# Variance

We have seen in the previous note that if we toss a coin $n$ times with bias $p$, 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 $np$?

**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. At each time step, the particle moves either one step to the right or one step to the left with equal probability, 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 at?

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. $X$ denote the position of the particle (relative to our starting point 0) after $n$ moves. Thus, we can write

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

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

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 we are really asking is: What is the expected value of $|X|$, the *distance* of the particle from 0? Rather than consider the r.v. $|X|$, which is a little difficult to work with 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 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 17.1.** *For the random variable $X$ defined in* (1)*, we have* $\mathrm{E}(X^2) = n$.

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

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

In the last line we have used linearity of expectation. To proceed, we need to compute $\mathrm{E}(X_i^2)$ and $\mathrm{E}(X_iX_j)$ (for $i \neq j$). Let's consider first $X_i^2$. Since $X_i$ can take on only values $\pm 1$, clearly $X_i^2 = 1$ always, so $\mathrm{E}(X_i^2) = 1$. What about $\mathrm{E}(X_iX_j)$? Well, $X_iX_j = +1$ 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 in the above calculation we used the fact that the events $X_i = +1$ and $X_j = +1$ are independent, and similarly the events $X_i = -1$ and $X_j = -1$ are independent. Thus $\mathbb{P}[X_iX_j = -1] = \frac{1}{2}$ as well, and hence $\mathrm{E}(X_iX_j) = 0$.

Plugging these values into the above equation gives

$$\mathrm{E}(X^2) = \sum_{i=1}^{n} 1 + \sum_{i \neq j} 0 = n,$$

as claimed. □

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 17.1** (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 17.1.** *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 above, we used linearity of expectation. Moreover, note that $\mu = \mathrm{E}(X)$ is a constant, so $\mathrm{E}(\mu X) = \mu\mathrm{E}(X) = \mu^2$ and $\mathrm{E}(\mu^2) = \mu^2$. □

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.

# 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 the previous note 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 17.1,

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

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:

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

   You should verify these as an exercise.

3. **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 the previous note that $\mathrm{E}(X) = 1$, regardless of $n$. To compute $\mathrm{E}(X^2)$, write $X = X_1 + X_2 + \cdots + X_n$, where $X_i = 1$ if $i$ is a fixed point, and $X_i = 0$ otherwise. Then as usual we have

   $$\mathrm{E}(X^2) = \sum_{i=1}^{n} \mathrm{E}(X_i^2) + \sum_{i\neq j} \mathrm{E}(X_iX_j). \tag{3}$$

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

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

   Make sure that you understand the last step here. Plugging this into equation (3) we get

   $$\mathrm{E}(X^2) = \sum_{i=1}^{n} \frac{1}{n} + \sum_{i\neq j} \frac{1}{n(n-1)} = (n \times \tfrac{1}{n}) + (n(n-1) \times \tfrac{1}{n(n-1)}) = 1 + 1 = 2.$$

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

# 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 the sequence, or a random variable $Y$ indicating the number of tails in a sequence of coin tosses, or a random variable $Z$ indicating whether the first is heads or not. Note, that for each sample point, any random variable has a specific value: for $HT$, $X = 1$, $Y = 1$, and $Z = 1$.

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

**Definition 17.2.** *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 for $X$, $Pr[X = a]$ is called the *marginal distribution* for $X$, and can be found by "summing out" over the values of $Y$. That is,

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

The marginal distribution for $Y$ is analagous, 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 $i$th roll of a sequence of $n$ die rolls) is $Pr[X_1 = x_1, \ldots, X_n = x_n]$ where $x_i \in S_i$ and $S_i$ 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.

# Independence of Random Variables

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

**Definition 17.3** (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*

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

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

This example motivates the commonly used phrase "*independent and identically distributed* **i.i.d.** set of random variables". In this example, the set of variables $\{X_i\}$ is a set of i.i.d. indicator random variables for a coin being Heads on each toss.

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 (i.e., they have the same distribution), then $\mathrm{Var}(X) = \sum_i \mathrm{Var}(X_i) = n \cdot \mathrm{Var}(X_1)$. This means that the standard deviation is $\sigma(X) = \sqrt{n} \cdot \sigma(X_1)$. Note that

by contrast, the expected value is $E[X] = n \cdot E[X_1]$. Intuitively this means that whereas the average value of $X$ grows proportionally to $n$, the spread of the distribution grows proportionally to $\sqrt{n}$, 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 17.2.** *For* independent *random variables $X, Y$, we have* $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$.

*Proof.* We have

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

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

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

**Theorem 17.3.** *For* independent *random variables $X, Y$, we have*

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

*Proof.* From the alternative formula for variance in Theorem 17.1, 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 17.2 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 $\mathbb{P}[X = 1] = p$, $\mathrm{E}(X^2) = p$ 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 $n, p$). As usual, we write $X = X_1 + X_2 + \cdots + X_n$, where $X_i = 1$ if the $i$-th toss is Head, and $X_i = 0$ otherwise.

We already know $\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(X_i) = np$. We can compute $\mathrm{Var}(X_i) = E(X_i^2) - E(X_i)^2 = p - p^2 = p(1-p)$. Since the $X_i$'s are independent, by Theorem 17.3 we get $\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1-p)$.

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 (2), where the standard deviation $\sigma(X) = \sqrt{\frac{n^2-1}{12}} \approx \frac{n}{\sqrt{12}}$ 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.