Lecture 7: Variance, Higher moments, and the Sum of a random variable number of random variables

1 Variance

Definition 1 *For a r.v. X, we define the* **variance of** X*, written as* $\mathbf{Var}(X)$ *as*

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

$\mathbf{Var}(X)$ is intended to measure how much X deviates from its mean. More precisely, $\mathbf{Var}(X)$ is the expected *squared* deviation between an instance of the r.v. X and its mean.

Example: Variance of Bernoulli

Suppose that $X \sim \text{Bernoulli}(p)$

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

Observe that $\mathbf{E}\{X\} = p$.

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

Commit the above formula to memory! You will use it again and again.

Question: What is the variance of a fair coin?

Answer: For a fair coin, $p = \frac{1}{2}$, so variance is $\frac{1}{4}$. This should maike sense because the difference between a flip and the mean of $\frac{1}{2}$ is always $\frac{1}{2}$, no matter what the flip. Hence the difference squared is always $\frac{1}{4}$.

2 Other notions of variance

Question: Why is the variance defined as the square of the difference, rather than just the difference?

Answer: A negative difference is still a deviation. By taking the square, we keep all the deviations positive. We could have instead used absolute values. However the square has certain nice properties that we will see later in Theorem 4.

People often try to "undo" the effect of the squaring, by looking at the *standard deviation* rather than the variance.

Definition 2 *The* **standard deviation** *of a random variable, X, is denoted by σ_X, where $\sigma_X = \sqrt{\mathbf{Var}(X)}$, the (positive) square root of the variance. Equivalently, it is common to denote* $\mathbf{Var}(X)$ *by* σ_X^2.

Question: Is the variance of a random variable influenced by scaling?

Hint: Consider the following two random variables:

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

Do they have the same variance? Do they have the same standard deviation?

Answer:

$$\begin{aligned} \mathbf{Var}(X) &= \frac{1}{3}(3-2)^2 + \frac{1}{3}(2-2)^2 + \frac{1}{3}(1-2)^2 \\ &= \frac{1}{3}(1) + \frac{1}{3}(0) + \frac{1}{3}(1) \\ &= \frac{2}{3} \end{aligned}$$

$$\begin{aligned} \mathbf{Var}(Y) &= \frac{1}{3}(100) + \frac{1}{3}(0) + \frac{1}{3}(100) \\ &= \frac{200}{3} \end{aligned}$$

The variances are very different for X and Y (same for the standard deviations). Both the variance and standard deviation are definitely influenced by the scaling.

Question: Can we somehow scale the variance to make the two answers the same?

Answer: Yes, a common metric used in computer systems is the squared coefficient of variation, defined below.

Definition 3 *The* **squared coefficient of variation of** X*, written* C_X^2*, is defined as:*

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

The squared coefficient of variation can be thought of as normalizing the variance. To see this, let's compute C_X^2 and C_Y^2 for the above r.v.s:

$$\begin{aligned} C_X^2 &= \frac{\mathbf{Var}(X)}{\mathbf{E}\{X\}^2} = \frac{\frac{2}{3}}{4} = \frac{1}{6} \\ C_Y^2 &= \frac{\mathbf{Var}(Y)}{\mathbf{E}\{Y\}^2} = \frac{\frac{200}{3}}{400} = \frac{1}{6} \end{aligned}$$

Thus, once we normalize the variance, X and Y in fact look to have the same variability, which is more in line with our intuition.

3 An easier way to compute variance

Returning to variance, we observe that the definition of variance can be rewritten in an alternative form:

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

The above expression is easy to derive from the original definition of variance via Linearity of Expectations as follows:

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

This equivalent definition of variance is often easier to use in practice. Let's consider a few examples.

Example: Variance of Geometric(p)

Let $X \sim$ Geometric(p). Then $\mathbf{E}\{X\} = \frac{1}{p}$.

The *hard way* to do this problem is to now write:

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

This sum can be solved by integrating twice. However you don't need to do all this...

The *easy way* to do this problem is to instead condition on the first flip:

$$\begin{aligned}
\mathbf{E}\left\{X^2\right\} &= 1 \cdot p + \mathbf{E}\left\{(1+X)^2\right\} \cdot (1-p) \\
&= 1 \cdot p + \mathbf{E}\left\{1 + 2X + X^2\right\} \cdot (1-p) \\
&= p + \left(1 + 2\mathbf{E}\{X\} + \mathbf{E}\left\{X^2\right\}\right)(1-p) \\
&= p + (1-p) + \frac{2}{p}(1-p) + \mathbf{E}\left\{X^2\right\}(1-p) \\
p\mathbf{E}\left\{X^2\right\} &= 1 + \frac{2(1-p)}{p} \\
\mathbf{E}\left\{X^2\right\} &= \frac{1}{p} + \frac{2(1-p)}{p^2} \\
&= \frac{1}{p} + \frac{2}{p^2} - \frac{2}{p} \\
&= \frac{2}{p^2} - \frac{1}{p} \\
&= \frac{2-p}{p^2}
\end{aligned}$$

$$\begin{aligned}
\mathbf{Var}(X) &= \mathbf{E}\left\{X^2\right\} - \mathbf{E}\{X\}^2 \\
&= \frac{2-p}{p^2} - \frac{1}{p^2} \\
&= \frac{1-p}{p^2}
\end{aligned}$$

4 Sums of variances

Recall that Linearity of Expectation tells us that:

$$\mathbf{E}\{X + Y\} = \mathbf{E}\{X\} + \mathbf{E}\{Y\}$$

Something similar can be said for variances, but we require that X and Y be independent.

Theorem 4 *Let X and Y be random variables where $X \perp Y$. Then*

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

Proof:

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

■

Observe that Theorem 4 does not hold for other notions of variance like the squared coefficient of variation, or a notion of variance where we take absolute values of the difference between a r.v. and its mean.

Example: Variance of Binomial

Let $X \sim \text{Binomial}(n, p)$.

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

Answer: We will use the same trick we used in deriving the mean of Binomial(n, p).

Let

$$X = \text{number of successes in n trials}$$

$$X = X_1 + X_2 + \cdots + X_n$$

where

$$\begin{aligned} X_i &= \begin{cases} 1 & \text{if trial i is successful} \\ 0 & \text{otherwise} \end{cases} \\ \mathbf{E}\{X_i\} &= p \\ \mathbf{Var}(X_i) &= \mathbf{E}\left\{X_i^2\right\} - \mathbf{E}\{X_i\}^2 \\ &= p - p^2 \\ &= p(1-p) \end{aligned}$$

Since the X_i's are independent, we have that:

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

Example: Downloading all songs – a.k.a. Coupon Collector

As a present for my brother, I decided to create a collection of songs from his favorite band. I needed to download 50 songs. 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.

Question: Compute $\mathbf{E}\{D\}$ and $\mathbf{Var}(D)$

Hint: Let D_i denote the number of downloads required to get the ith *new* song after the $i-1$th new song has been downloaded. How is D_i distributed?

Answer: Observe that

$$D_i \sim \text{Geometric}\left(\frac{51-i}{50}\right)$$

Thus

$$\mathbf{E}\{D_i\} = \frac{50}{51-i}$$

Furthermore

$$\mathbf{Var}(D_i) = \frac{1 - \frac{51-i}{50}}{\frac{(51-i)^2}{50^2}} = \frac{50i - 50}{(51-i)^2}$$

Also observe that the D_i's are independent.

$$\begin{aligned}
\mathbf{E}\{D\} &= \mathbf{E}\{D_1 + D_2 + \cdots + D_{50}\} \\
&= \mathbf{E}\{D_1\} + \mathbf{E}\{D_2\} + \cdots + \mathbf{E}\{D_{50}\} \\
&= 50\left(\frac{1}{50} + \frac{1}{49} + \cdots + 1\right) \\
&\approx 50\ln(50)
\end{aligned}$$

$$\begin{aligned}
\mathbf{Var}(D) &= \mathbf{Var}(D_1 + D_2 + \cdots + D_{50}) \\
&= \mathbf{Var}(D_1) + \mathbf{Var}(D_2) + \cdots + \mathbf{Var}(D_{50}) \\
&= \sum_{i=1}^{50} \frac{50i - 50}{(51-i)^2}
\end{aligned}$$

5 Higher moments

Definition 5 *The* **kth moment** *of a random variable, X, is defined as* $\mathbf{E}\left\{X^k\right\}$. *In the case where X is discrete, we have*

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

The **kth central moment** *is defined as* $\mathbf{E}\left\{(X - \mathbf{E}\{X\})^k\right\}$.

Just as it was difficult to understand intuition behind the definition of variance, it is even harder to understand what the higher moments tell us, but let's give it a try.

Remember that the centralized third moment is a sum of both negative and positive values, where the values are negative when X lies below the mean and positive when X lies above the mean.

Question: Suppose a distribution is symmetric about its mean. What is the 3rd central moment for such a distribution?

Answer: 0.

The 3rd central moment is related to how skewed a distribution is to the right or the left. If a distribution is skewed to the left, meaning that there are some points very

far from the mean on the left side, then the skew is negative (because those points contribute a large negative component to the sum). Likewise, if a distribution is skewed to the right, meaning that there are some points very far from the mean on the right side, then the skew is positive. Figure 1 illustrates this point.

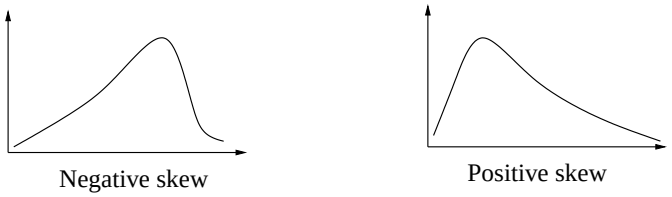

Figure 1: *Illustration of skewed distributions.*

One thing that's important to understand about moments is that the moments of a distribution define the distribution. Suppose, for example, that we have a discrete distribution with n values, denoted by the r.v. X. That is:

$$X = \begin{cases} x_1 & \text{w/prob } p_1 \\ x_2 & \text{w/prob } p_2 \\ x_3 & \text{w/prob } p_3 \\ \vdots & \\ x_n & \text{w/prob } p_n \end{cases}$$

where we don't know the x_i's or the p_i's.

Question: How many moments of X do you need to know before you've fully specified the distribution?

Answer: There are $2n - 1$ free variables. If we are given the first $2n - 1$ moments, that gives us $2n - 1$ equations involving these variables. In theory that should be enough to specify the distribution . . . although in practice it may still not be easy.

We will gain a better understanding of how moments define a distribution when we get to transforms (generating functions).

6 Sum of a random number of random variables

It is common in many applications that one needs to add up a number of *independent and identically distributed (i.i.d.)* random variables, where the number of these variables is itself a random variable. Specifically, we're talking about the quantity S below.

Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables. Let

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

where N is a non-negative, integer-valued random variable.

For example, N might be Geometrically distributed, meaning that we keep adding instances of random variable X, each time flipping a coin right after, until the coin comes up "heads," meaning that we can stop.

In this section we discuss how to derive quantities like $\mathbf{E}\{S\}$ and $\mathbf{E}\left\{S^2\right\}$.

Question: Why can't we directly apply linearity of expectations?

Answer: Linearity equations only apply when N is a constant.

Question: Does this give you any ideas?

Answer: Let's condition on the value of N, and then apply linearity of expectations.

$$\begin{aligned}
\mathbf{E}\{S\} = \mathbf{E}\left\{\sum_{i=1}^{N} X_i\right\} &= \sum_n \mathbf{E}\left\{\sum_{i=1}^{N} X_i \mid N = n\right\} \cdot \mathbf{P}\{N = n\} \\
&= \sum_n \mathbf{E}\left\{\sum_{i=1}^{n} X_i\right\} \cdot \mathbf{P}\{N = n\} \\
&= \sum_n n\mathbf{E}\{X\} \cdot \mathbf{P}\{N = n\} \\
&= \mathbf{E}\{X\} \cdot \mathbf{E}\{N\} \qquad (1)
\end{aligned}$$

Question: Can we use the same trick to get $\mathbf{E}\left\{S^2\right\}$?

Answer: The difficulty with conditioning on N is that we end up with a big sum that we need to square, and it's not obvious how to do that. Consider the following:

$$\begin{aligned}
\mathbf{E}\left\{S^2\right\} &= \sum_n \mathbf{E}\left\{S^2 \mid N = n\right\} \cdot \mathbf{P}\{N = n\} \\
&= \sum_n \mathbf{E}\left\{\left(\sum_{i=1}^{n} X_i\right)^2\right\} \cdot \mathbf{P}\{N = n\}
\end{aligned}$$

A better idea is to first derive $\mathbf{Var}(S \mid N = n)$ and then use that to get $\mathbf{E}\left\{S^2 \mid N = n\right\}$.

Question: What is $\mathbf{Var}(S \mid N = n)$?

Answer: By independence,

$$\mathbf{Var}(S \mid N = n) = n\mathbf{Var}(X)$$

Observe also that

$$\begin{aligned} n\mathbf{Var}(X) = \mathbf{Var}(S \mid N = n) &= \mathbf{E}\left\{S^2 \mid N = n\right\} - (\mathbf{E}\{S \mid N = n\})^2 \\ &= \mathbf{E}\left\{S^2 \mid N = n\right\} - (n\mathbf{E}\{X\})^2 \end{aligned}$$

From the above expression, we have that:

$$\mathbf{E}\left\{S^2 \mid N = n\right\} = n\mathbf{Var}(X) + n^2 (\mathbf{E}\{X\})^2$$

It follows that:

$$\begin{aligned} \mathbf{E}\left\{S^2\right\} &= \sum_n \mathbf{E}\left\{S^2 \mid N = n\right\} \cdot \mathbf{P}\{N = n\} \\ &= \sum_n \left(n\mathbf{Var}(X) + n^2 (\mathbf{E}\{X\})^2\right) \mathbf{P}\{N = n\} \\ &= \mathbf{E}\{N\}\,\mathbf{Var}(X) + \mathbf{E}\left\{N^2\right\} (\mathbf{E}\{X\})^2 \end{aligned}$$

Furthermore:

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

We have proven the following theorem:

Theorem 6 *Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables. Let*

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

Then

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

The variance trick was pretty cool. You may be wondering how we would get the third moment, $\mathbf{E}\left\{S^3\right\}$, if we ever needed it, given that the variance trick won't work there. The answer is to use transform analysis (generating functions), which will easily provide any moment of S. We'll get there soon!