5 Variance, Higher Moments, and Random Sums

In Chapter 4 we devoted a lot of time to computing the expectation of random variables. As we explained, the expectation is useful because it provides us with a single summary value when trading off different options. For example, in Example 4.1, we used the "expected earnings" in choosing between two startups.

However one might want more information than just the expected earnings. For example, two companies, say Microsoft and Startup X could both have expected earnings of 100K, but at Microsoft your earnings are unlikely to deviate much from 100K, whereas at Startup X your earnings could range from 0 to 1M. Although both companies offer the same expected earnings, Startup X feels "riskier." The purpose of this chapter is to formalize what we mean by "risk" or "variability." Before we start, it will be useful to go over the definition of moments.

5.1 Higher Moments

Definition 5.1 *For a random variable, X, we say that the* **kth moment** *of X is $\mathbf{E}\left[X^k\right]$. Observe that $\mathbf{E}[X]$ is the first moment of X.*

Example 5.2 (Second moment of Geometric)

Let $X \sim \text{Geometric}(p)$. What is $\mathbf{E}\left[X^2\right]$?

Formally,

$$\mathbf{E}\left[X^2\right] = \sum_{i=1}^{\infty} i^2 p_X(i) = \sum_{i=1}^{\infty} i^2 (1-p)^{i-1} p.$$

It is not obvious how to compute this sum.

Question: Can we use Theorem 4.8 to express $\mathbf{E}\left[X^2\right] = \mathbf{E}[X] \cdot \mathbf{E}[X]$?

Answer: No, because X is certainly not independent of X.

Fortunately, there is something we can do: Since $\mathbf{E}\left[X^2\right]$ is an expectation, we can compute it via conditioning. We will condition on the value of the first flip.

$$\begin{aligned}
\mathbf{E}\left[X^2\right] &= \mathbf{E}\left[X^2 \mid \text{1st flip is head}\right] \cdot p + \mathbf{E}\left[X^2 \mid \text{1st flip is tail}\right] \cdot (1-p) \\
&= 1^2 \cdot p + \mathbf{E}\left[(1+X)^2\right] \cdot (1-p) \\
&= p + \mathbf{E}\left[1 + 2X + X^2\right] \cdot (1-p) \\
&= p + \left(1 + 2\mathbf{E}\left[X\right] + \mathbf{E}\left[X^2\right]\right)(1-p) \\
&= p + (1-p) + 2(1-p) \cdot \frac{1}{p} + \mathbf{E}\left[X^2\right](1-p) \\
p\mathbf{E}\left[X^2\right] &= 1 + 2 \cdot \frac{1-p}{p} \\
\mathbf{E}\left[X^2\right] &= \frac{2-p}{p^2}.
\end{aligned}$$

Question: In the above, observe that we write:

$$\mathbf{E}\left[X^2 \mid \text{1st flip is tail}\right] = \mathbf{E}\left[(1+X)^2\right].$$

Why didn't we write $\mathbf{E}\left[X^2 \mid \text{1st flip is tail}\right] = 1 + \mathbf{E}\left[X^2\right]$?

Answer: Consider the random variable (r.v.) Y, where

$$Y = [X \mid \text{1st flip is tail}] = [X \mid X > 1].$$

That is, we define Y to be the r.v. X conditioned on the fact that we know that $X > 1$. Given that the first flip is a tail, everything starts from scratch, that is, we've wasted one flip, and we get a new draw of X. It thus makes sense that

$$Y \stackrel{d}{=} 1 + X, \tag{5.1}$$

that is, Y is equal in distribution to $1 + X$. In Exercise 5.18, we formally prove (5.1) by showing that $\mathbf{P}\{Y = i\} = \mathbf{P}\{1 + X = i\}$ for all i.

From (5.1), it follows that:

$$\begin{aligned}
\mathbf{E}\left[X \mid \text{1st flip is tail}\right] &= \mathbf{E}\left[Y\right] = \mathbf{E}\left[(1+X)\right] = 1 + \mathbf{E}\left[X\right] \\
\mathbf{E}\left[X^2 \mid \text{1st flip is tail}\right] &= \mathbf{E}\left[Y^2\right] = \mathbf{E}\left[(1+X)^2\right].
\end{aligned}$$

Question: Could we use the same approach to compute the third moment of X?

Answer: Sure:

$$\mathbf{E}\left[X^3\right] = 1^3 \cdot p + \mathbf{E}\left[(1+X)^3\right] \cdot (1-p).$$

Now expand out the cube and then again apply Linearity of Expectation.

5.2 Variance

We are often interested in how much an experiment is likely to deviate from its mean.

Definition 5.3 *The* **variance** *of r.v. X, written as* $\mathbf{Var}(X)$*, is the expected squared difference of X from its mean.*

$$\mathbf{Var}(X) = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right].$$

A depiction of variance is given in Figure 5.1.

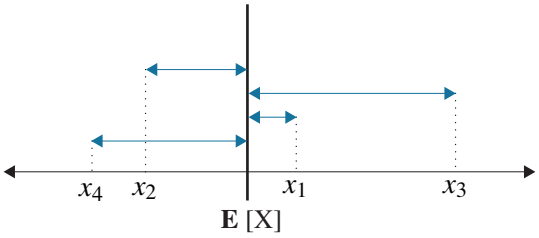


Figure 5.1 *Variance of X. For each value of X, we square its distance to $\mathbf{E}[X]$, and take the appropriate weighted average of these.*

Example 5.4 (Variance of Bernoulli)

Let $X \sim \text{Bernoulli}(p)$. Our goal is to determine $\mathbf{Var}(X)$. Here X represents a single flip of a coin, where the coin has probability p of heads. That is:

$$X = \begin{cases} 1 & \text{w/prob } p \\ 0 & \text{w/prob } 1-p \end{cases}.$$

Question: What is $\mathbf{E}[X]$? What is $\mathbf{Var}(X)$?

Answer:

$$\mathbf{E}[X] = p \cdot 1 + (1-p) \cdot 0 = p$$

$$\begin{aligned} \mathbf{Var}(X) &= \mathbf{E}\left[(X-p)^2\right] \\ &= \mathbf{E}\left[X^2 - 2Xp + p^2\right] \\ &= \mathbf{E}\left[X^2\right] - 2p\mathbf{E}[X] + p^2 \\ &= \left(p \cdot 1^2 + (1-p) \cdot 0^2\right) - 2p \cdot p + p^2 \\ &= p(1-p). \end{aligned} \tag{5.2}$$

Formula (5.2) is worth memorizing.

Question: Can we compute $\mathbf{Var}(X)$ via conditioning?

Answer: There is a right and a wrong way to do this.

The WRONG way is to say:

$$\begin{aligned}\mathbf{Var}(X) &= p \cdot \mathbf{Var}(X \mid X = 1) + (1-p) \cdot \mathbf{Var}(X \mid X = 0) \\ &= p \cdot 0 + (1-p) \cdot 0 = 0.\end{aligned}$$

This is incorrect, because no theorem says that we can condition on variance. We only have Theorem 4.24, which allows us to condition on expectation.

That said, if we can leave variance in the form of an expectation, then we can use conditioning. Here's the CORRECT way to condition:

$$\begin{aligned}\mathbf{Var}(X) &= \mathbf{E}\left[(X-p)^2\right] \\ &= \mathbf{E}\left[(X-p)^2 \mid X = 1\right] \cdot p + \mathbf{E}\left[(X-p)^2 \mid X = 0\right] \cdot (1-p) \\ &= (1-p)^2 \cdot p + p^2 \cdot (1-p) \\ &= p(1-p).\end{aligned}$$

Question: For any r.v. X, how does $\mathbf{Var}(-X)$ compare to $\mathbf{Var}(X)$?

Answer: Looking at Figure 5.1, we see that every value of X is now negated, including the mean of X. Thus the distance of each value to the mean doesn't change. Hence the sum of the squares of the distances doesn't change either. So $\mathbf{Var}(X) = \mathbf{Var}(-X)$.

5.3 Alternative Definitions of Variance

Question: How else might you want to define variance?

Answer: There are many answers possible. One thing that is bothersome about the existing definition is the squaring, since the units of $\mathbf{Var}(X)$ are then different from the units of X. One might instead choose to define $\mathbf{Var}(X)$ as

$$\mathbf{E}\left[X - \mathbf{E}\left[X\right]\right],$$

without the square term.

Question: What's wrong with this?

Answer: By Linearity of Expectation: $\mathbf{E}\left[X - \mathbf{E}\left[X\right]\right] = \mathbf{E}\left[X\right] - \mathbf{E}\left[X\right] = 0$. Hence this definition doesn't work.

Another possibility is to define $\mathbf{Var}(X)$ as

$$\mathbf{E}\left[\left|X - \mathbf{E}[X]\right|\right],$$

using the absolute value instead of the square.

This alternative definition is totally legitimate. The only problem is that it is missing the convenient linearity property that we'll see shortly in Theorem 5.8.

One more idea is to consider the square root of variance, which has the same units as X. This is actually so common that it has a name.

Definition 5.5 *We define the* **standard deviation of** X *as*

$$\sigma_X = \mathbf{std}(X) = \sqrt{\mathbf{Var}(X)}.$$

We often write

$$\mathbf{Var}(X) = \sigma_X^2.$$

There's something disturbing about the definition of variance: The same measurement taken in different scales will end up with different values of variance. For example, suppose that X and Y are measuring the same quantity, but X is measured in centimeters and Y is measured in millimeters. As a result, we find that:

$$X = \begin{cases} 3 & \text{w/prob } \frac{1}{3} \\ 2 & \text{w/prob } \frac{1}{3} \\ 1 & \text{w/prob } \frac{1}{3} \end{cases} \qquad Y = \begin{cases} 30 & \text{w/prob } \frac{1}{3} \\ 20 & \text{w/prob } \frac{1}{3} \\ 10 & \text{w/prob } \frac{1}{3} \end{cases}.$$

We would like to believe that X and Y have the same variance, in that they're measuring the same quantity, just in different units.

Question: How do $\mathbf{Var}(X)$ and $\mathbf{Var}(Y)$ compare?

Answer: $\mathbf{Var}(X) = \frac{2}{3}$, while $\mathbf{Var}(Y) = \frac{200}{3}$. Since units are not typically shown, we are left with very different values.

The problem is not fixed by switching to the standard deviation.

Question: How do $\mathbf{std}(X)$ and $\mathbf{std}(Y)$ compare?

Answer: $\mathbf{std}(X) = \sqrt{\frac{2}{3}}$, while $\mathbf{std}(Y) = \sqrt{\frac{200}{3}}$.

Again, this feels less than satisfactory. For these reasons, researchers use a normalized version of variance, which is *scale-invariant* (insensitive to scaling), called the squared coefficient of variation.

Definition 5.6 *The* **squared coefficient of variation** *of r.v. X is defined as*

$$C_X^2 = \frac{\mathbf{Var}(X)}{\mathbf{E}[X]^2}.$$

Question: How do C_X^2 and C_Y^2 compare?

Answer: Both are $\frac{1}{6}$.

Note that C_X^2 is not defined if $\mathbf{E}[X] = 0$. In practice, the C_X^2 metric is used when modeling empirical quantities like job sizes, flow durations, memory consumption, etc., whose values are typically positive with positive means.

5.4 Properties of Variance

Lemma 5.7 provides another way of computing variance.

Lemma 5.7 (Equivalent definition of variance) *The variance of r.v. X can equivalently be expressed as follows:*

$$\mathbf{Var}(X) = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2.$$

Proof:

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

■

Even with this easier formulation, variance is often hard to compute. Fortunately, the Linearity of Variance Theorem helps us break down the variance of a random variable into easier subproblems.

Theorem 5.8 (Linearity of Variance) *Let X and Y be random variables where $X \perp Y$. Then,*

$$\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y).$$

This generalizes to show that if $X_1, X_2, \ldots, X_n$ are independent, then

$$\mathbf{Var}(X_1 + X_2 + \cdots + X_n) = \mathbf{Var}(X_1) + \mathbf{Var}(X_2) + \cdots + \mathbf{Var}(X_n).$$

Proof: We prove the first statement of the theorem:

$$\begin{aligned}\mathbf{Var}(X+Y) &= \mathbf{E}\left[(X+Y)^2\right] - (\mathbf{E}\left[(X+Y)\right])^2 \\ &= \mathbf{E}\left[X^2\right] + \mathbf{E}\left[Y^2\right] + 2\mathbf{E}\left[XY\right] - (\mathbf{E}\left[X\right])^2 - (\mathbf{E}\left[Y\right])^2 - 2\mathbf{E}\left[X\right]\mathbf{E}\left[Y\right] \\ &= \mathbf{Var}(X) + \mathbf{Var}(Y) + \underbrace{2\mathbf{E}\left[XY\right] - 2\mathbf{E}\left[X\right]\mathbf{E}\left[Y\right]}_{\text{equals 0 if X} \perp \text{ Y}}.\end{aligned}$$

■

Theorem 5.8 is hugely powerful, assuming that X and Y are independent. One of the key reasons for the chosen definition of variance (as opposed to using absolute values, for example) is that the chosen definition lends itself to this linearity property.

It turns out that Theorem 5.8 can be extended to the case where the X_i's are not independent but rather only *pairwise independent*, which means that each *pair* of variables X_i and X_j are independent. This generalization is proven in Exercise 5.38.

We now present some examples of the benefits of Linearity of Variance.

Example 5.9 (Second moment of Binomial)

Let $X \sim \text{Binomial}(n, p)$. Our goal is to derive $\mathbf{E}\left[X^2\right]$.

If we work directly from the definition of the second moment, we have:

$$\mathbf{E}\left[X^2\right] = \sum_{i=0}^{n} i^2 \binom{n}{i} p^i (1-p)^{n-i}.$$

This is not an easy sum to work with. On the other hand, we can write X as a sum of indicator random variables, as we've done in the past:

$$X = \text{number of successes in } n \text{ trials} = X_1 + X_2 + \cdots + X_n,$$

where

$$X_i \sim \text{Bernoulli}(p) \qquad \text{and} \qquad \mathbf{E}\left[X_i\right] = p.$$

Then

$$\begin{aligned}\mathbf{Var}(X) &= \mathbf{Var}(X_1) + \mathbf{Var}(X_2) + \cdots + \mathbf{Var}(X_n) \\ &= n\mathbf{Var}(X_i) \\ &= np(1-p).\end{aligned}$$

Now, invoking Lemma 5.7, we have:

$$\mathbf{E}\left[X^2\right] = \mathbf{Var}(X) + \mathbf{E}\left[X\right]^2 = np(1-p) + (np)^2.$$

Question: Recall the drinks example from Section 4.2, where n people put their drinks on a table, and each picks up a random cup. Let X denote the number of people who get back their own cup. Can we use indicator random variables to derive $\mathbf{Var}(X)$?

Answer: We could define X_i to be an indicator r.v. on whether person i gets back their own drink or not. Unfortunately, these X_i's are *not* independent, so we can't apply the Linearity of Variance Theorem as we did in computing the variance of the Binomial. In Exercise 5.36 you will see that you can nonetheless deduce $\mathbf{Var}(X)$ by writing out $\mathbf{E}\left[X^2\right] = \mathbf{E}\left[(X_1 + X_2 + \cdots + X_n)^2\right]$ and reasoning about the $\mathbf{E}\left[X_i X_j\right]$ terms.

Example 5.10 (Sums versus copies)

Consider two independent and identically distributed (i.i.d.) random variables, X_1 and X_2, which are both distributed like X. Let

$$Y = X_1 + X_2 \qquad \text{and} \qquad Z = 2X.$$

Question: Do Y and Z have the same distribution?

Answer: No. Suppose, for example, that your experiment is flipping a fair coin, where heads is 1 and tails is 0. In the case of Y, you flip the coin two independent times and look at the sum. The possible values for Y are 0, 1, or 2. In the case of Z, you flip the coin one time, and return double your result. The only possible values for Z are 0 or 2.

Question: How do $\mathbf{E}\left[Y\right]$ and $\mathbf{E}\left[Z\right]$ compare?

Answer: They are the same. $\mathbf{E}\left[Y\right] = \mathbf{E}\left[Z\right] = 2\mathbf{E}\left[X\right]$. In the case of the coin experiment, $\mathbf{E}\left[Y\right] = \mathbf{E}\left[Z\right] = 2 \cdot \frac{1}{2} = 1$.

Question: How do $\mathbf{Var}(Y)$ and $\mathbf{Var}(Z)$ compare?

Answer: $\mathbf{Var}(Y) = 2\mathbf{Var}(X)$, but $\mathbf{Var}(Z) = 4\mathbf{Var}(X)$.

Question: Does it make sense that $\mathbf{Var}(Y)$ is smaller than $\mathbf{Var}(Z)$?

Answer: In the case of Y, you are adding two independent results, which tends to yield a result that is often closer to the average. By contrast, in the case of Z you are taking one result and doubling it. This yields more extreme values. The variance is higher when we see extreme values.

5.5 Summary Table for Discrete Distributions

It is worth memorizing the mean and variance of the common distributions, because they come up over and over again. Table 5.1 shows these quantities.

Distribution	p.m.f.	Mean	Variance
Bernoulli(p)	$p_X(0) = 1-p \;;\; p_X(1) = p$	p	$p(1-p)$
Binomial(n,p)	$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$	np	$np(1-p)$
Geometric(p)	$p_X(x) = (1-p)^{x-1} p, \quad x = 1, 2, 3, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$p_X(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$	λ	λ

Table 5.1 *Common discrete distributions.*

5.6 Covariance

Suppose we now have two random variables, X and Y.

Definition 5.11 *The* **covariance** *of any two random variables X and Y, denoted by* $\mathbf{Cov}(X,Y)$*, is defined by*

$$\mathbf{Cov}(X,Y) = \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\right].$$

Lemma 5.12 provides an alternative definition of covariance.

Lemma 5.12 $\mathbf{Cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y]$.

Proof:

$$\begin{aligned}
\mathbf{Cov}(X,Y) &= \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\right] \\
&= \mathbf{E}[XY] - \mathbf{E}\left[\mathbf{E}[X] \cdot Y\right] - \mathbf{E}\left[X \cdot \mathbf{E}[Y]\right] + \mathbf{E}\left[\mathbf{E}[X] \cdot \mathbf{E}[Y]\right] \\
&= \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y] - \mathbf{E}[X]\,\mathbf{E}[Y] + \mathbf{E}[X]\,\mathbf{E}[Y] \\
&= \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y]
\end{aligned}$$

■

Intuitively, the covariance between X and Y indicates something about the joint distribution between X and Y. If the larger-than-average values of X tend to

happen with the larger-than-average values of Y, then $(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])$ is positive on average, so the $\mathbf{Cov}(X,Y)$ is positive, and we say that the random variables X and Y are **positively correlated**. If the larger-than-average values of X mainly tend to happen together with the smaller-than-average values of Y, then $(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])$ is negative on average, so the $\mathbf{Cov}(X,Y)$ is negative, and we say that the random variables X and Y are **negatively correlated**.

Thus the *sign* of $\mathbf{Cov}(X,Y)$ tells us the direction of the relationship between X and Y. Note that the magnitude of $\mathbf{Cov}(X,Y)$ is meaningless because it is too influenced by the magnitudes of X and Y.

Question: What is a nice name for $\mathbf{Cov}(X,X)$?

Answer: $\mathbf{Var}(X)$.

We will explore properties of covariance in Exercises 5.13–5.17.

5.7 Central Moments

The variance of a r.v. X is the second moment of the difference of X from its mean. In the same way, we can define higher moments of the difference of X from its mean.

Definition 5.13 *The* ***k*th moment** *of a r.v. X is*

$$\mathbf{E}\left[X^k\right] = \sum_i i^k \cdot p_X(i).$$

The ***k*th central moment** *of a r.v. X is*

$$\mathbf{E}\left[(X - \mathbf{E}[X])^k\right] = \sum_i (i - \mathbf{E}[X])^k \cdot p_X(i).$$

Question: What do we call the second central moment?

Answer: Variance.

We've discussed the intuition behind the second central moment in terms of capturing the variability of the distribution. The third central moment is related to the "skew" of the distribution, namely whether it leans right or leans left.

Question: Consider the three distributions shown in Figure 5.2. Which have positive skew? Negative skew? Zero skew?

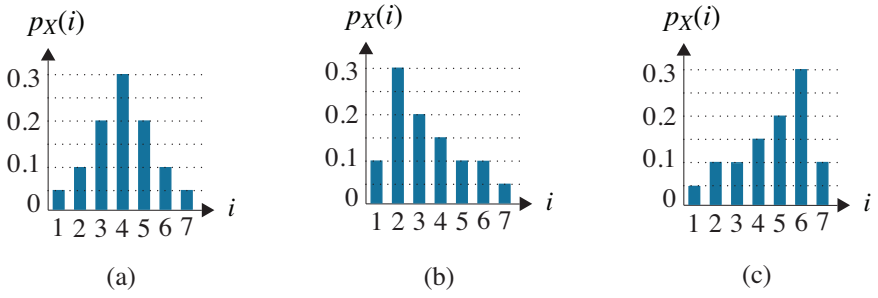


Figure 5.2 *Which of these distributions has positive/negative/zero skew?*

Answer: It is easy to see that the distribution in (a) has **zero skew**. Here, X is symmetric about its mean so $\mathbf{E}\left[(X - \mathbf{E}[X])^3\right] = 0$. The distribution in (b) has **positive skew** because it is "skewed" above its mean, so there will be more positive terms than negative ones in computing $\mathbf{E}\left[(X - \mathbf{E}[X])^3\right]$. Likewise the distribution in (c) has **negative skew** because it is "skewed" below its mean.

Question: Does having a zero third central moment guarantee that the distribution is symmetric?

Answer: No. This is why "skew" is not a perfect term. There are also plenty of distributions that don't look skewed one way or the other.

Question: Is there intuition behind the fourth central moment?

Answer: The fourth central moment is very similar to the second central moment, except that "outliers" count a lot more, because their difference from the mean is accentuated when raised to the fourth power.

5.8 Sum of a Random Number of Random Variables

In many applications one needs to add up a number of i.i.d. random variables, where the number of these variables is itself a r.v. Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables, where $X_i \sim X$. Let S denote the sum:

$$S = \sum_{i=1}^{N} X_i, \qquad \text{where } N \perp \{X_1, X_2, \ldots\},$$

where N is not a constant, but rather a non-negative, integer-valued r.v.

Figure 5.3 shows an example where a game show contestant gets a prize each day. Here, X_i represents the prize on day i. After receiving the prize, the wheel is

spun. If the wheel lands on STOP then the game ends; otherwise the contestant is invited to come back tomorrow. The number of times that the wheel is spun is a r.v., N. In this story $N \sim \text{Geometric}\left(\frac{1}{6}\right)$. The total earnings of the contestant is $S = \sum_{i=1}^{N} X_i$. We are interested in understanding $\mathbf{E}[S]$ and $\mathbf{Var}(S)$.

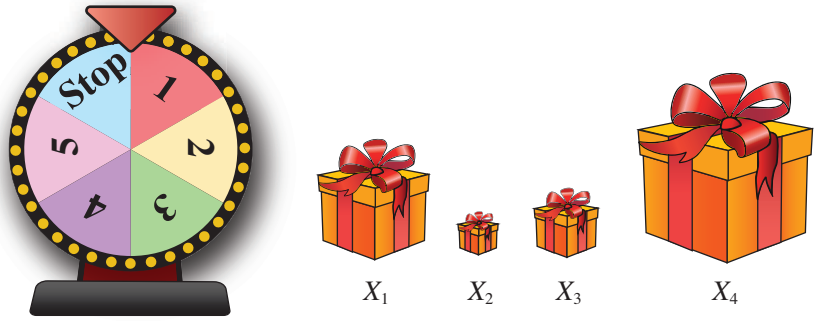


Figure 5.3 *Keep getting prizes until the wheel says STOP.*

Question: In computing $\mathbf{E}[S]$, why can't we directly apply Linearity of Expectation?

Answer: Linearity of Expectation only applies when N is a constant. But this suggests that we can condition on the value of N, and then apply Linearity of Expectation.

$$\begin{aligned}\mathbf{E}[S] = \mathbf{E}\left[\sum_{i=1}^{N} X_i\right] &= \sum_n \mathbf{E}\left[\sum_{i=1}^{N} X_i \;\middle|\; N = n\right] \cdot \mathbf{P}\{N = n\} \\ &= \sum_n \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \cdot \mathbf{P}\{N = n\} \\ &= \sum_n n\mathbf{E}[X] \cdot \mathbf{P}\{N = n\} \\ &= \mathbf{E}[X] \cdot \mathbf{E}[N]. \end{aligned} \tag{5.3}$$

Question: Let's try the same approach to get $\mathbf{Var}(S)$. What is $\mathbf{Var}(S \mid N = n)$?

Answer:

$$\mathbf{Var}(S \mid N = n) = n \cdot \mathbf{Var}(X), \qquad \text{by Linearity of Variance.}$$

Unfortunately, we there's no "Total Law of Variance" the way there's a "Total

Law of Expectation." So we cannot write:

$$\begin{aligned}\text{(WRONG)}\quad \mathbf{Var}(S) &= \sum_n \mathbf{Var}(S \mid N=n)\cdot \mathbf{P}\{N=n\}\\ &= \sum_n n\cdot \mathbf{Var}(X)\cdot \mathbf{P}\{N=n\}\\ &= \mathbf{E}[N]\cdot \mathbf{Var}(X).\end{aligned}$$

We can't use conditioning to get $\mathbf{Var}(S)$, but we can use it to get $\mathbf{E}\left[S^2\right]$:

$$\begin{aligned}\mathbf{E}\left[S^2\right] &= \sum_n \mathbf{E}\left[S^2 \mid N=n\right]\cdot \mathbf{P}\{N=n\}\\ &= \sum_n \mathbf{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right]\cdot \mathbf{P}\{N=n\}\\ &= \sum_n \mathbf{E}\left[(X_1+X_2+\cdots+X_n)^2\right]\cdot \mathbf{P}\{N=n\}\\ &= \sum_n \left(n\cdot \mathbf{E}\left[X_1^2\right]+(n^2-n)\cdot \mathbf{E}\left[X_1X_2\right]\right)\cdot \mathbf{P}\{N=n\}\\ &= \sum_n n\mathbf{E}\left[X^2\right]\mathbf{P}\{N=n\}+\sum_n (n^2-n)\mathbf{E}[X]^2\,\mathbf{P}\{N=n\}\\ &= \mathbf{E}[N]\,\mathbf{E}\left[X^2\right]+\mathbf{E}\left[N^2\right]\mathbf{E}[X]^2-\mathbf{E}[N]\,\mathbf{E}[X]^2\\ &= \mathbf{E}[N]\,\mathbf{Var}(X)+\mathbf{E}\left[N^2\right]\mathbf{E}[X]^2.\end{aligned}$$

Now,

$$\begin{aligned}\mathbf{Var}(S) &= \mathbf{E}\left[S^2\right]-\mathbf{E}[S]^2\\ &= \mathbf{E}[N]\,\mathbf{Var}(X)+\mathbf{E}\left[N^2\right]\mathbf{E}[X]^2-(\mathbf{E}[N]\,\mathbf{E}[X])^2\\ &= \mathbf{E}[N]\,\mathbf{Var}(X)+\mathbf{Var}(N)\mathbf{E}[X]^2.\end{aligned}$$

We have proven Theorem 5.14.

Theorem 5.14 *Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables, where $X_i \sim X$. Let*

$$S=\sum_{i=1}^{N} X_i, \qquad \text{where } N \perp \{X_1, X_2, \ldots\}.$$

Then,

$$\mathbf{E}[S]=\mathbf{E}[N]\,\mathbf{E}[X], \tag{5.4}$$

$$\mathbf{E}\left[S^2\right]=\mathbf{E}[N]\,\mathbf{Var}(X)+\mathbf{E}\left[N^2\right](\mathbf{E}[X])^2, \tag{5.5}$$

$$\mathbf{Var}(S)=\mathbf{E}[N]\,\mathbf{Var}(X)+\mathbf{Var}(N)\,(\mathbf{E}[X])^2. \tag{5.6}$$

While we were able to derive $\mathbf{E}\left[S^2\right]$, with some effort, you may be wondering how we would manage if we needed $\mathbf{E}\left[S^3\right]$, or some higher moment. It turns out that there's a much easier way to handle this type of analysis, by leveraging z-transforms, which we cover in Chapter 6.

Example 5.15 (Epidemic growth modeling)

A common way of modeling epidemic growth is via a tree.

Imagine that at time $t = 0$ we start with a single node (leaf). At each time step, every leaf independently either forks off two children with probability $\frac{1}{2}$, or stays inert (does nothing) with probability $\frac{1}{2}$.

We will be interested in

$$X_t = \text{Total number of leaves in the tree after } t \text{ steps.}$$

Specifically, what is $\mathbf{E}[X_t]$ and $\mathbf{Var}(X_t)$?

Figure 5.4 provides one example of how our tree might grow.

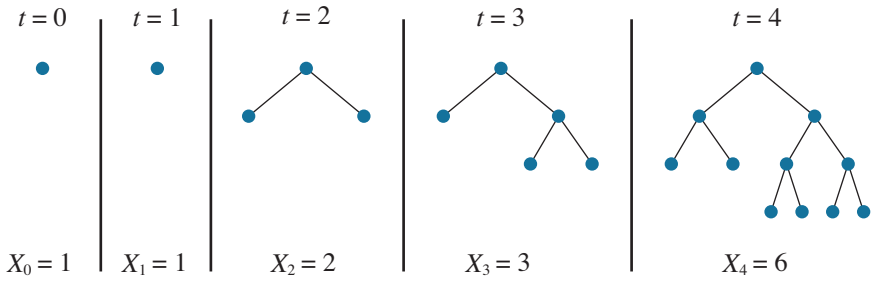


Figure 5.4 *Example of tree growth.*

Question: How can we model X_t?

It is tempting to try to write $X_t = C \cdot X_{t-1}$ for some C.

Question: Certainly the number of leaves at time t are related to the number of leaves at time $t-1$, so how can we relate X_t to X_{t-1}?

Hint: Think of X_t as a sum of a random number of random variables.

Answer: The key insight is that each of the X_{t-1} leaves contributes either 1 or 2 to X_t. Specifically, if the leaf is inert in the current round, then it contributes 1 to the next round. If the leaf forks children in the current round, then it contributes

2 to the next round. Thus we can write:

$$X_t = \sum_{i=1}^{X_{t-1}} Y_i,$$

where

$$Y_i \sim Y = \begin{cases} 1 & \text{w/prob } 0.5 \\ 2 & \text{w/prob } 0.5 \end{cases}$$

and where $X_0 = 1$.

Question: Do the conditions of Theorem 5.14 apply?

Answer: Yes, the Y_i's are all i.i.d. and are independent of X_{t-1}.

Observe that

$$\mathbf{E}[Y] = \frac{3}{2} \qquad \text{and} \qquad \mathbf{Var}(Y) = \frac{1}{4}.$$

Question: Applying Theorem 5.14, what are $\mathbf{E}[X_t]$ and $\mathbf{Var}(X_t)$?

Answer:

$$\mathbf{E}[X_t] = \mathbf{E}[X_{t-1}] \cdot \mathbf{E}[Y] = \mathbf{E}[X_{t-1}] \cdot \frac{3}{2}.$$

Therefore,

$$\mathbf{E}[X_t] = \mathbf{E}[X_0] \cdot \left(\frac{3}{2}\right)^t = \left(\frac{3}{2}\right)^t.$$

$$\begin{aligned} \mathbf{Var}(X_t) &= \mathbf{E}[X_{t-1}] \cdot \mathbf{Var}(Y) + \mathbf{Var}(X_{t-1}) \cdot \mathbf{E}[Y]^2 \\ &= \left(\frac{3}{2}\right)^{t-1} \cdot \frac{1}{4} + \mathbf{Var}(X_{t-1}) \cdot \frac{9}{4}. \end{aligned}$$

This recursion simplifies to:

$$\mathbf{Var}(X_t) = \left(\frac{9}{4}\right)^t \cdot \frac{1}{3}\left(1 - \left(\frac{2}{3}\right)^t\right).$$

5.9 Tails

The mean, the variance, and higher moments are all ways of summarizing a distribution. For a discrete r.v., X, when we refer to the **distribution associated**

with X, we are typically talking about either the p.m.f. of X, namely, $p_X(i) = \mathbf{P}\{X = i\}$ or the cumulative distribution function (c.d.f.) of X, namely, $F_X(i) = \mathbf{P}\{X \leq i\}$.

It is also common to talk about the **tail** of X, which is defined as

$$\overline{F}_X(i) = \mathbf{P}\{X > i\} = 1 - F_X(i).$$

The tail comes up in quality-of-service guarantees for computer systems and in capacity provisioning. Consider, for example, a router buffer that is designed to hold no more than 10,000 packets. We might be interested in the probability that the number of packets exceeds 10,000 and thus no longer fits within the buffer.

5.9.1 Simple Tail Bounds

A **tail bound** provides an upper bound on the tail of a distribution. We will spend considerable time on motivating and developing tail bounds in Chapter 18, but for now we only state the two simplest tail bounds. The first, Markov's inequality, relies only on the mean of the distribution, but requires the assumption that the distribution only takes on non-negative values.

Theorem 5.16 (Markov's inequality) *Let X be a non-negative r.v., with finite mean $\mu = \mathbf{E}[X]$. Then, $\forall a > 0$,*

$$\mathbf{P}\{X \geq a\} \leq \frac{\mu}{a}.$$

Proof:

$$\begin{aligned}
\mu &= \sum_{x=0}^{\infty} x p_X(x) \\
&\geq \sum_{x=a}^{\infty} x p_X(x) \\
&\geq \sum_{x=a}^{\infty} a p_X(x) \\
&= a \sum_{x=a}^{\infty} p_X(x) \\
&= a\mathbf{P}\{X \geq a\}.
\end{aligned}$$

■

The second tail bound, Chebyshev's inequality, is based on the variance of the

distribution. Chebyshev's inequality is derived by applying Markov's inequality to the deviation of a r.v. from its mean.

Theorem 5.17 (Chebyshev's inequality) *Let X be a r.v. with finite mean $\mu = \mathbf{E}[X]$ and finite variance $\mathbf{Var}(X)$. Then, $\forall a > 0$,*

$$\mathbf{P}\{|X - \mu| \geq a\} \leq \frac{\mathbf{Var}(X)}{a^2}.$$

Proof:

$$\begin{aligned}\mathbf{P}\{|X-\mu| \geq a\} &= \mathbf{P}\left\{(X-\mu)^2 \geq a^2\right\} \\ &\leq \frac{\mathbf{E}\left[(X-\mu)^2\right]}{a^2} \quad \text{(by Markov's inequality)} \\ &= \frac{\mathbf{Var}(X)}{a^2}.\end{aligned}$$

■

5.9.2 Stochastic Dominance

Question: Suppose that r.v. X and r.v. Y are defined on the same sample space, but $X \neq Y$ in distribution. Is it possible that

$$\mathbf{P}\{X > i\} \geq \mathbf{P}\{Y > i\} \quad \forall \text{ values of } i?$$

Answer: Yes! In fact this has a name.

Definition 5.18 *Given two random variables X and Y, if*

$$\mathbf{P}\{X > i\} \geq \mathbf{P}\{Y > i\}, \quad \forall i$$

we say that X **stochastically dominates** *Y. We write this as $X \geq_{st} Y$.*

Example 5.19 (Stochastic dominance)

Figure 5.5 illustrates stochastic dominance of X over Y. Let

$$X = \begin{cases} 1 & \text{w/prob } 0.3 \\ 2 & \text{w/prob } 0.2 \\ 3 & \text{w/prob } 0.4 \\ 4 & \text{w/prob } 0.1 \end{cases} \quad \text{and} \quad Y = \begin{cases} 1 & \text{w/prob } 0.4 \\ 2 & \text{w/prob } 0.4 \\ 3 & \text{w/prob } 0.2 \\ 4 & \text{w/prob } 0 \end{cases}.$$

When looking at the p.m.f. of X and the p.m.f. of Y, it is not at all obvious that X dominates Y. However, when looking at the tails of the distributions, we see

that the tail of X (purple function) is always above or equal to that of Y (red function).

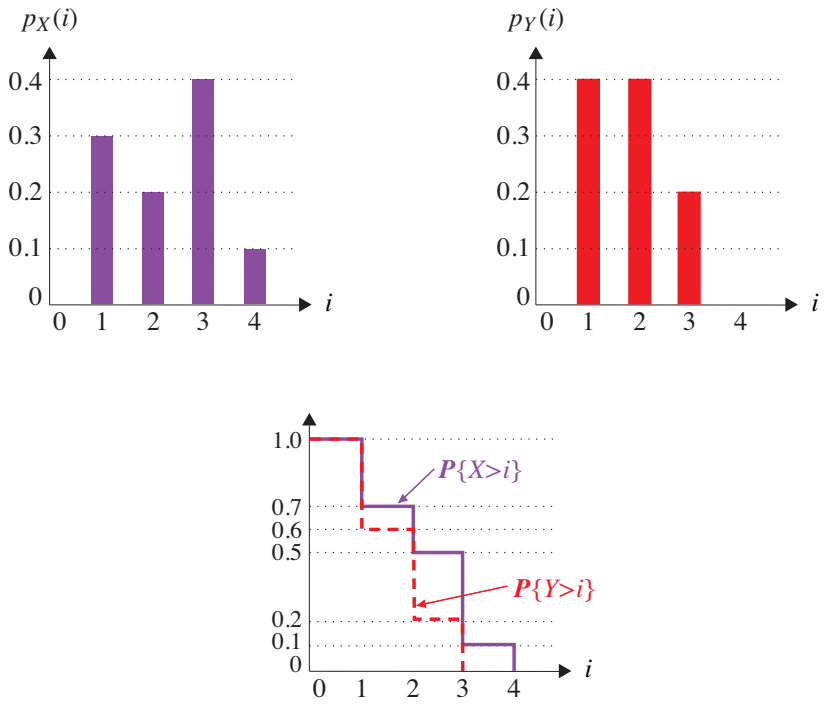


Figure 5.5 *$X \geq_{st} Y$, where X is shown in purple and Y is shown in red.*

Question: When looking at the tail part of Figure 5.5, what does the area under the red (dashed) $\mathbf{P}\{Y > i\}$ function represent?

Answer: The area under the red function is $\mathbf{E}[Y]$ and the area under the purple (solid) $\mathbf{P}\{X > i\}$ function is $\mathbf{E}[X]$. To understand this, recall Exercise 4.16.

Example 5.20 (More shoes are better!)

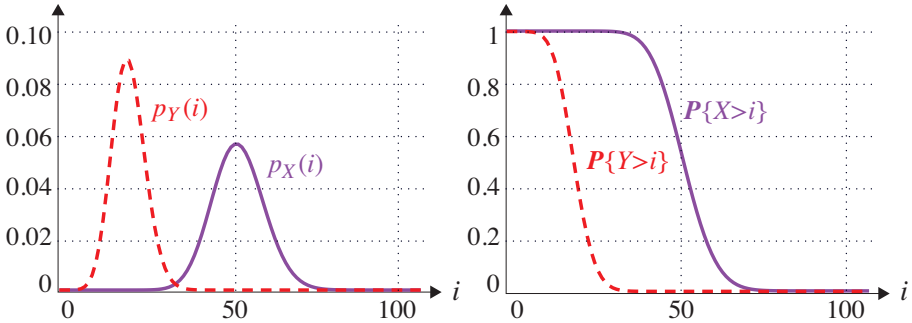
As another example of stochastic dominance, let's look at shoes. My husband likes to tell me that I own way too many shoes (Figure 5.6), but I argue that women stochastically dominate men when it comes to the number of pairs of shoes they own.[1] Let X be a random variable representing the number of pairs of shoes owned by women, where X is reasonably approximated by a Poisson distribution with mean 27. Similarly let $Y \sim \text{Poisson}(12)$ denote the number of pairs of shoes owned by men. While a given man might have more shoes than a

[1] According to a study of shoe brands in the United States, the average man owns 12 pairs of shoes, while the average woman owns 27 pairs [81].

Figure 5.6 *The shoes in my closet.*

given woman, the number of shoes owned by an arbitrary woman stochastically dominates the number owned by an arbitrary man.

Figure 5.7 shows an illustration of two Poisson distributions: $Y \sim \text{Poisson}(20)$ (red/dashed) and $X \sim \text{Poisson}(50)$ (purple/solid). In Figure 5.7(a), we see that $p_X(i)$ is above $p_Y(i)$ for large values of i (although it is below for small values of i). In Figure 5.7(b), we see that $\mathbf{P}\{X > i\}$ is always at least equal to $\mathbf{P}\{Y > i\}$. Thus, we say that Poisson(50) (purple/solid) stochastically dominates Poisson(20) (red/dashed).



(a) p.m.f. comparison (b) tail comparison

Figure 5.7 *The purple (solid) curve represents $X \sim \text{Poisson}(50)$ while the red (dashed) curve represents $Y \sim \text{Poisson}(20)$. The purple curve stochastically dominates the red one.*

Question: If X stochastically dominates Y, and both are non-negative, then it feels like the mean of X should be at least that of Y. Is this true? What about higher moments of X versus Y?

Answer: The answer is yes! See Exercise 5.37.

5.10 Jensen's Inequality

By the definition of variance, and the fact that it must be positive, we know that

$$\mathbf{E}\left[X^2\right] \geq \mathbf{E}\left[X\right]^2.$$

Question: Does it also hold that $\mathbf{E}\left[X^3\right] \geq \mathbf{E}\left[X\right]^3$? Is $\mathbf{E}\left[X^4\right] \geq \mathbf{E}\left[X\right]^4$?

Answer: Yes! Specifically, if X is a positive random variable, then

$$\mathbf{E}\left[X^a\right] \geq \mathbf{E}\left[X\right]^a, \qquad \forall a \in \mathbb{R}, \text{ where } a > 1. \tag{5.7}$$

The proof of (5.7) is given in Exercise 5.32 and follows immediately from Jensen's inequality (Theorem 5.23). Before we can describe Jensen's inequality, we need to review convex functions.

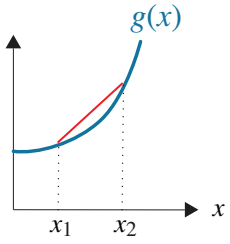


Figure 5.8 *Illustration of convex function $g(x)$.*

Informally a convex function is an upturned curve. More precisely, if we pick any two points on the curve and draw a line segment between these, then the line segment will lie above the curve (see Figure 5.8).

Definition 5.21 *A real-valued function $g(\cdot)$ defined on an interval $S \subseteq \mathbb{R}$ is said to be* **convex** *on S if, for any $x_1, x_2 \in S$ and any $\alpha \in [0,1]$, we have*

$$g\left(\alpha x_1 + (1-\alpha)x_2\right) \leq \alpha g(x_1) + (1-\alpha)g(x_2).$$

To visualize Definition 5.21, observe that

- $\alpha x_1 + (1-\alpha)x_2$ is a weighted average of x_1 and x_2; and
- $\alpha g(x_1) + (1-\alpha)g(x_2)$ is a weighted average of $g(x_1)$ and $g(x_2)$.

Thus Definition 5.21 is saying that if z is any weighted average of x_1 and x_2, then the point $g(z)$ on the curve will always lie below the corresponding point on the line, namely the weighted average of $g(x_1)$ and $g(x_2)$.

Suppose now that X is a r.v. where

$$X = \begin{cases} x_1 & \text{w/prob } p_X(x_1) \\ x_2 & \text{w/prob } p_X(x_2) \end{cases}.$$

Then, for any convex function $g(\cdot)$, Definition 5.21 says that

$$g\left(p_X(x_1)x_1 + p_X(x_2)x_2\right) \leq p_X(x_1)g(x_1) + p_X(x_2)g(x_2). \tag{5.8}$$

Question: What does (5.8) say about $g(\mathbf{E}[X])$?

Answer:

$$g(\mathbf{E}[X]) \leq \mathbf{E}[g(X)].$$

It is easy to generalize Definition 5.21 using induction to obtain Definition 5.22:

Definition 5.22 *A real-valued function $g(\cdot)$ defined on an interval $S \subseteq \mathbb{R}$ is said to be* **convex** *on S if, for any points $x_1, x_2, \ldots, x_n \in S$ and any $\alpha_1, \alpha_2, \ldots, \alpha_n \in [0,1]$, where $\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$, we have*

$$g(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) \leq \alpha_1 g(x_1) + \alpha_2 g(x_2) + \cdots + \alpha_n g(x_n).$$

Let X be a r.v. where

$$X = \begin{cases} x_1 & \text{w/prob } p_X(x_1) \\ x_2 & \text{w/prob } p_X(x_2) \\ \vdots & \\ x_n & \text{w/prob } p_X(x_n) \end{cases}.$$

Question: What does Definition 5.22 say about $\mathbf{E}[g(X)]$?

Answer: Again

$$g\left(p_X(x_1)x_1 + \cdots + p_X(x_n)x_n\right) \leq p_X(x_1)g(x_1) + \cdots + p_X(x_n)g(x_n),$$

so again $g(\mathbf{E}[X]) \leq \mathbf{E}[g(X)]$.

This is summarized by Jensen's inequality:

Theorem 5.23 (Jensen's inequality) *Let X be a r.v. that takes on values in an interval S, and let $g : S \to \mathbb{R}$ be convex on S. Then,*

$$g\left(\mathbf{E}[X]\right) \leq \mathbf{E}[g(X)]. \tag{5.9}$$

We have proven Theorem 5.23 in the case of a discrete r.v. X with finite support.

The theorem also generalizes to the case where X has infinite support and further to the case where X is a continuous r.v. We omit the proof.

Important: A useful method for determining that a function is convex is to check its second derivative. Specifically, $g(\cdot)$ is convex on S if and only if $g''(x) \geq 0$ for all $x \in S$. For example, $g(x) = x^2$ is convex over $\mathbb{R}$, because $g''(x) = 2 \geq 0$.

5.11 Inspection Paradox

We end this chapter by describing one of the more subtle consequences of variability, called the inspection paradox. The inspection paradox says that the mean seen by a random observer can be very different from the true mean. This is best illustrated via examples.

Example 5.24 (Waiting for the bus)

The 61C bus arrives at my bus stop every 10 minutes on average. Specifically, if S denotes the time between buses, then $\mathbf{E}[S] = 10$. I arrive at the bus stop at random times. I would expect that my average wait time for a bus is five minutes. However, I've been monitoring it, and my average wait time is actually eight minutes.

Question: How can this be?

Hint: The answer has to do with the variability of S, specifically its squared coefficient of variation, C_S^2.

Question: If $C_S^2 = 0$, what should the average waiting time of a random arrival be?

Answer: Five minutes, since the person is equally likely to arrive anywhere in $[0, 10]$.

Question: So what goes wrong when C_S^2 is high?

Hint: Looking at Figure 5.9, we see that there are short intervals and long intervals between buses. The average length of an interval is 10 minutes. But which interval is a random arriving person more likely to "land" in?

Answer: A random arriving person is more likely to land in a large interval, thus experiencing an extra-long waiting time. This difference between the true average and the average experienced by a randomly arriving person is what we call the inspection paradox. For a concrete example involving buses, see Exercise 5.20.

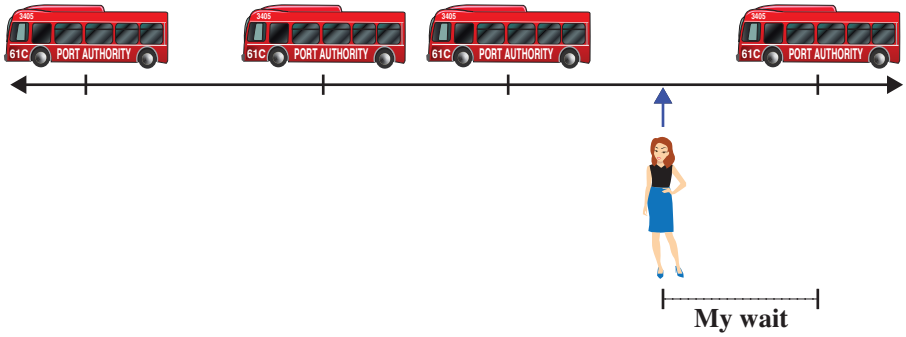


Figure 5.9 *Inspection paradox. The mean time between buses is 10 minutes, so why is my average wait so high?*

Example 5.25 (Class size)

As another example, suppose we ask every student at CMU about the sizes of their classes and we take the average of all these numbers. You will probably hear that the average is somewhere around 100 students in a class. But when you talk to the dean, the dean will tell you that the average class size is 30.

Question: Can the dean and the students both be right?

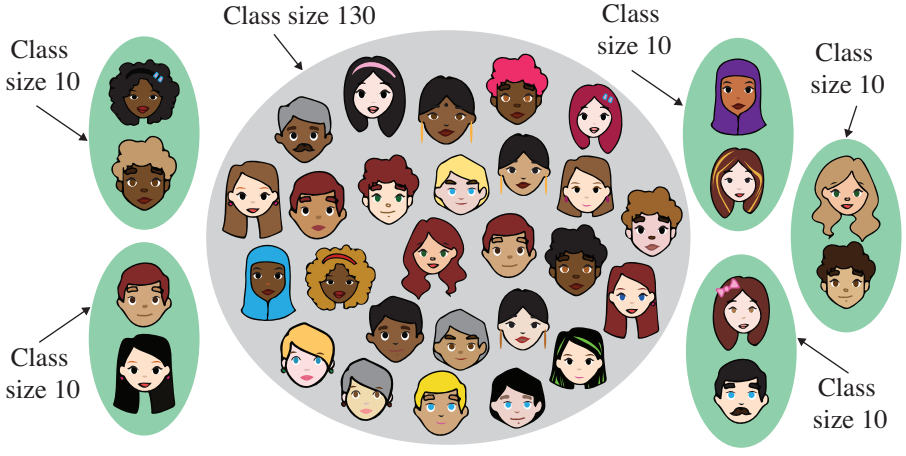


Figure 5.10 *Inspection paradox. The average class size is 30, so why is my average class size so large?*

Answer: Yes! This again is a classic example of the inspection paradox. Figure 5.10 provides an illustration. Say we have five classes of size 10 students and one class of size 130 students. The average across classes is indeed 30. However, most students are in the 130 person class, so they experience a high average.

Question: Suppose that each student takes just one class and there are $50+130 = 180$ students in the school. What is the average *observed* class size?

Answer: $\frac{50}{180}$ fraction of students observe a class size of 10, while $\frac{130}{180}$ fraction of students observe a class size of 130. Thus,

$$\text{Average observed class size } = \frac{50}{180} \cdot 10 + \frac{130}{180} \cdot 130 \approx 97.$$

Example 5.26 (My friends have more Facebook friends than I do!)

As a final example, we consider a study done by Allen Downey which samples 4,000 Facebook users [20]. For each person, p, the study computes the number of friends of p, and the number of friends of each of p's friends. The study found that an average user has 44 friends. However, your average friend has 104 friends. Furthermore, the probability that your friend is more popular than you is 76%.

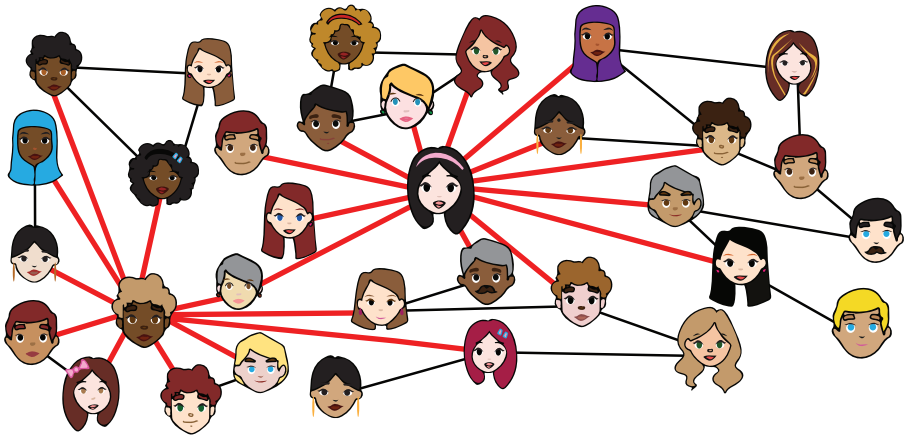


Figure 5.11 *Inspection paradox. Popular people (shown with red links) are more likely to be your friends.*

Question: How can this be?

Answer: This is again an inspection paradox, which stems from the fact that there is variability in the number of friends that people have. As shown in Figure 5.11, most people have a few friends, but a few people have a lot of friends. Let's call a person who has lots of friends a "popular" person. Popular people are simply counted more and thus are more visible to an observer. Consider two potential friends: one popular and one unpopular. The popular person is more likely to be included among your friends than the unpopular one, because the popular person has *lots* of friends. Now, whenever a popular person is included as one of your

friends, this ends up raising the average number of friends that your friends have. The friends phenomenon was originally studied by Scott Feld [26].

5.12 Exercises

5.1 **Simplifying variance**
Simplify each of the following expressions into its simplest form using either definition of variance. Also provide an interpretation of your result by explaining what changes in Figure 5.1.
(a) $\mathbf{Var}(X+5)$
(b) $\mathbf{Var}(X-5)$
(c) $\mathbf{Var}(5X)$
(d) $\mathbf{Var}(-X+3)$

5.2 **Difference of independent random variables**
Let X and Y be discrete random variables where $X \perp Y$. Express $\mathbf{Var}(X-Y)$ in terms of $\mathbf{Var}(X)$ and $\mathbf{Var}(Y)$. Prove it!

5.3 **Sums versus copies**
Let X, Y, and Z be i.i.d. random variables, all distributed as Bernoulli(p). Evaluate the following:
(a) $\mathbf{E}[X+Y+Z]$
(b) $\mathbf{E}[3X]$
(c) $\mathbf{E}[X+Y+Z]^2$
(d) $\mathbf{E}\left[(X+Y+Z)^2\right]$
(e) $\mathbf{E}\left[(3X)^2\right]$

5.4 **The coveted 212 area code**
There are eight million people in NYC. Suppose that each independently is given a phone number with a 212 area code with probability 2%. What is the standard deviation on the number of people who get the coveted 212 area code?

5.5 **Variance of Poisson**
Let $X \sim \text{Poisson}(\lambda)$. Derive $\mathbf{Var}(X)$.

5.6 **Die throws**
Let X_1 and X_2 be the results of two independent die throws. Which is larger $\mathbf{E}[X_1X_2]$ or $\mathbf{E}\left[X_1^2\right]$? Or are they the same? Compute each.

5.7 **Understanding variance and risk**
Let $X_1, X_2, \ldots, X_c$ be i.i.d. instances of r.v. X.

(a) Which is lower: $\mathbf{Var}(X_1 + X_2 + \cdots + X_c)$ or $\mathbf{Var}(cX)$? Compute each.
(b) A mutual fund allows you to buy a small piece of many different companies, as opposed to buying a large piece of a single company. It is said that investing in a mutual fund is less risky than investing in a single company. Explain this statement via your analysis in part (a).

5.8 **Grade of A**
The average grade on the first probability exam is 70%. The "A" grade cutoff is 90%. What is an upper bound on the fraction of students who get an "A"?
(a) Assume that we have no other knowledge, and use Markov's inequality.
(b) Assume that we know the standard deviation of grades is 5%, and apply Chebyshev's inequality.

5.9 **Chebyshev's inequality**
Show that Chebyshev's inequality guarantees that the probability of deviating from the mean by more than k standard deviations is less than $\frac{1}{k^2}$. Specifically, if X is any random variable with mean μ and finite variance σ^2, then for any real number $k > 0$,

$$\mathbf{P}\{|X - \mu| \geq k\sigma_X\} \leq \frac{1}{k^2}.$$

5.10 **Stochastic dominance of Geometrics**
Let $X \sim \text{Geometric}(0.2)$, $Y \sim \text{Geometric}(0.4)$, where $X \perp Y$. What is $\mathbf{P}\{X > Y\}$? Is $X \geq_{st} Y$?

5.11 **Applications of Jensen's inequality**
Let X be a positive random variable.
(a) How do $\mathbf{E}\left[X^{-1}\right]$ and $\mathbf{E}\left[X\right]^{-1}$ compare?
(b) How do $\mathbf{E}\left[e^X\right]$ and $e^{\mathbf{E}[X]}$ compare?

5.12 **Zero covariance**
(a) Prove that if X and Y are independent random variables, then $\mathbf{Cov}(X, Y) = 0$.
(b) Show that the converse is not true. That is, $\mathbf{Cov}(X, Y) = 0$ does *not* imply that $X \perp Y$. [Hint: Find a counter-example.]

5.13 **Using covariance to express variance of a sum**
Let X and Y be random variables. Prove that

$$\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y) + 2\mathbf{Cov}(X, Y). \tag{5.10}$$

Equation (5.10) can be generalized to:

$$\mathbf{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbf{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathbf{Cov}(X_i, X_j) \tag{5.11}$$

for random variables $X_1, X_2, \ldots, X_n$. You do not have to prove (5.11).

5.14 **Covariance and events**

Let X and Y be indicator random variables, where

$$X = \begin{cases} 1 & \text{if event } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases} \qquad Y = \begin{cases} 1 & \text{if event } B \text{ occurs} \\ 0 & \text{otherwise} \end{cases}.$$

Prove that $\mathbf{Cov}(X, Y) > 0$ if and only if events A and B are positively correlated (see Exercise 2.8 for the definition of positively correlated events).

5.15 **Cauchy–Schwarz inequality**

In this problem, you will prove the Cauchy–Schwarz inequality for random variables which says that for any two random variables X and Y,

$$\left|\mathbf{E}\left[XY\right]\right| \le \sqrt{\mathbf{E}\left[X^2\right]\mathbf{E}\left[Y^2\right]}. \tag{5.12}$$

Follow these steps:

(a) Let $Z = (X - cY)^2$, where c is a constant. Explain why $\mathbf{E}\left[Z\right] \ge 0$.

(b) Now substitute in $c = \frac{\mathbf{E}[XY]}{\mathbf{E}\left[Y^2\right]}$ and simplify until you get (5.12).

5.16 **Correlation coefficient**

Recall that $\mathbf{Cov}(X, Y)$ can be arbitrarily high or low depending on the magnitude of X and Y. In practice, it is common to use a normalized version on covariance called the correlation coefficient, $\rho(X, Y)$, where

$$\rho(X, Y) \equiv \frac{\mathbf{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y represent the standard deviations of X and Y respectively. Prove that the magnitude of $\mathbf{Cov}(X, Y)$ is bounded, specifically

$$-1 \le \rho(X, Y) \le 1.$$

[Hint 1: It helps to use the Cauchy–Schwarz inequality from Exercise 5.15.]
[Hint 2: Start by working with $V = \frac{1}{\sigma_X}(X - \mathbf{E}\left[X\right])$ and $W = \frac{1}{\sigma_Y}(Y - \mathbf{E}\left[Y\right])$.]

5.17 **Sampling without replacement: variance and covariance**

Suppose that we have an urn that contains b balls numbered $1, 2, \ldots b$. We draw $n \le b$ balls at random from the urn, one at a time, without replacement. Let X_i denote the number on the ith ball drawn.

(a) What is $\mathbf{P}\left\{X_i = k\right\}$?

(b) Show that

$$\mathbf{Var}(X_i) = \frac{(b-1)(b+1)}{12}. \tag{5.13}$$

[Hint: Use the identity that $\sum_{i=1}^{b} i^2 = \frac{1}{6}b(b+1)(2b+1)$.]

(c) Follow the steps below to show that:

$$\mathbf{Cov}\left(X_i, X_j\right) = -\frac{b+1}{12}. \tag{5.14}$$

(i) What is $\mathbf{P}\left\{X_i = k_1, X_j = k_2\right\}$?

(ii) Explain why $\mathbf{Var}\left(\sum_{i=1}^{b} X_i\right) = 0$.

(iii) Apply $\mathbf{Var}\left(\sum_{i=1}^{b} X_i\right) = \sum_{i=1}^{b} \mathbf{Var}(X_i) + 2\sum_{1\le i<j\le b} \mathbf{Cov}\left(X_i, X_j\right)$ from (5.11) to get $\mathbf{Cov}\left(X_i, X_j\right)$.

(iv) Explain why it makes sense that $\mathbf{Cov}\left(X_i, X_j\right)$ is negative.

5.18 **Memorylessness of Geometric**

Let $X \sim \text{Geometric}(p)$. Let $Y = [X \mid X > 1]$. You will prove that

$$Y \overset{d}{=} 1 + X.$$

(a) Argue that Y and $1 + X$ have the same sample space of possible values.

(b) Write a simple expression for $\mathbf{P}\{Y = i\}$, where $i \ge 2$.

(c) Write a simple expression for $\mathbf{P}\{1 + X = i\}$, where $i \ge 2$.

Your answers for parts (b) and (c) should be the same.

5.19 **Variance of the Geometric**

Let $X \sim \text{Geometric}(p)$. Derive $\mathbf{Var}(X) = \frac{1-p}{p^2}$. [Hint: Use conditioning.]

5.20 **Buses and the inspection paradox**

Suppose that half of all buses arrive 5 minutes after the previous bus, and half arrive 15 minutes after the previous bus. Let r.v. S denote the time between buses.

(a) What is $\mathbf{E}[S]$?

(b) If you arrive at a random time, what is the expected length of the inter-bus interval that you find yourself in?

(c) Let's consider a more extreme example where half of all buses arrive $\epsilon > 0$ minutes after the previous bus, while half arrive $20 - \epsilon$ minutes after the previous bus. How do your answers to (a) and (b) change? Derive the answers in the limit as $\epsilon \to 0$.

5.21 **Happy gambling**

At the Happy Casino, at every turn you earn a dollar with probability 0.6 and lose a dollar with probability 0.4. Let W denote your total money won after n games (this could be positive or negative).

(a) What is $\mathbf{E}[W]$?
(b) What is $\mathbf{Var}(W)$?

5.22 **Good chips versus lemons**
A chip supplier produces 95% good chips and 5% lemons (bad chips). The good chips fail with probability 0.0001 each day. The lemons fail with probability 0.01 each day. You buy a random chip. Let T be the time until your chip fails. Compute $\mathbf{E}[T]$ and $\mathbf{Var}(T)$.

5.23 **Napster**
As a present for my brother, I decided to create a collection of all 50 songs from his favorite band. Unfortunately, whenever I typed in the band name, I was sent a *random* song from the band. Let D denote the number of downloads required to get all 50 songs.
(a) What is $\mathbf{E}[D]$? Give a closed-form approximation.
(b) What is $\mathbf{Var}(D)$? (No need for closed-form here.)

5.24 **Ensuring Internet connectivity**
Janice manages the wireless Internet connection in a building. Let N denote the number of occupants in the building each day, where $\mathbf{E}[N] = 100$ and $\sigma_N = 10$. Each occupant needs Internet connectivity. Suppose that one wireless access point can serve $m = 10$ occupants. Janice wants to use as few access points as possible, while ensuring all occupants of the building get Internet connectivity.
(a) Suppose that on a given day Janice wants to ensure that, with probability at least 80%, all occupants get Internet connectivity. According to Markov's inequality, how many access points, n, does she need?
(b) Repeat part (a), this time using the Chebyshev bound.

5.25 **Hypothesis testing in data analysis**
In hypothesis testing, a decision between two alternatives, one of which is called the "null hypothesis" (H_0) and the other the "alternative hypothesis" (H_1), is to be made. You are given a coin with probability p of heads, and you want to test if it is fair or biased in favor of heads. Here,
H_0: Coin is fair (that is, $p = 0.5$).
H_1: Coin is biased toward heads (that is, $p > 0.5$).
You perform an experiment of tossing the coin $n = 10$ times and observe $k = 8$ heads. Based on this outcome, you have to decide whether to "reject H_0" (that is, choose H_1). A popular approach used in making such decisions is based on the "p-value." The p-value of an outcome is the probability that the observed outcome, *or something more extreme* than the observed outcome, occurs under the assumption that H_0 is true. Here, "more extreme" means more in favor of H_1.
(a) What is the p-value for the outcome of your experiment?

(b) To be more confident in choosing between hypotheses H_0 and H_1, should the associated p-value be higher or lower?
(c) To be confident of your decision, you set the p-value at 0.01. How many heads do you need to observe in the experiment in order to choose H_1.

5.26 **Mouse in a maze**

[Problem adapted from Sheldon Ross.] A mouse is trapped in a maze. Initially it has to choose one of two directions. If it goes to the right, then it will wander around in the maze for three minutes and will then return to its initial position. If it goes to the left, then with probability $\frac{1}{3}$ it will depart the maze after two minutes of traveling, and with probability $\frac{2}{3}$ it will return to its initial position after five minutes. The mouse is at all times equally likely to go to the left or the right. Let T denote the number of minutes that it will be trapped in the maze.
(a) What is $\mathbf{E}[T]$?
(b) What is $\mathbf{Var}(T)$?

5.27 **Central moments**

Recall that when X and Y are independent random variables, we have:

$$\mathbf{Var}(X+Y) = \mathbf{Var}(X) + \mathbf{Var}(Y).$$

Let $\mathbf{Skew}(X)$ denote the third central moment of X, that is,

$$\mathbf{Skew}(X) = \mathbf{E}\left[(X - \mathbf{E}[X])^3\right].$$

Either prove or disprove (via a counter-example) that, for independent X and Y:

$$\mathbf{Skew}(X+Y) = \mathbf{Skew}(X) + \mathbf{Skew}(Y).$$

[Hint: It may help to define $X' = X - \mathbf{E}[X]$ and $Y' = Y - \mathbf{E}[Y]$ and then restate the problem in terms of X' and Y'.]

5.28 **All I do is sleep and work**

A typical CMU student's life consists of alternating between home and school every hour, according to Figure 5.12. If the student is home, with probability p she will switch to school at the next hour (otherwise she will stay home). If the student is at school, with probability q she will switch to home at the next hour (otherwise she will stay at school). Assuming the student just got to school, let T be the time (in hours) until the student goes home. What is $\mathbf{Var}(T)$?

5.29 **Dominance**

[Proposed by Weina Wang] Suppose that X and Y represent the result of coin flips, where

$$X \sim \text{Bernoulli}(0.5) \qquad \text{and} \qquad Y \sim \text{Bernoulli}(0.6).$$

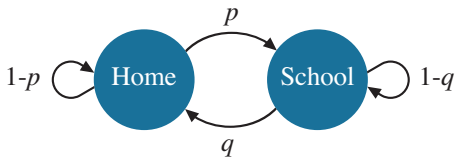


Figure 5.12 *Figure for Exercise 5.28.*

Clearly sometimes the value of X exceeds that of Y, although slightly more often the value of Y exceeds that of X. Define a joint probability distribution, $p_{X,Y}(x, y)$ where the marginal distributions are $p_X(x) \sim$ Bernoulli(0.5) and $p_Y(y) \sim$ Bernoulli(0.6), but $\mathbf{P}\{X \leq Y\} = 1$.

5.30 **All I do is sleep, work, and drink coffee**

Imagine a poor student caught in an endless cycle between sleeping, working, and drinking coffee at the coffee house. The student's life is described by Figure 5.13, where the student is always in one of three states, and every hour the student transitions (possibly back to the same state) with the probability shown. For example, after drinking a cup of coffee, the student will, at the next hour, with probability $\frac{1}{3}$ go back to work, or with probability $\frac{2}{3}$ stay to drink another cup of coffee. Assuming that the student is at work, let T denote the number of hours until she goes home to sleep.

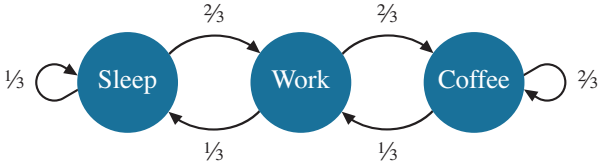


Figure 5.13 *Figure for Exercise 5.30.*

(a) What is $\mathbf{E}[T]$?
(b) What is $\mathbf{Var}(T)$?

5.31 **Average of random number of random variables**

Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables with distribution X. Let N be a positive, integer-valued r.v., where $N \perp X$. Let

$$A = \frac{1}{N} \sum_{i=1}^{N} X_i.$$

(a) Derive $\mathbf{E}[A]$.
(b) Derive $\mathbf{Var}(A)$.

5.32 **Higher moment inequalities**

Use Jensen's inequality to prove that for any positive r.v. X,

$$\mathbf{E}\left[X^a\right] \geq \mathbf{E}\left[X\right]^a, \qquad \forall a \in \mathbb{R}, \text{ where } a > 1.$$

5.33 **Summing up to a stopping time**

[Proposed by Tianxin Xu] Imagine I roll a fair three-sided die.

- If the die comes up 1, I give you one dollar, and I role again.
- If the die comes up 2, I give you two dollars, and I role again.
- If the die comes up 3, I give you three dollars, but we stop playing.

Let S denote the total amount of money that I give you during the game. Observe

$$S = \sum_{i=1}^{N} X_i,$$

where X_i is the result of the ith role, and N is the number of rolls until we see a 3 (inclusive). Your goal is to compute $\mathbf{E}\left[S\right]$ and $\mathbf{Var}(S)$.

(a) Explain why we can't apply Theorem 5.14.

(b) Compute $\mathbf{E}\left[S\right]$. [Hint: Condition on the first roll.] Is your answer the same as in Theorem 5.14?

(c) Now compute $\mathbf{Var}(S)$. Is your answer the same as in Theorem 5.14?

The r.v. N in this problem is called a "stopping time" because its value only depends on the X_i's that were seen so far, and not on the future. When N is a stopping time, and the X_i's are i.i.d. with $X_i \sim X$, an identity called *Wald's equation* says that $\mathbf{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbf{E}\left[N\right] \cdot \mathbf{E}\left[X\right]$ [74].

5.34 **Skewering the Binomial**

Let $\mathbf{Skew}(X) = \mathbf{E}\left[(X - \mathbf{E}\left[X\right])^3\right]$. If $Y \sim \text{Binomial}(n, p)$, what is $\mathbf{Skew}(Y)$?

5.35 **Race to win**

Obama and Romney are counting votes as they come in. Suppose that each incoming vote is for Obama with probability $p = 0.6$ and is for Romney with probability $1 - p = 0.4$. At the moment when Obama has 100 votes, we'd like to understand how many votes Romney has. Let R denote the number of Romney votes at the moment when Obama gets his 100th vote.[2]

(a) What is $p_R(i)$? (We want the probability of the event that there are i Romney votes *and* 100 Obama votes *and* that the last vote is for Obama.)

(b) What is $\mathbf{E}\left[R\right]$? [Hint: If you try to derive $\mathbf{E}\left[R\right]$ from $p_R(i)$, you will find it hard. Look for the much easier way. Hint: Linearity.]

(c) What is $\mathbf{Var}(R)$? [Hint: This should be easy after (b).]

[2] This is an instance of a Negative Binomial distribution.

5.36 **Cups at a party**

There are n people at a party. Each person puts their cup down on the table. Then they each pick up a random cup.

(a) What is the expected number of people who get back their own cup?

(b) Derive the variance of the number of people who get back their own cup.

5.37 **Stochastic dominance**

Let X and Y be non-negative, discrete, integer-valued random variables. We are given that $X \geq_{st} Y$.

(a) Prove that $\mathbf{E}[X] \geq \mathbf{E}[Y]$.

(b) Prove that $\mathbf{E}\left[X^2\right] \geq \mathbf{E}\left[Y^2\right]$.
[Hint: Compare $\sum_i i \cdot \mathbf{P}\{X > i\}$ with $\sum_i i \cdot \mathbf{P}\{Y > i\}$.]

5.38 **Pairwise independence**

Consider n random variables: $X_1, X_2, \ldots, X_n$. We say that these are *pairwise independent* if any two of these are independent, that is,

$$\forall i \neq j, \qquad \mathbf{P}\left\{X_i = i \;\&\; X_j = j\right\} = \mathbf{P}\{X_i = i\} \cdot \mathbf{P}\left\{X_j = j\right\}.$$

We will show that if $X_1, X_2, \ldots, X_n$ are pairwise independent, then:

$$\mathbf{Var}(X_1 + X_2 + \cdots + X_n) = \mathbf{Var}(X_1) + \mathbf{Var}(X_2) + \cdots + \mathbf{Var}(X_n).$$

(a) Prove the desired linearity theorem in the case where $\mathbf{E}[X_i] = 0, \forall i$.

(b) For the rest of the problem, assume that $\mathbf{E}[X_i] \neq 0$. Define $Y_i = X_i - \mathbf{E}[X_i]$. What is $\mathbf{E}[Y_i]$?

(c) What does your result from part (a) say about the linearity of $\mathbf{Var}(Y_1 + Y_2 + \cdots + Y_n)$? After writing down the linearity statement for the Y_i's, substitute back $Y_i = X_i - \mathbf{E}[X_i]$. You will see that you can claim a linearity result for the X_i's as well.

5.39 **Total variation distance**

We often want to express the "distance" between two distributions. There are many ways to define such a distance. One way is the total variation distance (TVD). Given two discrete distributions, X and Y, we write:

$$\mathbf{TVD}(X, Y) = \frac{1}{2} \sum_i \left|\mathbf{P}\{X = i\} - \mathbf{P}\{Y = i\}\right|.$$

Prove two properties of $\mathbf{TVD}(X, Y)$:

(a) Prove that $\mathbf{TVD}(X, Y) \leq 1$.

(b) Prove that $\mathbf{TVD}(X, Y) \leq \mathbf{P}\{X \neq Y\}$.