CS 70 Discrete Mathematics and Probability Theory
Fall 2013 Vazirani Lecture 13

Variance

Question: Let us return once again to the question of how many heads in a typical sequence of n coin flips. Recall that we used the Galton board to visualize this as follows: consider a token falling through the Galton board, deflected randomly to the right or left at each step. We think of these random deflections as taking place according to whether a fair coin comes up heads or tails. The expected position of the token when it reaches the bottom is right in the middle, but how far from the middle should we typically expect it to land?

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

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

where $X_i = \begin{cases} +1 & \text{if } i\text{th toss is Heads;} \\ -1 & \text{otherwise.} \end{cases}$

Now obviously we have $\mathrm{E}(X) = 0$. The easiest way to see this is to note that $\mathrm{E}(X_i) = (\frac{1}{2} \times 1) + (\frac{1}{2} \times (-1)) = 0$, so by linearity of expectation $\mathrm{E}(X) = \sum_{i=1}^n \mathrm{E}(X_i) = 0$. But of course this is not very informative, and is due to the fact that positive and negative deviations from 0 cancel out.

What the above question is really asking is: What is the expected value of $|X|$, the distance of the *distance* from 0? Rather than consider the r.v. $|X|$, which is a little awkward due to the absolute value operator, we will instead look at the r.v. X^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 traveled. However, because it is the *squared* distance, we will need to take a square root at the end.

Let's calculate $\mathrm{E}(X^2)$:

$$\begin{aligned} \mathrm{E}(X^2) &= \mathrm{E}((X_1 + X_2 + \cdots + X_n)^2) \\ &= \mathrm{E}(\textstyle\sum_{i=1}^n X_i^2 + \sum_{i \neq j} X_i X_j) \\ &= \textstyle\sum_{i=1}^n \mathrm{E}(X_i^2) + \sum_{i \neq j} \mathrm{E}(X_i X_j) \end{aligned}$$

In the last line here, we used linearity of expectation. To proceed, we need to compute $\mathrm{E}(X_i^2)$ and $\mathrm{E}(X_i X_j)$ (for $i \neq j$). Let's consider first X_i^2. Since X_i can take on only values ± 1, clearly $X_i^2 = 1$ always, so $\mathrm{E}(X_i^2) = 1$. What about $\mathrm{E}(X_i X_j)$? Well, $X_i X_j = +1$ when $X_i = X_j = +1$ or $X_i = X_j = -1$, and otherwise $X_i X_j = -1$. Also,

$$\Pr[(X_i = X_j = +1) \vee (X_i = X_j = -1)] = \Pr[X_i = X_j = +1] + \Pr[X_i = X_j = -1] = \frac{1}{4} + \frac{1}{4} = \frac{1}{2},$$

so $X_i X_j = 1$ with probability $\frac{1}{2}$. In the above calculation we used the fact that the events $X_i = +1$ and $X_j = +1$ are independent, so $\Pr[X_i = X_j = +1] = \Pr[X_i = +1] \times \Pr[X_j = +1] = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ (and similarly for $\Pr[X_i = X_j = -1]$). Therefore $X_i X_j = -1$ with probability $\frac{1}{2}$ also. Hence $\mathrm{E}(X_i X_j) = 0$.

Plugging these values into the above equation gives

$$E(X^2) = (n \times 1) + 0 = n.$$

So we see that our 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 $E(|X|) = \sqrt{E(X^2)} = \sqrt{n}$. (Why not?) We will see later in the lecture how to make precise deductions about $|X|$ from knowledge of $E(X^2)$.

For the moment, however, let's agree to view $E(X^2)$ as an intuitive measure of "spread" of the r.v. X. In fact, for a more general r.v. with expectation $E(X) = \mu$, what we are really interested in is $E((X-\mu)^2)$, the expected squared distance *from the mean*. In our random walk example, we had $\mu = 0$, so $E((X-\mu)^2)$ just reduces to $E(X^2)$.

Definition 13.1 (variance): For a r.v. X with expectation $E(X) = \mu$, the variance of X is defined to be

$$\mathrm{Var}(X) = E((X-\mu)^2).$$

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

The point of the standard deviation is merely to "undo" the squaring in the variance. Thus the standard deviation is "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 immediately. We shall usually use the variance. For the random walk example above, we have that $\mathrm{Var}(X) = n$, and the standard deviation of X, $\sigma(X)$, is $\sqrt{n}$.

The following easy observation gives us a slightly different way to compute the variance that is simpler in many cases.

Theorem 13.1: For a r.v. X with expectation $E(X) = \mu$, we have $\mathrm{Var}(X) = E(X^2) - \mu^2$.

Proof: From the definition of variance, we have

$$\mathrm{Var}(X) = E((X-\mu)^2) = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2.$$

In the third step here, we used linearity of expectation. □

Examples

Let's see some examples of variance calculations.

1. **Fair die.** Let X be the score on the roll of a single fair die. Recall from an earlier lecture that $E(X) = \frac{7}{2}$. So we just need to compute $E(X^2)$, which is a routine calculation:

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

   Thus from Theorem 16.1

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

   More generally, if X is a random variable that takes on values $1, \ldots, n$ with equal probability $1/n$ (i.e. X has a uniform distribution), the mean, variance and standard deviation of X are:

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

(You should verify these.)

2. **Number of fixed points.** Let X be the number of fixed points in a random permutation of n items (i.e., the number of students in a class of size n who receive their own homework after shuffling). We saw in an earlier lecture that $\mathrm{E}(X) = 1$ (regardless of n). To compute $\mathrm{E}(X^2)$, write $X = X_1 + X_2 + \cdots + X_n$, where $X_i = \begin{cases} 1 & \text{if } i \text{ is a fixed point;} \\ 0 & \text{otherwise} \end{cases}$
Then as usual we have

$$\mathrm{E}(X^2) = \sum_{i=1}^{n} \mathrm{E}(X_i^2) + \sum_{i \neq j} \mathrm{E}(X_i X_j). \tag{1}$$

Since X_i is an indicator r.v., we have that $\mathrm{E}(X_i^2) = \Pr[X_i = 1] = \frac{1}{n}$. Since both X_i and X_j are indicators, we can compute $\mathrm{E}(X_i X_j)$ as follows:

$$\mathrm{E}(X_i X_j) = \Pr[X_i = 1 \wedge X_j = 1] = \Pr[\text{both } i \text{ and } j \text{ are fixed points}] = \frac{1}{n(n-1)}.$$

[Check that you understand the last step here.] Plugging this into equation (1) we get

$$\mathrm{E}(X^2) = (n \times \tfrac{1}{n}) + (n(n-1) \times \tfrac{1}{n(n-1)}) = 1 + 1 = 2.$$

Thus $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = 2 - 1 = 1$. I.e., 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.

Independent Random Variables

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

Definition 13.2 (independent r.v.'s): 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

$$\Pr[X = a, Y = b] = \Pr[X = a]\Pr[Y = b] \quad \forall a, b.$$

Mutual independence of more than two r.v.'s is defined similarly. A very important example of independent r.v.'s is indicator r.v.'s for independent events. Thus, for example, if $\{X_i\}$ are indicator r.v.'s for the ith toss of a coin being Heads, then the X_i are mutually independent r.v.'s.

One of the most important and useful facts about variance is 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, then $\mathrm{Var}(X) = \sum_i \mathrm{Var}(X_i) = n \cdot varX_1$. This means that the standard deviation $\sigma(X) = \sqrt(n)\sigma(X_1)$. Note that by contrast, the expected value $E[X] = n \cdot E[X_1]$. Intuitively this means that whereas the average value of X grows proportionally to n, the spread of the distribution grows proportionally to $\sqrt{n}$. In other words the distribution of X tends to concentrate around its mean. Let us formalize these ideas:

Theorem 13.2: For any random variable X and constant c, we have

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

And for *independent* random variables X,Y, we have

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

Theorem 13.3: For *independent* random variables X,Y, we have $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$.

Proof: We have

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

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

We now use the above theorem to conclude the nice property of the variance of independent random variables stated in the theorem above, namely that for independent random variables X and Y, $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$:

Proof: From the alternative formula for variance in Theorem 16.1, we have, using linearity of expectation extensively,

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{E}((X+Y)^2) - \mathrm{E}(X+Y)^2 \\
&= \mathrm{E}(X^2) + \mathrm{E}(Y^2) + 2\mathrm{E}(XY) - (\mathrm{E}(X) + \mathrm{E}(Y))^2 \\
&= (\mathrm{E}(X^2) - \mathrm{E}(X)^2) + (\mathrm{E}(Y^2) - \mathrm{E}(Y)^2) + 2(\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)).
\end{aligned}$$

Now *because X,Y are independent*, by Theorem 18.1 the final term in this expression is zero. Hence we get our result. □

Note: The expression $\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$ appearing in the above proof is called the *covariance* of X and Y, and is a measure of the dependence between X,Y. It is zero when X,Y are independent.

It is very important to remember that **neither** of these two results is true in general, without the assumption that X,Y are independent. As a simple example, note that even for a 0-1 r.v. X with $\Pr[X=1] = p$, $\mathrm{E}(X^2) = p$ is not equal to $\mathrm{E}(X)^2 = p^2$ (because of course X and X are not independent!). This is in contrast to the case of the 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 *always*.

Example

Let's return to our motivating example of a sequence of n coin tosses. Let X the the number of Heads in n tosses of a biased coin with Heads probability p (i.e., X has the binomial distribution with parameters n,p). We already know that $\mathrm{E}(X) = np$. As usual, let $X = X_1 + X_2 + \cdots + X_n$, where $X_i = \begin{cases} 1 & \text{if } i\text{th toss is Head;} \\ 0 & \text{otherwise} \end{cases}$.

We can compute $\mathrm{Var}(X_i) = E(X_i^2) - E(X_i)^2 = p - p^2 = p(1-p)$. So $\mathrm{Var}(X) = np(1-p)$.

As an example, for a fair coin the expected number of Heads in n tosses is $\frac{n}{2}$, and the standard deviation is $\frac{\sqrt{n}}{2}$. 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, where the standard deviation

$$\sigma(X) = \sqrt{(n^2-1)/12} \approx n/\sqrt{12}$$

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.