Chebyshev’s inequality

There are two important numbers to calculate to describe a distribution: the mean and the variance. We have seen that, intuitively, the variance (or, more correctly the standard deviation) is a measure of "spread", or deviation from the mean. For instance if the variance is small, then the distribution of the variable is not very spread out meaning that the chance of getting an outcome that is far from the mean is small. For example, in the homework example from last lecture, since the mean and variance are both 1, we would expect that the number of students getting back their homeworks should be between 0 and 3 with significant probability.

Our next goal is to make this intuition quantitatively precise. What we can show is the following:

**Theorem 8.1: [Chebyshev’s Inequality]** For a random variable \( X \) with expectation \( \mathbb{E}[X] = \mu \), and for any \( a > 0 \),

\[
P(|X - \mu| \geq a) \leq \frac{\text{Var}(X)}{a^2}.
\]

We are interested in the quantity: \( P(|X - \mu| \geq a) \): what is the chance that the random variable is greater than \( \mu + a \) or smaller than \( \mu - a \)? The claim is that if the variance of the random variable is small, then this probability is small. Chebyshev’s inequality allows us to upper bound this probability by \( \frac{\text{Var}(X)}{a^2} \). If we know the distribution, we can calculate this number exactly. However, we might not have the whole distribution, or it might be too complicated. Even without knowing the whole distribution, Chebyshev tells you that you can bound the probability of this event. For example, in the homework problem, the probability that the number of students getting back their own homework is outside the range 0-3 is less than 1/4 (by setting \( a = 2 \) in the inequality).

Before proving the inequality, let’s pause to consider what it says. The upper bound decreases as \( a \) becomes larger. This means that the chance of deviating more and more from the mean gets smaller which seems reasonable. Moreover, the smaller the variance of the random variable, the smaller the probability of being far away from the mean which implies a smaller spread as we wanted to argue.

We now prove Chebyshev’s inequality. We define \( Y = |X - \mu| \). The inequality can be rewritten in terms of \( Y \) as

\[
P(Y \geq a) \leq \frac{\mathbb{E}[Y^2]}{a^2}.
\]

Consider two functions of \( Y \), \( f(Y) = Y^2/a^2 \) and \( g(Y) = 1_{Y \geq a} \). \( g(Y) \) is 1 whenever \( Y \geq a \) and is 0 otherwise. It is easy to verify that \( f(Y) \geq g(Y) \) for all possible \( Y \geq 0 \) (\( Y \) is always non-negative since it is the absolute value of \( X - \mu \)). Thus,

\[
\mathbb{E}[g(Y)] \leq \mathbb{E}[f(Y)]
\]

Now, the expectation of an indicator function is the probability of the corresponding set, i.e., \( \mathbb{E}[g(Y)] =
\( P(Y \geq a) \). Thus, we get,

\[
P(Y \geq a) \leq \frac{\mathbb{E}[Y^2]}{a^2}
\]

Geometric Distribution

Recall the geometric distribution discussed in last class, with \( X \) being the variable that counts the number of flips until and including the first Heads we obtain. The random variable \( X \) has the distribution:

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

where \( p \) is the probability that a toss is heads.

Now we compute the expectation of the geometric distribution.

\[
\mathbb{E}[X] = \sum_{i=1}^{\infty} i \times P(X = i)
\]

\[
= \sum_{i=1}^{\infty} i(1 - p)^{i-1}p
\]

\[
= \sum_{i=1}^{\infty} (i - 1 + 1)(1 - p)^{i-1}p
\]

\[
= \sum_{i=1}^{\infty} (i - 1)(1 - p)^{i-1}p + \sum_{i=1}^{\infty} (1 - p)^{i-1}p
\]

The second sum is 1 (since the probabilities of all possible \( i \)'s sum to 1). Changing the index of summation in the first summation to \( j = i - 1 \), we get,

\[
\mathbb{E}[X] = \sum_{j=0}^{\infty} j(1 - p)^j p + 1
\]

\( j = 0 \) is irrelevant as the term becomes 0,

\[
\mathbb{E}[X] = \sum_{j=1}^{\infty} j(1 - p)^j p + 1
\]

Pulling out \( 1 - p \) from the summation,

\[
\mathbb{E}[X] = (1 - p) \sum_{j=1}^{\infty} j(1 - p)^{j-1} p + 1
\]

But the summation is now precisely \( \mathbb{E}[X] \),

\[
\mathbb{E}[X] = (1 - p)\mathbb{E}[X] + 1
\]

Solving this, we get,

\[
\mathbb{E}[X] = \frac{1}{p}
\]
BitTorrent Servers’ Example

A video is broken down into \( n \) chunks. Each server has a random chunk, i.e. one out of \( n \) possible choices. We are interested in the number of servers we need to query to have the whole movie (meaning all \( n \) chunks). Let \( X \) be the number of servers we query before we get the \( n \) chunks. How do we compute \( E[X] \)? Of course we will need to query at least \( n \) servers. But because some servers will store the same chunk, we may need more. The question is, how many more on the average?

There is a natural way of breaking this random variable into a sum of simpler random variables. We make progress whenever we get a new chunk. We write \( X = Z_1 + \cdots + Z_n \) in order to use linearity of expectation. Let \( Z_1 \) be the time to get the first new chunk. The first server we query will give a new unobserved chunk no matter what. Therefore \( Z_1 = 1 \), and \( Z_1 \) is a deterministic random variable. \( Z_2 \) is the number of extra servers you query until you see the second new chunk, and similarly \( Z_n \) is the number of servers you query until you get the last chunk. The chunks are not ordered here, it does not matter which one comes first as long as we have all \( n \) of them at the end.

Let us now determine the distribution of \( Z_2 \) to compute its expectation. \( Z_2 \sim \text{Geom}(p) \) where \( p \) is the success probability. The probability of success is the probability of getting a new chunk. When you sample the next server, you will find a new chunk with a high probability: \( \frac{n-1}{n} \). Therefore, \( E[Z_2] = \frac{n}{n-1} \) which is slightly greater than 1. Similarly, \( Z_3 \sim \text{Geom}\left(\frac{n-2}{n}\right) \).

In general, \( X_i \sim \text{Geom}\left(\frac{n-i+1}{n}\right) \) and \( E[Z_i] = \frac{n}{n-i+1} \). In particular, \( Z_n \sim \text{Geom}\left(\frac{1}{n}\right) \).

Finally, we can compute the desired expectation

\[
E[X] = E[Z_1] + \cdots + E[Z_n] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \left[ \frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots + \frac{1}{1} \right].
\]

If \( n \) is very large, the number in brackets is roughly equal to \( \ln n \) (which can be seen by looking at the area under the curve \( \frac{1}{x} \)). This means we have to query a factor of \( \ln n \) more servers than the case where we know exactly where the chunks are (in which case we need only to query \( n \) servers.) This is the price of randomness in the protocol.

This is an important application of the geometric distribution. In general, once you figure out the relevant distribution of a random variable in a problem, you just need to determine the parameter of that distribution.

Some properties of variance

**Theorem 8.2:** Let \( X \) be a random variable and let \( c \) be a constant.

1. \( \text{Var}(X + c) = \text{Var}(X) \)
2. \( \text{Var}(cX) = c^2 \text{Var}(X) \)

**Proof:**

1. 
\[
\text{Var}(X + c) = E[(X + c)^2] - E[X + c]^2
\]
$$= \mathbb{E}[X^2 + 2cX + c^2] - (\mathbb{E}[X] + c)^2$$
$$= \mathbb{E}[X^2] + \mathbb{E}[2cX] + c^2 - (\mathbb{E}[X]^2 + c^2 + 2c\mathbb{E}[X])$$
$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$
$$= \text{Var}(X)$$

2.

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