CS 70 Discrete Mathematics and Probability Theory
Fall 2018 Alistair Sinclair and Yun Song Note 16

Random Variables: Variance and Independence

We have seen in the previous note that if we take a biased coin that shows heads with probability p and toss it n times, then the expected number of heads is np. What this means is that if we repeat the experiment multiple times, where in each experiment we toss the coin n times, then on average we get np heads. But in any single experiment, the number of heads observed can be any value between 0 and n. What can we say about how far off we are from the expected value? That is, what is the typical deviation of the number of heads from np?

1 Random Walk

Let us consider a simpler setting that is equivalent to tossing a fair coin n times, but is more amenable to analysis. Suppose we have a particle that starts at position 0 and performs a random walk in one dimension. At each time step, the particle moves either one step to the right or one step to the left with equal probability (this kind of random walk is called *symmetric*), and the move at each time step is independent of all other moves. We think of these random moves as taking place according to whether a fair coin comes up heads or tails. The expected position of the particle after n moves is back at 0, but how far from 0 should we typically expect the particle to end up?

Denoting a right-move by $+1$ and a left-move by -1, we can describe the probability space here as the set of all sequences of length n over the alphabet $\{\pm 1\}$, each having equal probability $\frac{1}{2^n}$. Let the r.v. S_n denote the position of the particle (relative to our starting point 0) after n moves. Thus, we can write

$$S_n = X_1 + X_2 + \cdots + X_n, \tag{1}$$

where $X_i = +1$ if the i-th move is to the right and $X_i = -1$ if the move is to the left.

The expectation of S_n can be easily computed as follows. Since $\mathbb{E}[X_i] = (\frac{1}{2} \times 1) + (\frac{1}{2} \times (-1)) = 0$, applying linearity of expectation immediately gives $\mathbb{E}[S_n] = \sum_{i=1}^{n} \mathbb{E}[X_i] = 0$. But of course this is not very informative, and is due to the fact that positive and negative deviations from 0 cancel out.

What we are really asking is: What is the expected value of $|S_n|$, the *distance* of the particle from 0? Rather than consider the r.v. $|S_n|$, which is a little difficult to work with due to the absolute value operator, we will instead look at the r.v. S_n^2. Notice that this also has the effect of making all deviations from 0 positive, so it should also give a good measure of the distance from 0. However, because it is the *squared* distance, we will need to take a square root at the end.

We will now show that the expected square distance after n steps is equal to n:

Claim 16.1. *For the random variable S_n defined in* (1)*, we have $\mathbb{E}\left[S_n^2\right] = n$.*

Proof. We use the expression (1) and expand the square:

$$\mathbb{E}\left[S_n^2\right] = \mathbb{E}\left[(X_1 + X_2 + \cdots + X_n)^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_i X_j\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] + 2\sum_{i<j} \mathbb{E}[X_i X_j]. \tag{2}$$

In the last equality we have used linearity of expectation. To proceed, we need to compute $\mathbb{E}\left[X_i^2\right]$ and $\mathbb{E}[X_iX_j]$ for $i \neq j$. Since X_i can take on only values ± 1, clearly $X_i^2 = 1$ always, so $\mathbb{E}\left[X_i^2\right] = 1$. To compute $\mathbb{E}[X_iX_j]$ for $i \neq j$, note $X_iX_j = +1$ when $X_i = X_j = +1$ or $X_i = X_j = -1$, and otherwise $X_iX_j = -1$. Therefore,

$$\begin{aligned}
\mathbb{P}[X_iX_j = 1] &= \mathbb{P}[(X_i = X_j = +1) \vee (X_i = X_j = -1)] \\
&= \mathbb{P}[X_i = X_j = +1] + \mathbb{P}[X_i = X_j = -1] \\
&= \mathbb{P}[X_i = +1] \times \mathbb{P}[X_j = +1] + \mathbb{P}[X_i = -1] \times \mathbb{P}[X_j = -1] \\
&= \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2},
\end{aligned}$$

where the second equality follows from the fact that the events $X_i = X_j = +1$ and $X_i = X_j = -1$ are mutually exclusive, while the third equality follows from the independence of the events $X_i = +1$ and $X_j = +1$, and likewise for the events $X_i = -1$ and $X_j = -1$. In a similar vein, one obtains $\mathbb{P}[X_iX_j = -1] = \frac{1}{2}$, and hence $\mathbb{E}[X_iX_j] = 0$.

Finally, plugging $\mathbb{E}\left[X_i^2\right] = 1$ and $\mathbb{E}[X_iX_j] = 0$, for $i \neq j$, into (2) gives $\mathbb{E}\left[S_n^2\right] = \sum_{i=1}^n 1 + 2\sum_{i<j} 0 = n$, as desired. □

So, for the symmetric random walk example, we see that the expected squared distance from 0 is n. One interpretation of this is that we might expect to be a distance of about $\sqrt{n}$ away from 0 after n steps. However, we have to be careful here: we **cannot** simply argue that $\mathbb{E}[|S_n|] = \sqrt{\mathbb{E}[S_n^2]} = \sqrt{n}$. (Why not?) We will see later in the course how to make precise deductions about $|S_n|$ from knowledge of $\mathbb{E}\left[S_n^2\right]$. For the moment, however, let us agree to view $\mathbb{E}\left[S_n^2\right]$ as an intuitive measure of "spread" of the r.v. S_n.

For a more general r.v. X with expectation $\mathbb{E}[X] = \mu$, what we are really interested in is $\mathbb{E}\left[(X-\mu)^2\right]$, the expected squared distance *from the mean*. In our symmetric random walk example, we had $\mu = 0$, so $\mathbb{E}\left[(X-\mu)^2\right]$ just reduced to $\mathbb{E}\left[X^2\right]$.

Definition 16.1 (Variance)**.** *For a r.v. X with expectation $\mathbb{E}[X] = \mu$, the variance of X is defined to be*

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

The square root $\sigma(X) := \sqrt{\mathrm{Var}(X)}$ is called the standard deviation of X.

The point of taking the square root of variance is to put the standard deviation "on the same scale" as the r.v. itself. Since the variance and standard deviation differ just by a square, it really doesn't matter which one we choose to work with as we can always compute one from the other. We shall usually use the variance. For the random walk example above, we have that $\mathrm{Var}(S_n) = n$, and the standard deviation $\sigma(S_n)$ of X is $\sqrt{n}$.

The following observation provides a slightly different way to compute the variance, which sometimes turns out to be simpler.

Theorem 16.1. *For a r.v. X with expectation $\mathbb{E}[X] = \mu$, we have $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mu^2$.*

Proof. From the definition of variance, we have

$$\mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left[X^2 - 2\mu X + \mu^2\right] = \mathbb{E}\left[X^2\right] - 2\mu\mathbb{E}[X] + \mu^2 = \mathbb{E}\left[X^2\right] - \mu^2.$$

In the third equality, we used linearity of expectation. We also used the fact that since $\mu = \mathbb{E}[X]$ is a constant, $\mathbb{E}[\mu X] = \mu\mathbb{E}[X] = \mu^2$ and $\mathbb{E}\left[\mu^2\right] = \mu^2$. □

Another important property that will come in handy is the following: For any random variable X and constant c, we have

$$\mathrm{Var}(cX) = c^2\mathrm{Var}(X).$$

The proof is simple and left as an exercise.

2 Variance Computation

Let us see some examples of variance calculations.

1. **Fair die.** Let X be the score on the roll of a single fair die. Recall from the previous note that $\mathbb{E}[X] = \frac{7}{2}$. So we just need to compute $\mathbb{E}\left[X^2\right]$, which is a routine calculation:

$$\mathbb{E}\left[X^2\right] = \frac{1}{6}\left(1^2+2^2+3^2+4^2+5^2+6^2\right) = \frac{91}{6}.$$

   Thus, from Theorem 16.1,

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.$$

2. **Uniform distribution.** More generally, if X is a uniform random variable on the set $\{1,\ldots,n\}$, so X takes on values $1,\ldots,n$ with equal probability $\frac{1}{n}$, then the mean, variance and standard deviation of X are given by:

$$\mathbb{E}[X] = \frac{n+1}{2}, \qquad \mathrm{Var}(X) = \frac{n^2-1}{12}, \qquad \sigma(X) = \sqrt{\frac{n^2-1}{12}}. \tag{3}$$

   You should verify these as an exercise.

3. **Fixed points of permutations.** Let X_n be the number of fixed points in a random permutation of n items (i.e., in the homework permutation example, X_n is the number of students in a class of size n who receive their own homework after shuffling). We saw in the previous note that $\mathbb{E}[X_n] = 1$, regardless of n. To compute $\mathbb{E}\left[X_n^2\right]$, write $X_n = I_1 + I_2 + \cdots + I_n$, where $I_i = 1$ if i is a fixed point, and $I_i = 0$ otherwise. Then as usual we have

$$\mathbb{E}\left[X_n^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[I_i^2\right] + 2\sum_{i<j} \mathbb{E}[I_i I_j]. \tag{4}$$

   Since I_i is an indicator r.v., we have that $\mathbb{E}\left[I_i^2\right] = \mathbb{P}[I_i = 1] = \frac{1}{n}$. For $i < j$, since both I_i and I_j are indicators, we can compute $\mathbb{E}[I_i I_j]$ as follows:

$$\mathbb{E}[I_i I_j] = \mathbb{P}[I_i I_j = 1] = \mathbb{P}[I_i = 1 \wedge I_j = 1] = \mathbb{P}[\text{both } i \text{ and } j \text{ are fixed points}] = \frac{1}{n(n-1)}.$$

   Make sure that you understand the last step here. Plugging this into equation (4) we get

$$\mathbb{E}\left[X_n^2\right] = \sum_{i=1}^{n} \frac{1}{n} + 2\sum_{i<j} \frac{1}{n(n-1)} = \left(n \times \frac{1}{n}\right) + \left[2\binom{n}{2} \times \frac{1}{n(n-1)}\right] = 1 + 1 = 2.$$

   Thus, $\mathrm{Var}(X_n) = \mathbb{E}\left[X_n^2\right] - (\mathbb{E}[X_n])^2 = 2 - 1 = 1$. That is, the variance and the mean are both equal to 1. Like the mean, the variance is also independent of n. Intuitively at least, this means that it is unlikely that there will be more than a small number of fixed points even when the number of items, n, is very large.

3 Multiple Random Variables

Often one is interested in multiple random variables on the same sample space. Consider, for example, the sample space of flipping two coins. One could define many random variables: for example a random variable X indicating the number of heads in a sequence of coin tosses, or a random variable Y indicating the number of tails, or a random variable Z indicating whether the first is H or not. Note that for each sample point, any random variable has a specific value: e.g., for $\omega = HTT$, we have $X(\omega) = 1$, $Y(\omega) = 2$, and $Z(\omega) = 1$.

The concept of a distribution can then be extended to probabilities for the combination of values for multiple random variables.

Definition 16.2. *The joint distribution for two discrete random variables X and Y is the collection of values $\{((a,b), \mathbb{P}[X = a, Y = b]) : a \in \mathscr{A}, b \in \mathscr{B}\}$, where $\mathscr{A}$ is the set of all possible values taken by X and $\mathscr{B}$ is the set of all possible values taken by Y.*

When given a joint distribution for X and Y, the distribution $\mathbb{P}[X = a]$ for X is called the *marginal distribution* for X, and can be found by "summing" over the values of Y. That is,

$$\mathbb{P}[X = a] = \sum_{b \in \mathscr{B}} \mathbb{P}[X = a, Y = b].$$

The marginal distribution for Y is analogous, as is the notion of a joint distribution for any number of random variables.

A joint distribution over random variables $X_1, \ldots, X_n$ (for example, X_i could be the value of the ith roll of a sequence of n die rolls) is $\mathbb{P}[X_1 = a_1, \ldots, X_n = a_n]$, where $a_i \in \mathscr{A}_i$ and $\mathscr{A}_i$ is the set of possible values for X_i. The marginal distribution for X_i is simply the distribution for X_i and can be obtained by summing over all the possible values of the other variables, but in some cases can be derived more simply. We proceed to one such case.

4 Independence of Random Variables

Independence for random variables is defined in an analogous fashion to independence for events:

Definition 16.3 (Independence)**.** *Random variables X and Y on the same probability space are said to be* independent *if the events $X = a$ and $Y = b$ are independent for all values a, b. Equivalently, the joint distribution of independent r.v.'s decomposes as*

$$\mathbb{P}[X = a, Y = b] = \mathbb{P}[X = a]\mathbb{P}[Y = b], \quad \forall a, b.$$

Mutual independence of more than two r.v.'s is defined similarly.

A very important example of independent random variables are indicator random variables for independent events. If I_i denotes the indicator r.v. for the i-th toss of a coin being H, then $I_1, \ldots, I_n$ are mutually independent random variables. This example motivates the commonly used phrase "*independent and identically distributed (i.i.d.)* set of random variables." In this example, $\{I_1, \ldots, I_n\}$ is a set of i.i.d. indicator random variables.

One of the most important and useful facts about variance is that if a random variable X is the sum of *independent* random variables $X = X_1 + \cdots + X_n$, then its variance is the sum of the variances of the individual

r.v.'s. In particular, if the individual r.v.'s X_i are identically distributed (i.e., they have the same distribution), then $\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i) = n \cdot \text{Var}(X_1)$. This means that the standard deviation is $\sigma(X) = \sqrt{n} \cdot \sigma(X_1)$. Note that by contrast, the expected value is $\mathbb{E}[X] = n \cdot \mathbb{E}[X_1]$. Intuitively this means that whereas the average value of X grows proportionally to n, the spread of the distribution grows proportionally to $\sqrt{n}$, which is much smaller than n. In other words the distribution of X tends to concentrate around its mean.

Let us now formalize these ideas. First, we have the following result which states that the expected value of the product of two independent random variables is equal to the product of their expected values.

Theorem 16.2. *For* independent *random variables X, Y, we have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

Proof. We have

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_a \sum_b ab \times \mathbb{P}[X = a, Y = b] \\
&= \sum_a \sum_b ab \times \mathbb{P}[X = a] \times \mathbb{P}[Y = b] \\
&= \left(\sum_a a \times \mathbb{P}[X = a] \right) \times \left(\sum_b b \times \mathbb{P}[Y = b] \right) \\
&= \mathbb{E}[X] \times \mathbb{E}[Y],
\end{aligned}$$

as required. In the second line here we made crucial use of independence. □

We now use the above theorem to conclude the nice property that the variance of the sum of independent random variables is equal to the sum of their variances.

Theorem 16.3. *For* independent *random variables X, Y, we have*

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

Proof. From the alternative formula for variance in Theorem 16.1 and linearity of expectation, we have

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

Since X, Y are independent, Theorem 16.2 implies that the final term in this expression is zero. □

It is very important to remember that **neither** of the above two results is true in general when X, Y are not independent. As a simple example, note that even for a $\{0, 1\}$-valued r.v. X with $\mathbb{P}[X = 1] = p$, $\mathbb{E}\left[X^2\right] = p$ is not equal to $\mathbb{E}[X]^2 = p^2$ (because of course X and X are not independent!). This is in contrast to linearity of expectation, where we saw that the expectation of a sum of r.v.'s is the sum of the expectations of the individual r.v.'s, regardless of whether or not the r.v.'s are independent.

Example

Let us return to our motivating example of a sequence of n coin tosses. Let X_n denote the number of Heads in n tosses of a biased coin with Heads probability p (i.e., $X_n \sim \text{Binomial}(n,p)$). As usual, we write $X_n = I_1 + I_2 + \cdots + I_n$, where $I_i = 1$ if the i-th toss is H, and $I_i = 0$ otherwise.

We already know $\mathbb{E}[X_n] = \sum_{i=1}^n \mathbb{E}[I_i] = np$. We can compute $\text{Var}(I_i) = \mathbb{E}\left[I_i^2\right] - \mathbb{E}[I_i]^2 = p - p^2 = p(1-p)$. Since the I_i's are independent, by Theorem 16.3 we get $\text{Var}(X_n) = \sum_{i=1}^n \text{Var}(I_i) = np(1-p)$.

As an example, for a fair coin ($p = \frac{1}{2}$) the expected number of Heads in n tosses is $\frac{n}{2}$, and the standard deviation is $\sqrt{\frac{n}{4}} = \frac{\sqrt{n}}{2}$. Note that since the maximum number of Heads is n, the standard deviation is much less than this maximum number for large n. This is in contrast to the previous example of the uniformly distributed random variable (3), where the standard deviation $\sigma(X) = \sqrt{\frac{n^2-1}{12}} \approx \frac{n}{\sqrt{12}}$ (for large n) is of the same order as the largest value, n. In this sense, the spread of a binomially distributed r.v. is much smaller than that of a uniformly distributed r.v.

5 Covariance and Correlation

The expression $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ in the proof of Theorem 16.3 is a measure of association between X, Y, and is called the *covariance*:

Definition 16.4 (Covariance)**.** *The* covariance *of random variables X and Y, denoted* $\text{Cov}(X,Y)$*, is defined as*

$$\text{Cov}(X,Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],$$

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.

Remarks. We note some important facts about covariance.

1. If X, Y are independent, then $\text{Cov}(X,Y) = 0$. However, the converse is **not** true.
2. $\text{Cov}(X,X) = \text{Var}(X)$.
3. Covariance is *bilinear*; i.e., for any collection of random variables $\{X_1, \ldots, X_n\}, \{Y_1, \ldots, Y_m\}$ and fixed constants $\{a_1 \ldots, a_n\}, \{b_1, \ldots, b_m\}$,

$$\text{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(X_i, Y_j).$$

For general random variables X and Y,

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y).$$

While the sign of $\text{Cov}(X,Y)$ is informative of how X and Y are associated, its magnitude is difficult to interpret. A statistic that is easier to interpret is *correlation*:

Definition 16.5 (Correlation)**.** *Suppose X and Y are random variables with $\sigma(X) > 0$ and $\sigma(Y) > 0$. Then, the correlation of X and Y is defined as*

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

Correlation is more useful than covariance because the former always ranges between -1 and $+1$, as the following theorem shows:

Theorem 16.4. *For any pair of random variables X and Y with $\sigma(X) > 0$ and $\sigma(Y) > 0$,*

$$-1 \leq \text{Corr}(X,Y) \leq +1.$$

Proof. Let $\mathbb{E}[X] = \mu_X$ and $\mathbb{E}[Y] = \mu_Y$, and define $\widetilde{X} = (X - \mu_X)/\sigma(X)$ and $\widetilde{Y} = (Y - \mu_Y)/\sigma(Y)$. Then, $\mathbb{E}[\widetilde{X}^2] = \mathbb{E}[\widetilde{Y}^2] = 1$, so

$$\begin{aligned} 0 \leq \mathbb{E}[(\widetilde{X} - \widetilde{Y})^2] &= \mathbb{E}[\widetilde{X}^2] + \mathbb{E}[\widetilde{Y}^2] - 2\mathbb{E}[\widetilde{X}\widetilde{Y}] = 2 - 2\mathbb{E}[\widetilde{X}\widetilde{Y}] \\ 0 \leq \mathbb{E}[(\widetilde{X} + \widetilde{Y})^2] &= \mathbb{E}[\widetilde{X}^2] + \mathbb{E}[\widetilde{Y}^2] + 2\mathbb{E}[\widetilde{X}\widetilde{Y}] = 2 + 2\mathbb{E}[\widetilde{X}\widetilde{Y}], \end{aligned}$$

which implies $-1 \leq \mathbb{E}[\widetilde{X}\widetilde{Y}] \leq +1$. Now, noting that $\mathbb{E}[\widetilde{X}] = \mathbb{E}[\widetilde{Y}] = 0$, we obtain $\text{Corr}(X,Y) = \text{Cov}(\widetilde{X},\widetilde{Y}) = \mathbb{E}[\widetilde{X}\widetilde{Y}]$. Hence, $-1 \leq \text{Corr}(X,Y) \leq +1$. □

Note that the above proof shows that $\text{Corr}(X,Y) = +1$ if and only if $\mathbb{E}[(\widetilde{X} - \widetilde{Y})^2] = 0$, which implies $\widetilde{X} = \widetilde{Y}$ with probability 1. Similarly, $\text{Corr}(X,Y) = -1$ if and only if $\mathbb{E}[(\widetilde{X} + \widetilde{Y})^2] = 0$, which implies $\widetilde{X} = -\widetilde{Y}$ with probability 1. In terms of the original random variables X, Y, this means the following: if $\text{Corr}(X,Y) = \pm 1$, then there exist constants a and b such that, with probability 1,

$$Y = aX + b,$$

where $a > 0$ if $\text{Corr}(X,Y) = +1$ and $a < 0$ if $\text{Corr}(X,Y) = -1$.