Variance

The expectation is like the center of the distribution, it gives a rough idea of where the values are. But we are still missing something important. The mean does not describe the randomness of the phenomenon, meaning the variation around the expectation of the random variable. Consider again the scores on an exam and suppose the mean is given to be 50: all the students could have gotten exactly 50, but the scores can also vary from 0 to 100 as long as they still satisfy the mean constraint. The spread of the random variable around its mean is related to how much uncertainty there is in the random variable. We will now define a notion of spread of the random variable.

We are interested in summarizing the variation of the random variable around its mean. Let $\mu = \mathbb{E}[X]$. One possible measure of deviation from the mean of $X$ is $X - \mu$. The deviation can be small or large depending on the values the random variable takes on. So $X - \mu$ is itself a random variable, and we can compute its mean to get the average of the deviation. However, $\mathbb{E}[X - \mu]$ is equal to zero by linearity. This happens because it takes on positive and negative values which cancel out. So actually $X - \mu$ is not a good measure of deviation. Instead, we could square $X - \mu$ or take its absolute value as a better measure of deviation, as then it will be non-negative. Unfortunately, computing the expected value of $|X - \mu|$ turns out to be a little awkward, due to the absolute value operator. Therefore we consider the random variable $(X - \mu)^2$ as a measure of deviation of $X$ around the mean. We take its expectation and get the variance of the random variable to be:

**Definition 8.1 (variance):** The variance of a discrete random variable $X$ is defined as

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

If $X$ measures the height in cm, the unit of the variance is $cm^2$. So we also define the standard deviation of a random variable.

**Definition 8.2 (Standard deviation):** The standard deviation of a discrete random variable $X$ is defined as

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

which has the same unit as the random variable $X$.

A random variable with 0 variance assigns every outcome to the same number (and therefore not random). In general, the variance will be a positive number which models the spread of the variable. The larger the variance, the larger the spread of the random variable. The mean of a random variable is interpreted as being the center of mass, and the variance represents the moment of inertia.

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

**Theorem 8.1:** For a r.v. $X$ with mean $\mathbb{E}[X] = \mu$, we have $\text{Var}(X) = \mathbb{E}[X^2] - \mu^2$. 

\[ \text{Proof:} \ \text{From the definition of variance, we have} \]
\[ \text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu \mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2. \]

In the second step here, we used linearity of expectation and that \( E[\mu^2] = \mu^2 \) because \( \mu \) is constant. \( \square \)

\[ \mathbb{E}[X^2] \] is called the second moment of the random variable. So the variance can be written in terms of the first and second moments of the random variable.

Let us now look at some examples:

1. \textbf{Uniform distribution.} \( X \) 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 of \( X \) is equal to:

\[ \mathbb{E}[X] = \frac{n + 1}{2} \]

To compute the variance, we just need to compute \( \mathbb{E}[X^2] \), which is a routine calculation:

\[ \mathbb{E}[X^2] = \sum_{a=1}^{n} a^2 \times P_X(a) = \sum_{a=1}^{n} a^2 \times \frac{1}{n} = \frac{(n+1)(2n+1)}{6}. \]

Finally we get that

\[ \text{Var}(X) = \frac{n^2 - 1}{12}, \quad \sigma_X = \sqrt{\frac{n^2 - 1}{12}}. \]

(You should verify these.) Note that the variance is proportional to \( n^2 \), and the standard deviation is proportional to \( n \), the range of the r.v..

2. \textbf{Binomial distribution.} Consider \( X \sim \text{Bin}(n, p) \). Then \( \mu = \mathbb{E}[X] = np \).

To compute \( \text{Var}(X) \), let us start by directly trying to compute it from the distribution:

\[ \text{Var}(X) = \sum_{a} (a - \mu)^2 P(X = a) = \sum_{a=0}^{n} (a - np)^2 \left( \frac{n}{i} \right) p^a (1 - p)^{n-a}. \]

where we have used the fact that \( X \) has a binomial distribution.

This sum doesn’t seem easy to evaluate. Let’s try the trick we did for computing the mean, by writing \( X = X_1 + X_2 + \cdots + X_n \) where

\[ X_i = \begin{cases} 
1 & \text{if } \text{ith toss is Heads}, \\
0 & \text{otherwise.}
\end{cases} \]

\( X_i \) is called the \textit{indicator r.v.} for the event "the \( i \)th toss is Heads"; it indicates when the event occurs. Now, we know that

\[ \mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]. \]

Is it also true that

\[ \text{Var}(X) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)? \]

Turns out that while this is not true in general, it \textit{is} true if the \( X_i \)’s are independent. We have the following facts:

\textbf{Lemma 8.1:} \textit{Let} \( X \) \textit{and} \( Y \) \textit{be two independent random variables. Then}
You will be asked to verify these two facts in the homework. Using the second fact, we have

\[ \text{Var}(X) = n \text{Var}(X_1). \]

By direct computation, \( \text{Var}(X_1) = p(1-p) \). Hence,

\[ \text{Var}(X) = np(1-p) \]

and

\[ \sigma_X = \sqrt{np(1-p)}. \]

**Note:** Both the uniform and binomial random variables take values in the range from 0 to \( n \). For the binomial distribution, the mean and variance are linear in \( n \), and the standard deviation is linear in \( \sqrt{n} \). For the uniform distribution, the mean is linear in \( n \), the variance linear in \( n^2 \) and the standard deviation linear in \( n \). The standard deviation of the binomial is shrunk relative to its range. This is because there is a mass concentration in a band around its mean (See figure 1). However, in the case of the uniform distribution, the standard deviation is more spread so of order \( n \). The probability mass concentration for the binomial distribution is an important phenomenon, and will be studied in greater depth in upcoming lectures.

As another example to work out on your own: compute the variance of the number of students that get their homework back.

**Some Important Distributions**

We have already discussed an important distribution: \( X \sim \text{Bin}(n, p) \). We will now give some more examples of important distributions: \( X \sim \text{Geom}(p) \) and \( X \sim \text{Poiss}(\lambda) \).

**Geometric Distribution:** Consider the experiment where we flip a coin until we see a head.

\[ \Omega = \{H, TH, TTH, TTTTH, \cdots \} \]
Let $X$ denote the variable that counts the number of flips until and including the first Heads we obtain. There is a one to one correspondence between the outcomes and the random variable.

Assuming the results of the flips are independent, $P(H) = p$, $P(TH) = (1-p)p$. More generally,

$$P(i-1 \text{ Tails before a Heads}) = (1-p)^{i-1}p.$$ 

So, the random variable $X$ has the following distribution:

$$P(X = i) = (1-p)^{i-1}p \text{ for } i = 1, 2, \ldots$$

A chance of getting a large value with a Geometric random variable is very small (See figure 2).

Now if we go back to the bit torrent servers example. A video is broken down into $m$ chunks. Each server has a random chunk, i.e. one out of $m$ possible choices. We are interested in the first time in which we have the whole movie (meaning all the $m$ chunks). Let $X$ be the number of servers we need to query before we get all $m$ chunks. How do we analyze this problem? How do we compute $E[X]$? There is a natural way of breaking this random variable into easier steps. You are making progress whenever you get a new chunk. Let $X_1$ be the time to get a new chunk. Then $E[X_1] = X_1 = 1$. Let $X_2$ be the time until we get the second new chunk after we already get one chunk. Can you proceed from here?