X be a random variable with

$$\text{Range}(X) = \{..., -16, -8, -4, 4, 8, 16, ...\} = \{-2^k : k = 2, 3, ...\} \cup \{2^k : k = 2, 3, ...\}$$

and pmf

$$P\left(X = \pm 2^k\right) = \frac{1}{2^k}, \quad k = 2, 3, ...$$

Is this a valid pmf? Yes. First, observe that

$$\sum_{x \in \text{Range}(X)} P(X = x) = \sum_{k=2}^{\infty} P(X = 2^k) + \sum_{k=2}^{\infty} P(X = -2^k) = \sum_{k=2}^{\infty} \frac{1}{2^k} + \sum_{k=2}^{\infty} \frac{1}{2^k}.$$

If (and only if!) these two terms are finite, then we can add them together. So we have to check the finiteness first:

$$\begin{aligned} \sum_{k=2}^{\infty} \frac{1}{2^k} &= \sum_{k=2}^{\infty} \left(\frac{1}{2}\right)^k \\ &= \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^{i+2} && \text{reindex} \\ &= \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^i \\ &= \sum_{i=0}^{\infty} \frac{1}{4} \left(\frac{1}{2}\right)^i \\ &= \frac{1/4}{1 - \frac{1}{2}} && \text{geometric series} \\ &= \frac{1/4}{1/2} \\ &= 1/2. \end{aligned}$$

Great! Now we can add:

$$\sum_{x\in\text{Range}(X)} P(X=x) = \sum_{k=2}^{\infty} P(X=2^k) + \sum_{k=2}^{\infty} P(X=-2^k) = \sum_{k=2}^{\infty}\frac{1}{2^k} + \sum_{k=2}^{\infty}\frac{1}{2^k} = \frac{1}{2}+\frac{1}{2} = 1.$$

Even so, the expected value is undefined:

$$\begin{aligned}\text{“}E(X)\text{”} = \sum_{x\in\text{Range}(X)} x\cdot P(X=x) &= \sum_{k=2}^{\infty} -2^k P(X=-2^k) + \sum_{k=2}^{\infty} 2^k P(X=2^k)\\ &= \sum_{k=2}^{\infty} -2^k\frac{1}{2^k} + \sum_{k=2}^{\infty} 2^k\frac{1}{2^k}\\ &= \sum_{k=2}^{\infty}(-1) + \sum_{k=2}^{\infty} 1\\ &= -\infty+\infty.\end{aligned}$$

Having reached an $\infty - \infty$ indeterminate form, we conclude that the expected value simply does not exist.

3.3.2 The expected value of a transformation

A random variable X is a real number, so you can apply transformations to it: X^2, $\ln(X)$, $aX+b$, etc. Because X is random, these transformations will also be random variables with their own distributions. Later on, we will learn how to compute the entire distribution of a transformation of a random variable. But for now, we restrict ourselves to a simpler task: given some function g, how do you compute $E[g(X)]$? Here's an example of doing it the long way:

Example 3.9. Start with a random variable X, where $\text{Range}(X)=\{-1,\,0,\,1\}$, $P(X=-1)=2/10$, $P(X=0)=5/10$, and $P(X=1)=3/10$. Next, define a new random variable $Y=g(X)=X^2$. This new random variable has $\text{Range}(Y)=\{0,\,1\}$, with $P(Y=0)=P(X=0)=1/2$ and

$$P(Y=1)=P(X\in\{-1,\,1\})=P(X=-1)+P(X=1)=\frac{2}{10}+\frac{3}{10}=\frac{1}{2}.$$

So it turns out that $Y\sim\text{Bern}(1/2)$, and so $E(Y)=E[g(X)]=E(X^2)=0.5$.

In order to compute the expected value of X^2 in this example, we first derived the entire distribution of X^2, and then applied the definition of the expected value to the transformed random variable. That was simple enough in this toy example, but in other contexts this could be hard, and if possible we would like to skip this intermediate step of deriving the whole distribution before computing one measly expected value. Theorem 3.1 provides the shortcut we crave.

Theorem 3.1. *If X is a discrete random variable and $g:\text{Range}(X)\to\mathbb{R}$ is a transformation, then*

$$E[g(X)]=\sum_{i=1}^{\infty} g(x_i)P(X=x_i). \tag{3.7}$$

The upshot of Theorem 3.1 is that it allows you to compute $E[g(X)]$ without having to first derive the range and distribution of the new random variable $g(X)$. The statement of $E[g(X)]$ only refers to Range(X) and the pmf of X. This is very convenient.

Remark 3.3. Theorem 3.1 is called the **law of the unconscious statistician** (**LOTUS**) because equation 3.7 seems too good to be true. It seems like it's a mistake – the kind that an *unconscious statistician* might carelessly make in a bout of wishful thinking. Sort of like if you weren't paying close attention and blithely wrote that $(x + y)^2$ equaled $x^2 + y^2$. But unlike this, and fortunately for us, LOTUS is as true as the sky is blue.

Example 3.10. Consider the same X from Example 3.9. Theorem 3.1 now tells us that

$$\begin{aligned} E[g(X)] = E(X^2) &= (-1)^2\frac{2}{10} + 0^2\frac{5}{10} + 1^2\frac{3}{10} \\ &= \frac{2}{10} + \frac{3}{10} \\ &= \frac{1}{2}. \end{aligned}$$

So we get the same answer as before, but without the intermediate step of deriving the entire distribution of X^2. We only had to make use of the range and pmf of the original X.

Theorem 3.2. Let $a, b \in \mathbb{R}$ be constants, and consider any random variable X for which $E(X)$ exists and is finite. Then

$$E(aX + b) = aE(X) + b. \tag{3.8}$$

In other words, "the expected value of a linear transformation is the linear transformation of the expected value."

Partial proof. Assume X is discrete with Range$(X) = \{x_1, x_2, ...\}$. The theorem does not require this, but at the moment this is all we can prove with the tools we possess. We are interested in $E[g(X)]$ for $g(x) = ax + b$, so LOTUS tells us that

$$\begin{aligned} E(aX + b) &= \sum_{i=1}^{\infty}(ax_i + b)P(X = x_i) \\ &= \sum_{i=1}^{\infty}[ax_iP(X = x_i) + bP(X = x_i)] \\ &= a\underbrace{\sum_{i=1}^{\infty} x_iP(X = x_i)}_{\text{Definition 3.5}} + b\underbrace{\sum_{i=1}^{\infty} P(X = x_i)}_{1} \\ &= aE(X) + b. \end{aligned}$$

□

Theorem 3.3. If $a_1, a_2, ..., a_n \in \mathbb{R}$ are constants and $X_1, X_2, ..., X_n$ are random variables for which $E(X_i)$ all exist and are finite, then

$$E\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_iE(X_i). \tag{3.9}$$

In other words, "the expected value of a linear combination is the linear combination of the expected values."

Once we have more tools, we will see a partial proof of this result in the case where $n = 2$. The general case then follows by induction.

Remark 3.4. Theorem 3.3 establishes that the expected value is a *linear operator*. That vocabulary may be foreboding, but you've encountered many linear operators in your life. The derivative is a linear operator:

$$\frac{\mathrm{d}}{\mathrm{d}x}[a \cdot f(x) + b \cdot g(x)] = a\frac{\mathrm{d}}{\mathrm{d}x}f(x) + b\frac{\mathrm{d}}{\mathrm{d}x}g(x).$$

The integral is a linear operator:

$$\int_0^1 [a \cdot f(x) + b \cdot g(x)]\mathrm{d}x = a\int_0^1 f(x)\,\mathrm{d}x + b\int_0^1 g(x)\,\mathrm{d}x.$$

We just mean it distributes over sums and you can pull constants out.

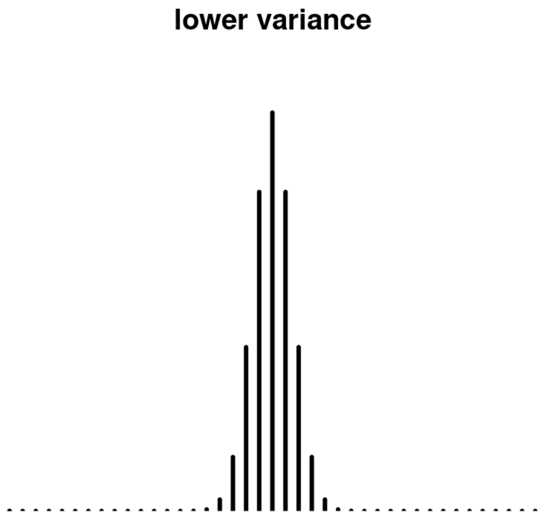
Remark 3.5. Theorem 3.3 is very general. It applies to any collection of random variables, regardless their type (discrete, continuous or neither) or their dependence structure. As long as each has finite mean, the theorem holds.

Remark 3.6. Does it work for infinite sums? Can we take $n = \infty$ and say $E(\sum_{i=1}^{\infty} a_i X_i) = \sum_{i=1}^{\infty} a_i E(X_i)$. Not always. And the same goes for the derivative and integral. They do not automatically pass through infinite sums either.

Example 3.11. If $X \sim \text{Binom}(n,\ p)$, we saw that $E(X) = np$. But the calculation was ugly. We can use linearity to clean it up. Recalling Example 3.3, a binomial random variable can be expressed as a sum of Bernoulli indicators. That is, for $i = 1,\ 2,\ ...,\ n$, let $I_i \sim \text{Bern}(p)$. Furthermore, assume these are all independent. Since the I_i all have the same p, we can summarize this situation by writing $I_i \overset{\text{iid}}{\sim} \text{Bern}(p)$, where iid stands for **independent and identically distributed**. Each I_i represents a 0/1 trial, they are all independent, and they have the same probability of success. As such, a new random variable X that counts the number of successes in these n trials is equivalent to $X = \sum_{i=1}^{n} I_i \sim \text{Binom}(n,\ p)$. Recalling that the expectation of a Bernoulli is $E(I_i) = p$, we have by the linearity of expectation that

$$E(X) = E\left(\sum_{i=1}^{n} I_i\right) = \sum_{i=1}^{n} E(I_i) = \sum_{i=1}^{n} p = np.$$

Much easier! Granted, we have yet to prove that the expected value is actually a linear operator. If doing so were a prerequisite for performing this particular calculation, perhaps you would not regard it as particularly easy. But taking linearity for granted, this is a far more elegant solution than when we proceeding from the definition.



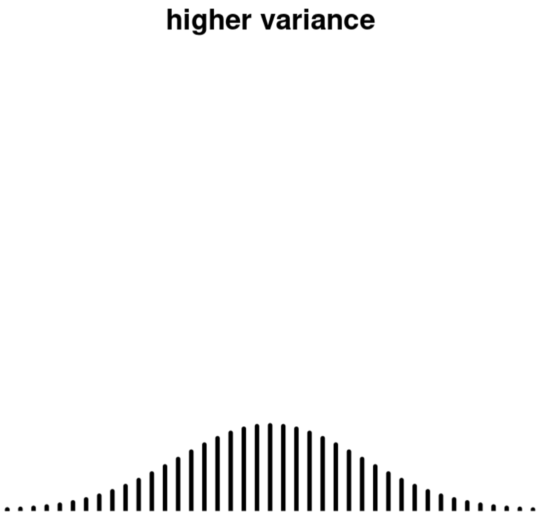


Figure 3.10

3.3.3 Variance

The expected value $E(X)$ of a random variable is a convenient single number summary that describes the "average" or "typical" behavior of a random variable. It captures the location or central tendency of the distribution. Here is another summary that describes the spread of the distribution:

Definition 3.11. Let X be *any* random variable with finite mean. Then the **variance** of X is

$$\text{var}(X) = E\left[(X - E(X))^2\right]. \tag{3.10}$$

The **standard deviation** of X is the square root of the variance: $\text{sd}(X) = \sqrt{\text{var}(X)}$.

Recalling LOTUS, we see that the variance is simply the expected value of a particular transformation of X: $g(X) = (X - E(X))^2$. This transformation g is the squared distance between X and its mean, and so the variance answers the question "how far away is X from its mean, on average?" If the answer is "not that far," then the distribution of X is not that spread out. If the answer is "pretty far," then the distribution is more spread out. Figure 3.10 displays a cartoon of this. So the variance is a single number ranging from 0 to ∞ that summarizes how variable or surprising X is. If X has low variance near 0, you don't expect to be surprised by its realizations. It's basically giving you results close to the mean. If X has high variance, you expect to regularly be surprised by what it delivers. $E(X)$ may be the typical value, but values quite far away from $E(X)$ remain fair game.

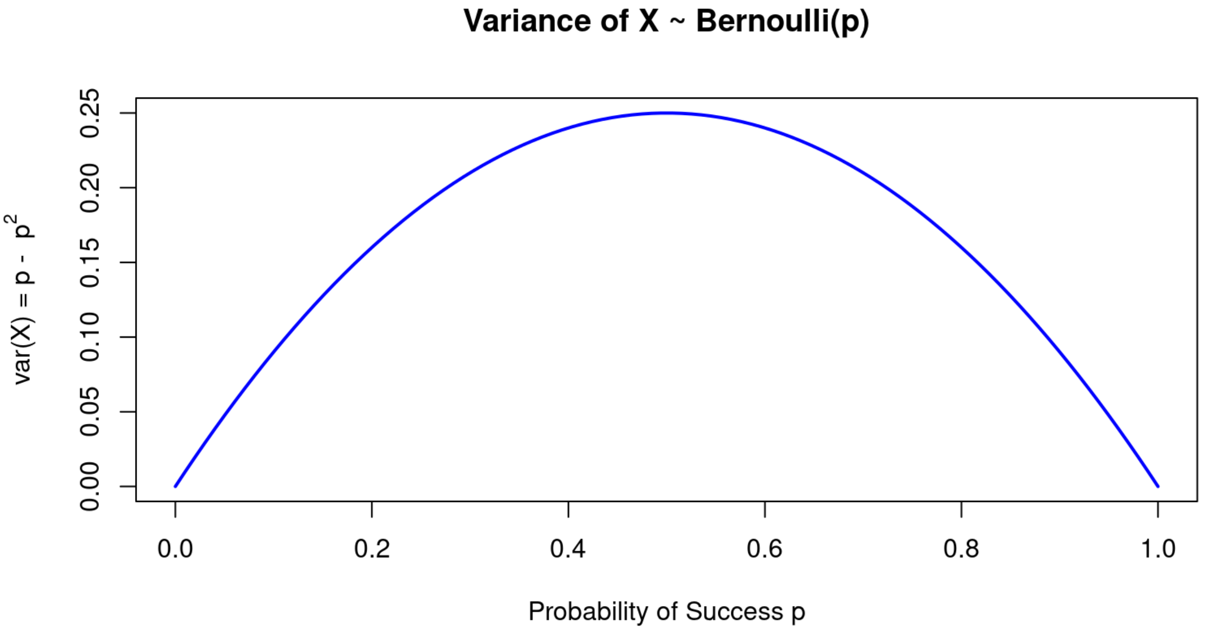


Figure 3.11

Example 3.12. Consider $X \sim \text{Bern}(p)$. We know $E(X) = p$, so by LOTUS, we know that

$$\begin{aligned}
\text{var}(X) &= E[(X - E(X))^2] \\
&= E[(X - p)^2] \\
&= \sum_{k=0}^{1}(k - p)^2 P(X = k) \\
&= (0 - p)^2(1 - p) + (1 - p)^2 p \\
&= p^2(1 - p) + (1 - p)^2 p \\
&= p(1 - p)(p + 1 - p) \\
&= p(1 - p) \\
&= p - p^2.
\end{aligned}$$

Figure 3.11 plots the variance of the Bernoulli as a function of the probability of success p. We observe a few things: if $p = 0$, meaning $X = 0$ is guaranteed, or $p = 1$, meaning $X = 1$ is guaranteed, then the variance is 0, meaning X is perfectly predictable and there are no surprises. If $p = 1/2$, then the variance is maximized. Thinking of X as a coin flip then, this implies that a fair coin is the most surprising, or least predictable, which makes sense. It could go either way with equal probability.

In the Bernoulli example, we computed the variance using LOTUS. That is the last time we will do that. The following **computation formula** is much more convenient.

Theorem 3.4. Let X be any random variance with finite mean and variance. Then

$$\text{var}(X) = E(X^2) - E(X)^2. \tag{3.11}$$

Proof. This calculation is an exercise in applying the linearity of expectation, whilst remembering that, whatever its value happens to be, $E(X)$ is itself just a constant. So you should treat it like one. Observe:

$$\begin{aligned}
\text{var}(X) &= E[(X - E(X))^2] \\
&= E\left[X^2 - 2E(X)X + E(X)^2\right] \\
&= E(X^2) - E[2E(X)X] + E[E(X)^2] \\
&= E(X^2) - 2E(X)E(X) + E(X)^2 \\
&= E(X^2) - 2E(X)^2 + E(X)^2 \\
&= E(X^2) - E(X)^2.
\end{aligned}$$

□

Remark 3.7. The quantity $E(X^2)$ is called the **second moment** of X. But there is nothing special about the number two, and in general $E(X^n)$ is called the ***n*th moment** of X.

Example 3.13. Let $X \sim \text{Bern}(p)$ again. We can use LOTUS to compute the second moment:

$$E(X^2) = \sum_{k=0}^{1} k^2 P(X = k) = 0^2(1 - p) + 1^2 p = p.$$

With this, Theorem 3.4 gives that

$$\text{var}(X) = E(X^2) - E(X)^2 = p - p^2 = p(1 - p).$$

This is the same result that we got in Example 3.12, but the computation was cleaner.

Example 3.14. The second moment of $X \sim \text{Poisson}(\lambda)$ is

$$\begin{aligned}
E(X^2) &= \sum_{k=0}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} \\
&= \sum_{k=1}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} \\
&= \lambda e^{-\lambda}\sum_{k=1}^{\infty} k^2 \frac{\lambda^{k-1}}{k!} \\
&= \lambda e^{-\lambda}\sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda}\sum_{k=1}^{\infty} (k-1+1) \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda}\sum_{k=1}^{\infty} \left[(k-1)\frac{\lambda^{k-1}}{(k-1)!} + \frac{\lambda^{k-1}}{(k-1)!}\right] \\
&= \lambda e^{-\lambda}\left[\sum_{k=1}^{\infty}(k-1)\frac{\lambda^{k-1}}{(k-1)!} + \sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}\right] \\
&= \lambda e^{-\lambda}\left[\sum_{k=2}^{\infty}(k-1)\frac{\lambda^{k-1}}{(k-1)!} + \sum_{i=0}^{\infty}\frac{\lambda^{i}}{i!}\right] \\
&= \lambda e^{-\lambda}\left[\sum_{k=2}^{\infty}\frac{\lambda^{k-1}}{(k-2)!} + e^{\lambda}\right] \\
&= \lambda e^{-\lambda}\left[\lambda\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} + e^{\lambda}\right] \\
&= \lambda e^{-\lambda}\left[\lambda\sum_{j=0}^{\infty}\frac{\lambda^{j}}{j!} + e^{\lambda}\right] \\
&= \lambda e^{-\lambda}\left[\lambda e^{\lambda} + e^{\lambda}\right] \\
&= \lambda^2 e^{-\lambda}e^{\lambda} + \lambda e^{-\lambda}e^{\lambda} \\
&= \lambda^2 + \lambda.
\end{aligned}$$

So $\text{var}(X) = E(X^2) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$, and strangely, the Poisson distribution has $E(X) = \text{var}(X)$. We saw this in Figure 3.8, where the pmf shifted rightward *and widened* as λ increased. This property renders the Poisson distribution less useful for modeling count-valued data than we might hope, because we cannot separately manipulate the center and the spread of the distribution. An entire literature has emerged in statistics that attempts to modify the Poisson so that it does not have this funky property.

When studying the expected value, we wondered how we might compute the expectation of a transformation $E[g(X)]$, and we came up with LOTUS. Is there a convenient, LOTUS-like formula for var$[g(X)]$? In general, no. But an important special case is the linear transformation $g(X) = aX + b$. We know that the expected value is a linear operator, and so $E(aX + b) = aE(X) + b$. There is a clean formula for var$(aX + b)$, but it's *not* linear:

Theorem 3.5. Let X be any random variance with finite mean and variance, and let $a, b \in \mathbb{R}$ be arbitrary constants. Then

$$\text{var}(aX + b) = a^2\text{var}(X). \tag{3.12}$$

Proof.

$$\begin{aligned}
\text{var}(aX + b) &= E[(aX + b)^2] - E(aX + b)^2 \\
&= E(a^2X^2 + 2abX + b^2) - [aE(X) + b]^2 \\
&= E(a^2X^2 + 2abX + b^2) - [a^2E(X)^2 + 2abE(X) + b^2] \\
&= a^2E(X^2) + 2abE(X) + b^2 - a^2E(X^2) - 2abE(X) - b^2 \\
&= a^2E(X^2) - a^2E(X)^2 \\
&= a^2[E(X^2) - E(X)^2] \\
&= a^2\text{var}(X).
\end{aligned}$$

□

Remark 3.8. As we see, it is not the case that var$(aX + b)$ is equal to $a$var$(X) + b$, and so the variance *is not* a linear operator. But this makes sense. The variance is a measure of spread, and merely shifting the location of a random variable with $X + b$, for instance, should not have an effect on how spread out it is. The spread remains the same.

Theorem 3.6. Let $X_1, X_2, ..., X_n$ be *independent* random variables each with finite mean and variance (possibly all different), and let $a_1, a_2, ..., a_n \in \mathbb{R}$ be arbitrary constants. Then

$$\text{var}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\text{var}(X_i). \tag{3.13}$$

Example 3.15. If $I_i \overset{\text{iid}}{\sim} \text{Bern}(p)$, then $X = \sum_{i=1}^{n} I_i \sim \text{Binom}(n, p)$, and we used this fact together with the linearity of expectation to show that $E(X) = np$. We can now use Example 3.12 and Theorem 3.6 to perform a quick derivation of the variance of a binomial random variable:

$$\text{var}(X) = \text{var}\left(\sum_{i=1}^{n} I_i\right) = \sum_{i=1}^{n} \text{var}(I_i) = \sum_{i=1}^{n} p(1-p) = np(1-p).$$