CS 30: Discrete Math in CS (Winter 2020): Lecture 17

Date: 7th February, 2020 (Friday)

Topic: Probability: Independence, Variance

Disclaimer: These notes have not gone through scrutiny and in all probability contain errors. Please discuss in Piazza/email errors to deeparnab@dartmouth.edu

---

1. **Independent Random Variables.** Two random variables X and Y are independent, if for any $x \in \text{range}(X)$ and any $y \in \text{range}(Y)$,

$$\mathbf{Pr}[X = x, Y = y] = \mathbf{Pr}[X = x] \cdot \mathbf{Pr}[Y = y]$$

Examples:

- If we roll two dice, and X_1 and X_2 indicate the value of the rolls, then X_1 and X_2 are independent.
- If we have two independent events $\mathcal{A}$ and $\mathcal{B}$, then their indicator random variables $\mathbf{1}_{\mathcal{A}}$ and $\mathbf{1}_{\mathcal{B}}$ are independent.
- Consider a random variable X taking value $+1$ if a toss of a coins is head, and -1 if its tails. Such random variables are called *Rademacher random variables*. Suppose we toss the coin twice and X_1 and X_2 are the corresponding random variables. Then X_1 and X_2 are independent.

A set of k random variables $X_1, \ldots, X_k$ are *mutually independent* if for any $x_1, x_2, \ldots, x_k$ with $x_i \in \text{range}(X_i)$, we have

$$\mathbf{Pr}[X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k] = \prod_{i=1}^{k} \mathbf{Pr}[X_i = x_i]$$

Theorem 1. If X and Y are two independent random variables, then

$$\mathbf{Exp}[XY] = \mathbf{Exp}[X] \cdot \mathbf{Exp}[Y]$$

Proof.

$$\begin{aligned}
\mathbf{Exp}[XY] &= \sum_{x \in \text{range}(x), y \in \text{range}(y)} (xy) \cdot \mathbf{Pr}[X = x, Y = y] && \text{Definition of Expectation} \\
&= \sum_{x \in \text{range}(x), y \in \text{range}(y)} (xy) \cdot \mathbf{Pr}[X = x] \cdot \mathbf{Pr}[Y = y] && \text{Independence} \\
&= \left(\sum_{x \in \text{range}(x)} x \cdot \mathbf{Pr}[X = x] \right) \cdot \left(\sum_{y \in \text{range}(y)} y \cdot \mathbf{Pr}[Y = y] \right) && \text{Algebra} \\
&= \mathbf{Exp}[X] \cdot \mathbf{Exp}[Y] && \text{Definition of Expectation}
\end{aligned}$$

□

Of course, there is no need to stick to two random variables. The theorem easily generalizes (do you see how?) to mutually independent random variables as follows.

Theorem 2. If $X_1, X_2, \ldots, X_k$ are mutually independent random variables, then

$$\mathbf{Exp}\left[\prod_{i=1}^{k} X_i\right] = \prod_{i=1}^{k} \mathbf{Exp}\left[X_i\right]$$

Examples.

- Let X_i and X_j be two independent Rademacher random variables. Recall, X_i takes $+1$ with probability $1/2$ and -1 with probability $1/2$. Then note (a) $\mathbf{Exp}[X_i] = \mathbf{Exp}[X_j] = 0$, (b) $\mathbf{Exp}[X_i \cdot X_i] = \mathbf{Exp}[X_j \cdot X_j] = 1$, and (c) $\mathbf{Exp}[X_i X_j] = \mathbf{Exp}[X_i] \cdot \mathbf{Exp}[X_j] = 0$. This is a very useful fact.
- Consider rolling a die n times, independently. Let Z be the random variable indicating the *product* of all the numbers seen. What is $\mathbf{Exp}[Z]$? To solve this, let X_i be the roll of the ith die. We know that $\mathbf{Exp}[X_i] = 3.5$ for all i. We also know $X_1, X_2, \ldots, X_n$ are all independent random variables. Thus, $\mathbf{Exp}[Z] = (3.5)^n$.

2. **Variance and Standard Deviation.**

The expectation of a random variable is some sort of an "average behavior" of a random variable. However, the true value of a random variable may be no where close to the expectation. For instance, consider a random variable which takes the value 10000 with probability $1/2$, and -10000 with probability $1/2$. What is $\mathbf{Exp}[X]$? Yes, it is 0. Thus, there is significant *deviation* of X from its expectation.

The variance and standard deviation try to capture this deviation. In particular, the variance of a random variable is the *expected value of the square of the deviation*.

Let X be a random variable. The variance of X is defined to be

$$\mathbf{Var}[X] := \mathbf{Exp}\left[(X - \mathbf{Exp}[X])^2\right]$$

That is, if we define another random variable $D := (X - \mathbf{Exp}[X])^2$, then $\mathbf{Var}[X]$ is the expected value of this new deviation random variable D.

The *standard deviation* $\sigma(X)$ is defined to be $\sqrt{\mathbf{Var}(X)}$.

Theorem 3. $\mathbf{Var}[X] = \mathbf{Exp}[X^2] - (\mathbf{Exp}[X])^2$.

Proof.

$$\mathbf{Var}[X] = \mathbf{Exp}[(X - \mathbf{Exp}[X])^2] = \mathbf{Exp}[X^2 - 2X\mathbf{Exp}[X] + (\mathbf{Exp}[X])^2]$$

Then, we apply linearity of expectation to get

$$\mathbf{Var}[X] = \mathbf{Exp}[X^2] - 2\mathbf{Exp}[X] \cdot \mathbf{Exp}[X] + (\mathbf{Exp}[X])^2 = \mathbf{Exp}[X^2] - (\mathbf{Exp}[X])^2$$

□

A useful corollary (something we observed in the last lecture notes):

Theorem 4. For any random variable $\mathbf{Exp}[X^2] \geq (\mathbf{Exp}[X])^2$.

Proof. $\mathbf{Var}[X]$ is the expected value of $(X - \mathbf{Exp}[X])^2$. That is, $\mathbf{Var}[X]$ is the expected value of a random variable which is always non-negative. In particular, $\mathbf{Var}[X]$ is non-negative. Which in turn means $\mathbf{Exp}[X^2] - (\mathbf{Exp}[X])^2 \geq 0$. Rearranging implies the corollary. □

Examples

- *Roll of a die.* Let X be the roll of a fair 6-sided die. We know that $\mathbf{Exp}[X] = 3.5$. To calculate the variance, we can use the deviation $D := (X - \mathbf{Exp}[X])^2 = (X - 3.5)^2$. Usinhg this, we get

$$\mathbf{Var}[X] = \mathbf{Exp}[D] = \frac{1}{6}\left((2.5)^2 + (1.5)^2 + (0.5)^2 + (0.5)^2 + (1.5)^2 + (2.5)^2\right) = \frac{35}{12}$$

- *Toss of a biased coin.* Let X be a Bernoulli random variable taking value 1 if a coin tosses heads, and 0 otherwise. Suppose the probability of heads was p. Recall, $\mathbf{Exp}[X] = p$. Also note since X is a indicator random variable, $X^2 = X$. Thus, $\mathbf{Exp}[X^2] = p$ as well. We can calculate the variance as

$$\mathbf{Var}[X] = \mathbf{Exp}[X^2] - (\mathbf{Exp}[X])^2 = p - p^2 = p(1-p)$$

- *Indicator Random Variable.* Using the above toss of a biased coin example, we see that for any event $\mathcal{E}$, the variance of the indicator random variable is

$$\mathbf{Var}[\mathbf{1}_{\mathcal{E}}] = \mathbf{Pr}[\mathcal{E}] \cdot (1 - \mathbf{Pr}[\mathcal{E}]) = \mathbf{Pr}[\mathcal{E}] \cdot \mathbf{Pr}[\neg\mathcal{E}]$$

Theorem 5. If X is a random variable, and c is a "scalar" (a constant), then $Z = c \cdot X$ is another random variable. $\mathbf{Var}[c \cdot X] = c^2 \cdot \mathbf{Var}[X]$.

Proof.

$$\mathbf{Var}[c \cdot X] = \mathbf{Exp}[c^2X^2] - (\mathbf{Exp}[cX])^2 = c^2\mathbf{Exp}[X^2] - c^2\left(\mathbf{Exp}[X]\right)^2 = c^2\mathbf{Var}[X]$$

□

The next theorem is a *linearity of variance* result for ***independent*** random variables.

Theorem 6. For any two ***independent*** random variables X and Y, let $Z := X + Y$. Then,

$$\mathbf{Var}[Z] = \mathbf{Var}[X] + \mathbf{Var}[Y]$$

Proof. By definition of variance, we get

$$\mathbf{Var}[X + Y] = \mathbf{Exp}[(X + Y)^2] - (\mathbf{Exp}[X] + \mathbf{Exp}[Y])^2 \tag{1}$$

Now, we expand the first term in the RHS to get

$$\begin{aligned}\mathbf{Exp}[(X+Y)^2] &= \mathbf{Exp}[X^2+2XY+Y^2] \\ &= \mathbf{Exp}[X^2]+2\mathbf{Exp}[XY]+\mathbf{Exp}[Y^2] && \text{Linearity of Expectation} \\ &= \mathbf{Exp}[X^2]+2\mathbf{Exp}[X]\mathbf{Exp}[Y]+\mathbf{Exp}[Y^2] && \text{Since } X \text{ and } Y \text{ are independent.}\end{aligned} \tag{2}$$

Next, we expand the second term in the RHS of (1), to get

$$(\mathbf{Exp}[X]+\mathbf{Exp}[Y])^2 = (\mathbf{Exp}[X])^2+2\mathbf{Exp}[X]\mathbf{Exp}[Y]+(\mathbf{Exp}[Y])^2 \tag{3}$$

Subtracting (3) from (2), we get

$$\begin{aligned}\mathbf{Var}[X+Y] &= \left(\mathbf{Exp}[X^2]-(\mathbf{Exp}[X])^2\right)+\left(\mathbf{Exp}[Y^2]-(\mathbf{Exp}[Y])^2\right) \\ &= \mathbf{Var}[X]+\mathbf{Var}[Y]\end{aligned} \tag{4}$$

□

We can generalize the above proof to many random variables. In particular, we can say that if $X_1, X_2, \ldots, X_k$ are mutually independent random variables, then the variance of the sum is the sum of the variances. However, we *don't need mutual independence*. Pairwise independence suffices. The proof is given as a solution to the UGP; perhaps you can try it. There is nothing more than the algebra above except there are k things adding up.

Theorem 7. For any k ***pairwise independent*** (and therefore also for mutually independent) random variables $X_1, X_2, \ldots, X_k$,

$$\mathbf{Var}\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} \mathbf{Var}[X_i]$$

3. **Deviation Inequalities**

We have seen an example that $\mathbf{Exp}[X]$ may not be anywhere close to what values X can take (recall the $X = 10000$ with 0.5 probability and -10000 with 0.5 probability). Deviation inequalities try to put an *upper bound* on the probability that a random walk deviates too far from the expectation.

The mother of all deviation inequalities is the following:

Theorem 8. (Markov's Inequality)

Let X be a random variable whose range is *non-negative reals*. Then for any $t > 0$, we have

$$\mathbf{Pr}[X \geq t] \leq \frac{\mathbf{Exp}[X]}{t}$$

Before we embark on to the proof of Markov's inequality, let us actually understand what it says. For simplicity, assume the probability distribution is uniform (so the expectation is the usual "average"). And also let's fix $t = 2$. Also, just for concreteness, let X denote the height of a random person in a group of people. Then, Markov states that the fraction of people whose height is *at least* twice the average is *at most* $1/2$. Indeed, if not, then more than $1/2$ the fraction will be more than 2 times the average, but that will just drive the average up. The proof below is basically this argument for general probability distributions.

Proof. By definition of expectation, we have

$$\mathbf{Exp}[X] = \sum_{k \in \mathbb{R}} k \cdot \mathbf{Pr}[X = k] = \sum_{0 \le k < t} k \cdot \mathbf{Pr}[X = k] + \sum_{k \ge t} k \cdot \mathbf{Pr}[X = k]$$

The first summation $\sum_{0 \le k < t} k \cdot \mathbf{Pr}[X = k] \ge 0$ since all terms are non-negative. The second summation is $\sum_{k \ge t} k \cdot \mathbf{Pr}[X = k] \ge t \cdot \sum_{k \ge t} \mathbf{Pr}[X = k] = t \cdot \mathbf{Pr}[X \ge t]$.

Putting it all together, we get

$$\mathbf{Exp}[X] \ge t \cdot \mathbf{Pr}[X \ge t]$$

which gives what we want by rearrangement. □

Markov's inequality only talks about non-negative random variables. Indeed, the example in the beginning of this bullet point shows that it cannot be true for general random variables. This is where *variance* comes to play. The following is one of the most general forms of deviation inequalities.

Theorem 9. (Chebyshev's Inequality)
Let X be a random variable. Then for any $t > 0$, we have

$$\mathbf{Pr}[|X - \mathbf{Exp}[X]| \ \ge \ t] \le \frac{\mathbf{Var}[X]}{t^2}$$

Proof. We first note that

$$\mathbf{Pr}[|X - \mathbf{Exp}[X]| \ge t] = \mathbf{Pr}[(X - \mathbf{Exp}[X])^2 \ge t^2]$$

Then we notice that $D := (X - \mathbf{Exp}[X])^2$ is a non-negative random variable, and therefore we can apply Markov's inequality on it to get

$$\mathbf{Pr}[|X - \mathbf{Exp}[X]| \ge t] = \mathbf{Pr}[D \ge t^2] \le \frac{\mathbf{Exp}[D]}{t^2} = \frac{\mathbf{Var}[X]}{t^2}$$

□

Theorem 10. A useful corollary to the above, and one which is often used as rule of

thumb, is obtained by setting $t = c\sigma(X)$ for some $c \geq 0$. One gets,

$$\mathbf{Pr}[|X - \mathbf{Exp}[X]| \geq c\sigma(X)] \leq \frac{1}{c^2}$$

Proof. When $t = c\sigma(X)$ is substituted in Chebyshev's inequality, one gets the RHS in the above corollary by reminding oneself that $\sigma(X) = \sqrt{\mathbf{Var}(X)}$. □

Example

- Suppose we toss 1000 fair coins. What are the chances that we see more than 600 heads? In this case, let Z be the random variable which evaluates to the number of heads seen in the toss of 1000 coins. We are interested in the question

  $$\mathbf{Pr}[Z \geq 600]?$$

  To evaluate this, we define random variables $X_1, X_2, \ldots, X_{1000}$, where X_i is the indicator random variable for the ith toss; that is, it is defined to be 1 if the ith toss is heads, and it is defined to be 0 if the ith toss is tails. We observe four *crucial* things:
  - $Z = X_1 + X_2 + \cdots + X_{1000}$.
  - $\mathbf{Exp}[X_i] = 0.5$ for all $1 \leq i \leq 1000$. This is because the coins are fair.
  - $X_1, X_2, \ldots, X_{1000}$ are (mutually) *independent*.
  - $\mathbf{Var}[X_i] = 0.25$ (see variance example above – with $p = 0.5$)

  Linearity of expectation gives us

  $$\mathbf{Exp}[Z] = \sum_{i=1}^{1000} \mathbf{Exp}[X_i] = 1000 \cdot 0.5 = 500$$

  The fact that the X_i's are (mutually) independent, allows us to use linearity of variance (Theorem 7), to get

  $$\mathbf{Var}[Z] = \sum_{i=1}^{1000} \mathbf{Var}[X_i] = 1000 \cdot 0.25 = 250$$

  Finally, we can apply Chebyshev's inequality as follows

  $$\begin{aligned}
  \mathbf{Pr}[Z \geq 600] &= \mathbf{Pr}[Z - 500 \geq 100] && \text{We have subtracted the expectation from both sides} \\
  &\leq \mathbf{Pr}[|Z - 500| \geq 100] && \text{if } Z - 500 \geq 100, \text{ surely the absolute value is.} \\
  &\leq \frac{\mathbf{Var}(Z)}{100^2} && \text{Chebyshev's Inequality} \\
  &= \frac{1}{40} && \text{Substituting } \mathbf{Var}[Z] = 250.
  \end{aligned}$$

  Thus, the chances we see more than 600 heads is *at most* 2.5%.

Remark: *The* true *answer to the question of what is the probability we see more than* 600 *heads is in fact* much, much *lower. The reason is that when a random variable can be written as a sum of mutually independent random variables, then the rule of thumb for the deviations is*

The probability X *is more than* c *standard deviations away is of the order of* $e^{-c^2/2}$

The above statement is qualitative rather than quantitative (and therefore I use the term "order of"). But one can see in the above coins example, the standard deviation is $\sqrt{250} \approx 16$*. Thus seeing more than* 100 *heads than the mean is being off by more than* 6 *standard deviations. The chances of this is roughly* $e^{-6^2/2}$ *which is roughly* 1 *in* 100 *million! Way smaller than* 2.5%*.*

You should use a computer to check it out.

✍

Exercise: *Do the following exercises mimicking the above example.*

- *Suppose every email I get independently is spam with probability* 1%*. I receive* 100 *emails. What is the probability that more than* 7 *of them are spam?*
- *Suppose I roll* 100 *normal dice, and add the sum up. What is the probability that the total sum is less than* 100*?*