CS 70	Discrete Mathematics and Probability Theory
Fall 2012	Vazirani Lecture 16

Variance

Question: At each time step, I flip a fair coin. If it comes up Heads, I walk one step to the right; if it comes up Tails, I walk one step to the left. How far do I expect to have traveled from my starting point after n steps?

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 $\mathrm{E}(X) = \sum_{i=1}^n \mathrm{E}(X_i) = 0$. Thus after n steps, my expected position is 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 $|X|$, our *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

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

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

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

Definition 16.1 (variance): For a r.v. X with expectation $\mathrm{E}(X) = \mu$, the variance of X is defined to be

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

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

Proof: From the definition of variance, we have

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

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

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 $\mathrm{E}(X) = \frac{7}{2}$. So we just need to compute $\mathrm{E}(X^2)$, which is a routine calculation:

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

   Thus from Theorem 16.1

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

   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. **Biased coin.** 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 $\mathrm{E}(X) = np$. Writing as usual $X = X_1 + X_2 + \cdots + X_n$, where $X_i = \begin{cases} 1 & \text{if } i\text{th toss is Head;} \\ 0 & \text{otherwise} \end{cases}$ we have

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

In the third line here, we have used the facts that $\mathrm{E}(X_i^2) = p$, and that

$$\mathrm{E}(X_i X_j) = \Pr[X_i = X_j = 1] = \Pr[X_i = 1] \cdot \Pr[X_j = 1] = p^2,$$

(since $X_i = 1$ and $X_j = 1$ are independent events). Note that there are $n(n-1)$ pairs i, j with $i \neq j$.

Finally, we get that $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = np(1-p)$. Notice that in fact $\mathrm{Var}(X) = \sum_i \mathrm{Var}(X_i)$, and the same was true in the random walk example. This is in fact no coincidence. We will explore for what kinds of random variables this is true later in the next lecture.

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.

3. **Poisson distribution.** What is the variance of a Poisson r.v. X?

$$\mathrm{E}(X^2) = \sum_{i=0}^{\infty} i^2 \mathrm{e}^{-\lambda} \frac{\lambda^i}{i!} = \lambda \sum_{i=1}^{\infty} i\mathrm{e}^{-\lambda} \frac{\lambda^{i-1}}{(i-1)!} = \lambda \left(\sum_{i=1}^{\infty} (i-1) \mathrm{e}^{-\lambda} \frac{\lambda^{i-1}}{(i-1)!} + \sum_{i=1}^{\infty} \mathrm{e}^{-\lambda} \frac{\lambda^{i-1}}{(i-1)!} \right) = \lambda(\lambda + 1).$$

[Check you follow each of these steps. In the last step, we have noted that the two sums are respectively $\mathrm{E}(X)$ and $\sum_i \Pr[X = i] = 1$.]

Finally, we get $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \lambda$. So, for a Poisson random variable, the expectation and variance are equal.

4. **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 $\mathrm{E}(X_iX_j)$ as follows:

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

[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.

Chebyshev's Inequality

We have seen that, intuitively, the variance (or, more correctly the standard deviation) is a measure of "spread", or deviation from the mean. Our next goal is to make this intuition quantitatively precise. What we can show is the following:

Theorem 16.2: **[Chebyshev's Inequality]** For a random variable X with expectation $\mathrm{E}(X) = \mu$, and for any $\alpha > 0$,

$$\Pr[|X - \mu| \geq \alpha] \leq \frac{\mathrm{Var}(X)}{\alpha^2}.$$

Before proving Chebyshev's inequality, let's pause to consider what it says. It tells us that the probability of any given deviation, α, from the mean, either above it or below it (note the absolute value sign), is at most $\frac{\mathrm{Var}(X)}{\alpha^2}$. As expected, this deviation probability will be small if the variance is small. An immediate corollary of Chebyshev's inequality is the following:

Corollary 16.3: For a random variable X with expectation $\mathrm{E}(X) = \mu$, and standard deviation $\sigma = \sqrt{\mathrm{Var}(X)}$,

$$\Pr[|X - \mu| \geq \beta\sigma] \leq \frac{1}{\beta^2}.$$

Proof: Plug $\alpha = \beta\sigma$ into Chebyshev's inequality. □

So, for example, we see that the probability of deviating from the mean by more than (say) two standard deviations on either side is at most $\frac{1}{4}$. In this sense, the standard deviation is a good working definition of the "width" or "spread" of a distribution.

We should now go back and prove Chebyshev's inequality. The proof will make use of the following simpler bound, which applies only to *non-negative* random variables (i.e., r.v.'s which take only values ≥ 0).

Theorem 16.4: **[Markov's Inequality]** For a *non-negative* random variable X with expectation $\mathrm{E}(X) = \mu$, and any $\alpha > 0$,

$$\Pr[X \geq \alpha] \leq \frac{\mathrm{E}(X)}{\alpha}.$$

Proof: From the definition of expectation, we have

$$\begin{aligned}\mathrm{E}(X) &= \textstyle\sum_a a \times \Pr[X = a] \\ &\geq \textstyle\sum_{a \geq \alpha} a \times \Pr[X = a] \\ &\geq \alpha \textstyle\sum_{a \geq \alpha} \Pr[X = a] \\ &= \alpha \Pr[X \geq \alpha].\end{aligned}$$

The crucial step here is the second line, where we have used the fact that X takes on only non-negative values. (Why is this step not valid otherwise?) □

There is an intuitive way of understanding Markov's inequality through an analogy of a seasaw. Imagine that the distribution of a non-negative random variable X is resting on a fulcrum, $\mu = \mathrm{E}(X)$. We are trying to find an upper bound on the percentage of the distribution which lies beyond $k\mu$, i.e. $\Pr[X \geq k\mu]$. In other words, we seek to add as much weight m_2 as possible on the seesaw at $k\mu$ while minimizing the effect it has on the seesaw's balance. This weight will represent the upper bound we are searching for. To minimize the weight's effect, we must imagine that the weight of the distribution which lies beyond $k\mu$ is concentrated at exactly $k\mu$. However, to keep things stable and maximize the weight at $k\mu$, we must add another weight m_1 as far left to the fulcrum as we can so that m_2 is as large as it can be. The farthest we can go to the left is 0, since X is non-negative. Moreover, the two weights m_1 and m_2 must add up to 1, since they represent the area under the distribution curve:

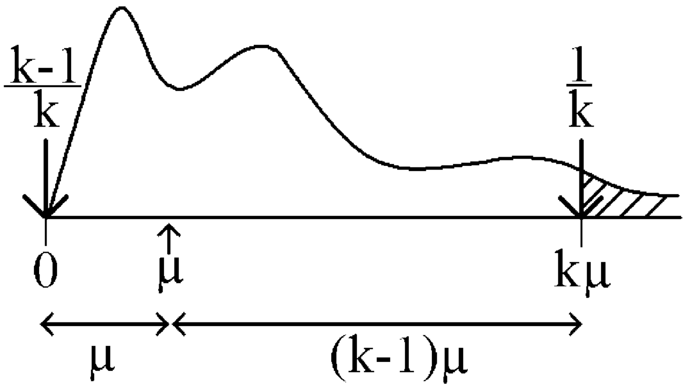

Figure 1: Markov's inequality interpreted as balancing a seasaw.

Since the lever arms are in the ratio $k-1$ to 1, a unit weight at $k\mu$ balances $k-1$ units of weight at 0. So the weights should be $\frac{k-1}{k}$ at 0 and $\frac{1}{k}$ at $k\mu$, which is exactly Markov's bound.

Now we can prove Chebyshev's inequality quite easily.

Proof of Theorem 16.2 Define the r.v. $Y = (X-\mu)^2$. Note that $\mathrm{E}(Y) = \mathrm{E}((X-\mu)^2) = \mathrm{Var}(X)$. Also, notice that the probability we are interested in, $\Pr[|X-\mu| \geq \alpha]$, is exactly the same as $\Pr[Y \geq \alpha^2]$. (Why?) Moreover, Y is obviously non-negative, so we can apply Markov's inequality to it to get

$$\Pr[Y \geq \alpha^2] \leq \frac{\mathrm{E}(Y)}{\alpha^2} = \frac{\mathrm{Var}(X)}{\alpha^2}.$$

This completes the proof. □

Let's apply Chebyshev's inequality to answer our question about the random walk at the beginning of this lecture note. Recall that X is our position after n steps, and that $\mathrm{E}(X) = 0$, $\mathrm{Var}(X) = n$. Corollary 16.3 says that, for any $\beta > 0$, $\Pr[|X| \geq \beta\sqrt{n}] \leq \frac{1}{\beta^2}$. Thus for example, if we take $n = 10^6$ steps, the probability that we end up more than 10000 steps away from our starting point is at most $\frac{1}{100}$.

Here are a few more examples of applications of Chebyshev's inequality (you should check the algebra in them):

1. **Coin tosses.** Let X be the number of Heads in n tosses of a fair coin. The probability that X deviates from $\mu = \frac{n}{2}$ by more than $\sqrt{n}$ is at most $\frac{1}{4}$. The probability that it deviates by more than $5\sqrt{n}$ is at most $\frac{1}{100}$.
2. **Poisson distribution.** Let X be a Poisson r.v. with parameter λ. The probability that X deviates from λ by more than $2\sqrt{\lambda}$ is at most $\frac{1}{4}$.
3. **Fixed points.** Let X be the number of fixed points in a random permutation of n items; recall that $\mathrm{E}(X) = \mathrm{Var}(X) = 1$. Thus the probability that more than (say) 10 students get their own homeworks after shuffling is at most $\frac{1}{100}$, however large n is.

In some special cases, including the coin tossing example above, it is possible to get much tighter bounds on the probability of deviations from the mean. However, for general random variables Chebyshev's inequality is essentially the only tool. Its power derives from the fact that it can be applied to *any* random variable.