Chapter 3. Discrete Random Variables

3.3: Variance

(From "Probability & Statistics with Applications to Computing" by Alex Tsun)

3.3.1 Linearity of Expectation with Indicator RVs

We've seen how useful and important linearity of expectation was (e.g., with the frog example). We'll now see how to apply it in a very clever way that is very commonly used to solve seemingly difficult problems.

Suppose there are 7 mermaids in the sea. Below is a table that represents these mermaids and the colors of their hair.

Mermaid	1	2	3	4	5	6	7
Color	RED	BLUE	PURPLE	RED	BLACK	YELLOW	RED
1 / 0	1	0	0	1	0	0	1

Each column in the third row of the table is a variable, X_i, that is 1 if the i-th mermaid has red hair and 0 otherwise. We call these sorts of variables *indicator variables* because they are either 1 or 0, and their values indicate the truth of a boolean (red hair or not).

Let the variable X represent how many of the 7 mermaids have red hair. If I only gave you this third row ($X_1, X_2, \ldots, X_7$ of 1's and 0's), how could you compute X?

Well, you would add them all up! $X = X_1 + X_2 + ... + X_7 = 3$. So, there are 3 mermaids in the sea that have red hair. This might seem like a trivial result, but let's go over a more complicated an example to illustrate the usefulness of indicator random variables!

Example(s)

Suppose n people go to a party and leave their hat with the hat-check person. At the end of the party, she returns hats randomly and uniformly because she does not care about her job. Let X be the number of people who get their original hat back. What is $\mathbb{E}[X]$?

Solution Your first instinct might be to approach this problem with brute force. Such an approach would involve enumerating the range, $\Omega_X = \{0, 1, 2, ..., n-2, n\}$ (all the integers from 0 to n, except $n-1$), and computing the probability mass function for each of its elements. However, this approach will get very complicated (give it a shot). So, let's use our new friend, linearity of expectation.

Quick Observation: Does it matter where you are in line?
If we are first in line, $\mathbb{P}(\text{gets hat back}) = \frac{1}{n}$, because there are n in total and each is equally likely.
If we are last in line, $\mathbb{P}(\text{gets hat back}) = \frac{1}{n}$, because there is one left and its just as likely to be yours as any

other hat after giving away $n-1$.
(Similar logic applies to the other positions in between as well). So actually, no, the probability that someone will get their original hat back does NOT depend on where they are in line. Each person gets their hat back with probability $\frac{1}{n}$.

(Another way to think of this is: the sample space of all ways to give n hats back has size $n!$. If we want person i to get their hat back, then there are $(n-1)!$ ways to do so, so the probability is $\dfrac{(n-1)!}{n!} = \dfrac{1}{n}$.)

Let's use linearity with indicator random variables! For $i = 1, ..., n$, let

$$X_i = \begin{cases} 1 & \text{if } i\text{-th person got their hat back} \\ 0 & \text{otherwise} \end{cases}.$$

Then the total number of people who get their hat back is $X = \sum_{i=1}^{n} X_i$. (Why?)

The expected value of each individual indicator random variable can be found as follows, since it can only take on the values 0 and 1:

$$\mathbb{E}[X_i] = 1 \cdot \mathbb{P}(X_i = 1) + 0 \cdot \mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = \mathbb{P}(i\text{th person got their hat back}) = \frac{1}{n}$$

From here, we will use linearity of expectation:

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[X_i] && \text{[linearity of expectation]} \\ &= \sum_{i=1}^{n} \frac{1}{n} \\ &= n \cdot \frac{1}{n} \\ &= 1 \end{aligned}$$

So, the expected number of people to get their hats back is 1 (doesn't even depend on n)! It is worth noting that these indicator random variables are *not* "independent" (we'll define this formally later). One of the reasons why is because if we know that a particular person did not get their own hat back, then the original owner of that hat will have a probability of 0 that they get that hat back. □

Theorem 3.3.1: Linearity of Expectation with Indicators

If asked only about the expectation of a random variable X (and not its PMF), then you may be able to write X as the sum of possibly dependent indicator random variables, and apply linearity of expectation. This technique is used when X is counting something (the number of people who get their hat back). Finding the PMF for this random variable is extremely complicated, and linearity makes computing the expectation easy (or at least easier than directly finding the PMF).

For an indicator random variable X_i,

$$\mathbb{E}[X_i] = 1 \cdot \mathbb{P}(X_i = 1) + 0 \cdot \mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1)$$

Example(s)

Suppose we flip a coin $n = 100$ times independently, where the probability of getting a head on each flip is $p = 0.23$. What is the expected number of heads we get? Before doing any computation, what do you think it might be?

Solution You might expect $np = 100 \cdot 0.23 = 23$ heads, and you would be absolutely correct! But we do need to prove/show this.
Let X be the number of heads total, so $\Omega_X = \{0, 1, 2, \ldots, 100\}$. The "normal" approach might be to try to find this PMF, which could be a bit complicated (we'll actually see this in the next section)! But let's try to use what we just learned instead, and define indicators.

For $i = 1, 2, \ldots, 100$, let $X_i = 1$ if the i-th flip is heads, and $X_i = 0$ otherwise. Then, $X = \sum_{i=1}^{100} X_i$ is the total number of heads (why?). To use linearity, we need to find $\mathbb{E}[X_i]$.
We showed earlier that

$$\mathbb{E}[X_i] = \mathbb{P}(X_i = 1) = p = 0.23$$

and so

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{100} X_i\right] && \text{[def of } X\text{]} \\
&= \sum_{i=1}^{100} \mathbb{E}[X_i] && \text{[linearity of expectation]} \\
&= \sum_{i=1}^{100} 0.23 \\
&= 100 \cdot 0.23 = 23
\end{aligned}$$

□

3.3.2 Variance

We've talked about the expectation (average/mean) of a random variable, and some approaches to computing this quantity. This provides a nice "summarization" of a random variable, as something we often want to know about it (sometimes even in place of its PMF). But we might want to know another summary quantity: how "variable" the random variable is, or how much it deviates from its mean. This is called the *variance* of a random variable, and we'll start with a motivating example below!

Consider the following two games. In both games we flip a fair coin. In Game 1, if a heads is flipped you pay me \$1, and if a tails is flipped I pay you \$1. In Game 2, if a heads is flipped you pay me \$1000, and if a tails is flipped I pay you \$1000.

Both games are fair, in the sense that the expected values of playing both games is 0.

$$\mathbb{E}[G_1] = -1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0 = -1000 \cdot \frac{1}{2} + 1000 \cdot \frac{1}{2} = \mathbb{E}[G_2]$$

Which game would you rather play? Maybe the adrenaline junkies among us would be willing to risk it all on Game 2, but I think most of us would feel better playing Game 1. As shown above, there is no difference in the expected value of playing these two games, so we need another metric to explain why Game 1 feels safer than Game 2.

We can measure this by calculating how far away a random variable is from its mean, on average. The quantity $X - \mathbb{E}[X]$ is the difference between a rv and its mean, but we want a distance, a positive value. So we will look at the squared difference $(X - \mathbb{E}[X])^2$ instead (another option would have been the absolute difference $|X - \mathbb{E}[X]|$, but someone chose the squared one instead). This is still a random variable (a nonnegative one, since it is squared), and so to get a number (the average distance from the mean), we take the *expectation* of this new rv, $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$. This is called the variance of the original random variable. The definition goes as follows:

Definition 3.3.1: Variance

The variance of a random variable X is defined to be

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

The variance is always nonnegative since we take the expectation of a nonnegative random variable $(X - \mathbb{E}[X])^2$. The first equality is the *definition* of variance, and the second equality is a more useful identity for doing computation that we show below.

Proof of Variance Identity.
Let $\mu = \mathbb{E}[X]$ as a shorthand. Then,

$$\begin{aligned}
\mathsf{Var}(X) &= \mathbb{E}\left[(X-\mu)^2\right] && \text{[def of variance]} \\
&= \mathbb{E}\left[X^2 - 2\mu X + \mu^2\right] && \text{[algebra]} \\
&= \mathbb{E}\left[X^2\right] - 2\mu\mathbb{E}[X] + \mu^2 && \text{[linearity of expectation]} \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 && [\mu = \mathbb{E}[X]] \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2
\end{aligned}$$

Notice that $\mathbb{E}\left[X^2\right] \neq \mathbb{E}[X]^2$ - this is a perfect time to point this out again. If these were equal, variance would always be zero and this would be a useless construct. □

The reason that someone chose the squared definition instead of the absolute value definition is because it has this nice splitting property $\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$ (the absolute value definition wouldn't have something this nice), and because the squaring function $g(t) = t^2$ is differentiable but the absolute value function $g(t) = |t|$ is not.

There is one problem though - if X is the height of someone in feet for example, then the average $\mathbb{E}[X]$ is also in units of feet, but the variance is in terms of square feet (since we square X). We'd like to say something like: the height of adults is generally 5.5 feet plus or minus 0.3 feet. To correct for this, we define the standard deviation to be the square root of the variance, which "undoes" the squaring.

Definition 3.3.2: Standard Deviation

Another measure of a random variable X's spread is the **standard deviation**, which is

$$\sigma_X = \sqrt{\mathsf{Var}(X)}$$

This measure is also useful, because the units of variance are squared in terms of the original variable X, and this essentially "undoes" our squaring, returning our units to the same as X.

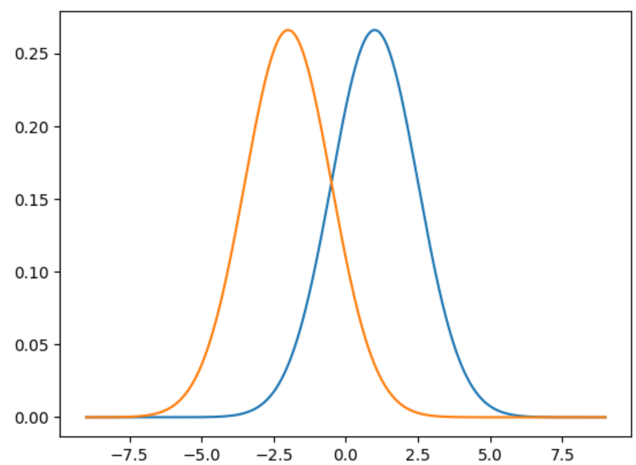
We had something nice happen for the random variable $aX+b$ when computing its expectation: $\mathbb{E}[aX+b] = a\mathbb{E}[X]+b$, called linearity of expectation. Is there a similar nice property for the variance as well?

Theorem 3.3.2: Property of Variance

We can also show that for any scalar $a, b \in \mathbb{R}$,

$$\mathsf{Var}(aX+b) = a^2\mathsf{Var}(X)$$

Before proving this, let's think about and try to understand why a came out squared, and what happened to the b. The reason a is squared is because variance involved squaring the random variable, so the a had to come out squared. It might not be a great intuitive reason, but we'll prove it below algebraically. The second (b disappearing) has a nice intuition behind it. Which of the two distributions (random variables) below do you think should have higher variance?



You might agree with me that they have the same variance! Why?

The idea behind variance is that it measures the "spread" of the values that a random variable can take on. The two graphs of random variables (distributions) above have the same "spread", but one is shifted slightly to the right. Since these graphs have the same "spread", we want their variance to reflect this similarity. Thus, shifting a random variable by some constant does not change the variance of that random variable. That is, $\mathsf{Var}(X+b) = \mathsf{Var}(X)$: that's why the b got lost!

Proof of Variance Property:. $\mathsf{Var}(aX+b) = a^2\mathsf{Var}(X)$
First, we show variance is unaffected by shifts; that is, $\mathsf{Var}(X+b) = \mathsf{Var}(X)$ for any scalar b. We use the original definition that $\mathsf{Var}(Y) = \mathbb{E}\left[(Y-\mathbb{E}[Y])^2\right]$, with $Y = X+b$.

$$\begin{aligned}
\mathsf{Var}(X+b) &= \mathbb{E}\left[((X+b)-\mathbb{E}[X+b])^2\right] && \text{[def of variance]}\\
&= \mathbb{E}\left[(X+b-\mathbb{E}[X]-b)^2\right] && \text{[linearity of expectation]}\\
&= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] \\
&= \mathsf{Var}(X) && \text{[def of variance]}
\end{aligned}$$

Then, we use this result to get the final one:

$$\begin{aligned}
\mathsf{Var}(aX+b) &= \mathsf{Var}(aX) && \text{[shifts don't matter]} \\
&= \mathbb{E}\left[(aX)^2\right] - (\mathbb{E}[aX])^2 && \text{[property of variance]} \\
&= \mathbb{E}\left[a^2X^2\right] - (a\mathbb{E}[X])^2 && \text{[linearity of expectation]} \\
&= a^2\mathbb{E}\left[X^2\right] - a^2\mathbb{E}[X]^2 && \text{[linearity of expectation]} \\
&= a^2(\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2) && \\
&= a^2\mathsf{Var}(X) && \text{[def of variance]}
\end{aligned}$$

□

Example(s)

Let X be the outcome of a fair 6-sided die roll. Recall that $\mathbb{E}[X] = 3.5$. What is $\mathsf{Var}(X)$?
Let's say you play a casino game, where you must pay $10 to roll this die once, but earn twice the value of the roll. What are the expected value and variance of your earnings?

Solution Recall that one of the equations for variance is

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

Computing the expected value of X we get

$$\mathbb{E}[X] = \sum_{x\in\Omega_X} xp_X(x) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + ... + 6\cdot\frac{1}{6} = 3.5$$

And using LOTUS, we can compute the expected value of X^2 to be

$$\mathbb{E}\left[X^2\right] = \sum_{x\in\Omega_X} x^2p_X(x) = 1^2\cdot\frac{1}{6} + 2^2\cdot\frac{1}{6} + ... + 6^2\cdot\frac{1}{6} = \frac{91}{6}$$

Putting these together with our definition above gives

$$\mathsf{Var}(X) = \frac{91}{6} - (3.5)^2 = \frac{35}{12}$$

Now, let Y denote our earnings; then $Y = 2X - 10$. So by linearity of expectation,

$$\mathbb{E}[2X-10] = 2\mathbb{E}[X] - 10 = 2\cdot 3.5 - 10 = -3$$

By the property of variance, we get

$$\mathsf{Var}(2X-10) = 2^2\mathsf{Var}(X) = 2^2\cdot\frac{35}{12} = \frac{35}{3}$$

□

Now you might wonder, what about the variance of a sum $\mathsf{Var}(X+Y)$? You might hope that $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$, but this unfortunately is only true when the random variables are independent (we'll define this in the next section, but you can kind of guess what it means)! It is so important to remember that we made no independence assumptions for linearity of expectation - it's always true!

3.3.3 Exercises

1. Suppose you studied hard for a 100-question multiple-choice exam (with 4 choices per question) so that you believe you know the answer to about 80% of the questions, and you guess the answer to the remaining 20%. What is the expected number of questions you answer correctly?

   Solution: For $i = 1, \dots, 100$, let X_i be the indicator rv which is 1 if you got the i^{th} question correct, and 0 otherwise. Then, the total number of questions correct is $X = \sum_{i=1}^{100} X_i$. To compute $\mathbb{E}[X]$ we need $\mathbb{E}[X_i]$ for each $i = 1, \dots, 100$.

   $\mathbb{E}[X_i] = 1 \cdot \mathbb{P}(X_i = 1) + 0 \cdot \mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = \mathbb{P}(\text{correct on question } i) = 1 \cdot 0.8 + 0.25 \cdot 0.2 = 0.85$

   where the second last step was using the law of total probability, conditioning on whether we know the answer to a question or not. Hence,

   $$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{100} X_i\right] = \sum_{i=1}^{100} \mathbb{E}[X_i] = \sum_{i=1}^{100} 0.85 = 85$$

   This kind of makes sense - I should be guaranteed 80 out of 100, and if I guess on the other 20, I would get about 5 (a quarter of them) right, for a total of 85.

2. Recall exercise 2 from 3.1, where we had a random variable X with PMF

   $$p_X(k) = \begin{cases} 1/100 & k = -2 \\ 18/100 & k = 0 \\ 81/100 & k = 2 \end{cases}$$

   The expectation was

   $$\mathbb{E}[X] = \sum_{k \in \Omega_X} k \cdot p_X(k) = -2 \cdot \frac{1}{100} + 0 \cdot \frac{18}{100} + 2 \cdot \frac{81}{100} = 1.6$$

   Compute $\mathsf{Var}(X)$.

   Solution: Since $\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$, we need to use LOTUS to compute the first part.

   $$\mathbb{E}\left[X^2\right] = \sum_{k \in \Omega_X} k^2 \cdot p_X(k) = (-2)^2 \cdot \frac{1}{100} + 0^2 \cdot \frac{18}{100} + 2^2 \cdot \frac{81}{100} = 3.28$$

   Hence,

   $$\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = 3.28 - 1.6^2 = 0.72$$