CS 30: Discrete Math in CS (Winter 2020): Lecture 16, 17

Date: 6th February, 2020 (X-hour) + 7th February, 2020 (Friday)
Topic: Probability: Random Variables, Expectation, 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. **Random Variable.**

   Given a random experiment with outcomes Ω, a *real valued random variable* X defined over this experiment is a mapping $X : \Omega \to \mathbb{R}$. An *integer valued random variable* X is a mapping from $X : \Omega \to \mathbb{Z}$.

   Examples:

   - We toss a fair coin. $X(\text{heads}) = 0$ and $X(\text{tails}) = 1$. This is a $\{0, 1\}$-random variable, or a Boolean random variable. Also called a *Bernoulli* random variable.
   - We roll a fair die. X takes the value on the face of the die.
   - We roll *two* fair dice. X takes the value of the sum. In this case, $X = Y + Z$ where Y, Z are two *identical* random variables of the kind from the previous bullet point.
   - We toss 1000 fair coins. Z takes the value of the number of heads we see.
   - Given any event $\mathcal{E}$, there is an associated random variable called the *indicator random variable* denoted as $\mathbf{1}_{\mathcal{E}}$, where $\mathbf{1}_{\mathcal{E}}(\omega) = 1$ if $\omega \in \mathcal{E}$, and 0 otherwise.

2. **Events associated with random variables.**

   Given a random variable X, we can associate many events and ask for their probabilities. For instance, we can ask $\mathbf{Pr}[X = x]$. More precisely, this is a shorthand for saying $\sum_{\omega \in \Omega : X(\omega) = x} \mathbf{Pr}[\omega]$.

   Similarly, $\mathbf{Pr}[X \geq k]$ is a shorthand for saying $\sum_{\omega \in \Omega : X(\omega) \geq k} \mathbf{Pr}[\omega]$.

3. **"Shape" of a Random Variable.**

   Since X is real valued (or integer valued), one can plot how the $\mathbf{Pr}[X = x]$ looks like with respect to X. The following plots show a couple of examples. The first set of figures (Figure 1) is related to dice. We roll N dice, each independent of one another, and we use X to denote the sum of the numbers seen. The plots show how $\mathbf{Pr}[X = x]$ changes with x, as x goes from 0 to $6N + 1$. As you can see, when $N = 1$, the probabilities are the same for each number, and equals $1/6$th. However, the distribution becomes less and less uniform as N grows.

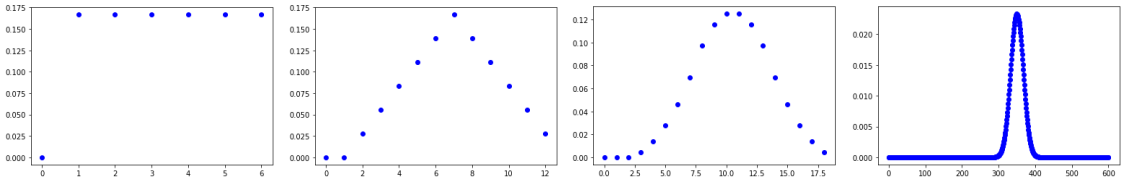

Figure 1: *The above graphs plot the probability of seeing a particular sum on the Y-axis against the possible sums on the X-axis. From left to right, the number of dice is* $1, 2, 3$ *and* 100.

The next set of figures (Figure 2) relate to coin tosses. We toss N coins and Z denotes the number of heads we see. The plots in blue (the ones to the left) are the plots of tosses of fair coins which turn up heads 50-50. The plots in green (the ones to the right) are for biased coins which come up heads with probability 0.3.

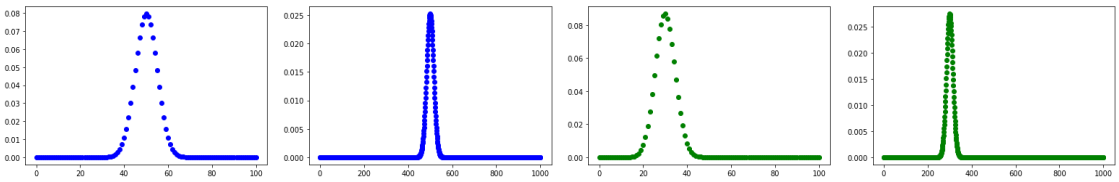

Figure 2: *The above graphs plot the probability of seeing a particular number of heads on the Y-axis against the reals on the X-axis. The first two figures (in blue) on the left are for fair coins, with* $N = 100$ *coins tossed and* $N = 1000$ *coins tossed. The two figures in the right (in green) are for biased coins which come heads with* 0.3 *probability. The number of coins are* $N = 100$ *and* $N = 1000$ *respectively.*

Remark: *A few points are noteworthy*

- *Note the shapes become "narrower" as the number of coins/dice grow.*
- *Note that the shape of fair coin is similar to the shape of biased coins with just a shift.*
- *Note that the* 100 *dice shape looks quite similar to the shape with* 1000 *coins.*

All of these happen for a very important reason (which we will not cover, unfortunately). The reason, informally, states that if we take many, many independent copies of the same random variable (dice, coin, whatever), and add them all up, their shape (or "distribution" more formally) all tend to look the same (like a bell curve). This unifying shape is called the "normal distribution" or the "Gaussian distribution".

4. **Expectation of a Random Variable.**

The expectation of a random variable X is defined to be

$$\mathbf{Exp}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbf{Pr}[\omega] = \sum_{x \in \text{range}(X)} x \cdot \mathbf{Pr}[X = x]$$

Remark: *Do you see why the second summation equals the first summation? Here is how:*

$$\begin{aligned}
\sum_{x\in\text{range}(X)} x\cdot\mathbf{Pr}[X=x] &= \sum_{x\in\text{range}(X)} x\cdot\left(\sum_{\omega\in\Omega:X(\omega)=x}\mathbf{Pr}[\omega]\right) && \text{by definition of } \mathbf{Pr}[X=x]\\
&= \sum_{\omega\in\Omega}\mathbf{Pr}[\omega]\cdot\sum_{x:X(\omega)=x} x && \text{A swap of summations}\\
&= \sum_{\omega\in\Omega}\mathbf{Pr}[\omega]\cdot X(\omega) && \text{There is only one } x \text{ which } X(\omega) \text{ evaluates to}
\end{aligned}$$

I hope this didn't confuse you more ...

Remark: *The expectation is therefore often thought of as an inner-product (aka dot-product) of two vectors. These vectors have $|\Omega|$ dimensions. One vector is $(X(\omega):\omega\in\Omega)$, and the other is $(\mathbf{Pr}[\omega]:\omega\in\Omega)$. This dot-product view is often useful (although, sadly, we may not see its ramifications in this course).*

Examples:

- *We toss a fair coin. $X(\text{heads})=0$ and $X(\text{tails})=1$. This is a $\{0,1\}$-random variable, or a Boolean random variable. Also called a* Bernoulli *random variable.*

  $$\mathbf{Exp}[X]=0\cdot\mathbf{Pr}[X=0]+1\cdot\mathbf{Pr}[X=1]=1/2$$

  Indeed, if the coin were not fair, and the probability that tails would come with probability p, then $\mathbf{Exp}[X]=p$.
- *We roll a fair die. X takes the value on the face of the die.*

  $$\mathbf{Exp}[X]=1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6}=3.5$$
- *We roll two fair dice. X takes the value of the sum. In this case, $X=Y+Z$ where Y,Z are random variables of the kind from the previous bullet point.*

  This is requires a little work. The range of X is $\{2,3,4,5,6,7,8,9,10,11,12\}$. We can calculate the probabilities for each (remember, it is not uniform), and then do the calculation. ✍

  Exercise: *Please do the calculation.*

  We get the answer 7. Did you?
- *We toss a fair coin 100 times. Z is the number of heads.*

  This is a lot more work. First, we observe the $\text{range}(Z)=\{0,1,2,\ldots,100\}$. Then, we try to figure out $\mathbf{Pr}[Z=k]$. This is $\frac{1}{2^{100}}\cdot\binom{100}{k}$. (Do you see how? There are 2^{100} possible outcomes, each equally likely coz the coins are fair, and $\binom{100}{k}$ have exactly k heads.). Therefore,

  $$\mathbf{Exp}[Z]=\sum_{k=0}^{100}k\cdot\binom{100}{k}\cdot\frac{1}{2^{100}}$$

  Phew!

- *Given any event $\mathcal{E}$, there is an associated random variable called the* indicator random variable *denoted as $\mathbf{1}_{\mathcal{E}}$, where $\mathbf{1}_{\mathcal{E}}(\omega) = 1$ if $\omega \in \mathcal{E}$, and 0 otherwise.*

  $$\mathbf{Exp}[\mathbf{1}_{\mathcal{E}}] = 0 \cdot \mathbf{Pr}[\neg\mathcal{E}] + 1 \cdot \mathbf{Pr}[\mathcal{E}] = \mathbf{Pr}[\mathcal{E}]$$

  This is quite important. Why? Because it turns a probability calculation (the RHS) into an expectation calculation. As we show below, calculating expectations is often easier than calculating probabilities. ✍

Exercise: *Suppose you have a fair coin. Construct the following random variable Z whose range is $\mathbb{N}$. You keep tossing the fair coin till you get a heads. Z is the number of times you have tossed the coin. What is $\mathbf{Exp}[Z]$? To do this, figure out what is $\mathbf{Pr}[Z = k]$. Then write the expectation as a sum. Then see if you can simplify the sum.*

5. **Multiplication by a scalar.** If X is a random variable, and c is a "scalar" (a constant), then $Z = c \cdot X$ is another random variable. $\mathbf{Exp}[c \cdot X] = c \cdot \mathbf{Exp}[X]$. ✍

   Exercise: *Prove this.*

6. **Linearity of Expectation.** This is one of the most powerful equations in all of probability. Literally. It states the following. It literally has a four line proof.

   Theorem 1. For any two random variables X and Y, let $Z := X + Y$. Then,

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

   Proof.

   $$\begin{aligned}
   \mathbf{Exp}[Z] &= \sum_{\omega\in\Omega} Z(\omega)\mathbf{Pr}[\omega] && \text{Definition of Expectation}\\
   &= \sum_{\omega\in\Omega} (X(\omega) + Y(\omega))\,\mathbf{Pr}[\omega] && \text{Definition of } Z\\
   &= \sum_{\omega\in\Omega} X(\omega)\mathbf{Pr}[\omega] + \sum_{\omega\in\Omega} Y(\omega)\cdot\mathbf{Pr}[\omega] && \text{Distributivity}\\
   &= \mathbf{Exp}[X] + \mathbf{Exp}[Y] && \text{Definition of Expectation}
   \end{aligned}$$

   □

   As a corollary, we get:

   Theorem 2. For any k random variables $X_1, X_2, \ldots, X_k$,

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

   Examples of applications.

(a) *We roll two fair dice. X takes the value of the sum. In this case, $X = Y + Z$ where Y, Z are random variables of the kind from the previous bullet point.*

Tailor-made application. $\mathbf{Exp}[Y] = \mathbf{Exp}[Z] = 3.5$, the expected value of a single roll of a die. Thus, $\mathbf{Exp}[X] = \mathbf{Exp}[Y + Z] = 7$ by linearity of expectation.

(b) *We have a biased coin which lands heads with probability p. We toss it 100 times. Let Z be the number of heads we see. What is $\mathbf{Exp}[Z]$? Note that earlier we had the question for $p = 0.5$.*

> **Remark:** *Try doing this the "first-principle" way. That is, for each $0 \le k \le 100$, figure out the probability $\mathbf{Pr}[X = k]$ (that is, the probability we get exactly k heads), and then sum $\sum_{k=0}^{100} k \cdot \mathbf{Pr}[X = k]$. Please try it; feel the sweat needed to do this. It will make you appreciate the next three lines more!*

Define new random variables; define X_i to take the value 1 if the ith toss is heads, and 0 otherwise. Note, $X = X_1 + X_2 + \cdots + X_{100}$. Note, $\mathbf{Exp}[X_i] = p$ (it is a Bernoulli random variable). Thus, linearity of expectation gives $\mathbf{Exp}[X] = 100p$.

(c) *n people checked in their hats, but on their way out, were handed back hats randomly. What is the expected number of people who get their correct hats?*

Define X_i to be 1 if the ith person gets his or her back correctly. What is $\mathbf{Exp}[X_i]$? It is $1/n$; it is the probability that $\sigma(i) = i$ for a random ordering σ. This question was there in the UGP. Let $Z = \sum_{i=1}^{n} X_i$. Note, Z is the number of people who get their correct hats. By linearity of expectation, $\mathbf{Exp}[Z] = 1$.

(d) *In a party of n people there are some pairs of people who are friends, and some pairs who are not. In all there are m pairs of friends. The host randomly divides the party by taking each person and sending them left or right at the toss of a fair coin. How many friends, in expectation, are sundered apart?*

> **Remark:** *In terms of graphs (which we will see soon) the question is: a graph with m edges is randomly partitioned. How many edges, in expectation, have endpoints in different parts?*

For each pair of friends (u, v), define X_{uv} which takes the value 1 if u and v are split, and takes the value 0 if u and v are not split. The probability u and v are split is $1/2$ (either u is sent left, v is sent right, or vice-versa – do you see this?). Thus, $\mathbf{Exp}[X_{uv}] = 1/2$. Define $Z = \sum_{(u,v)\text{: friends}} X_{uv}$; Z is the number of friends sent apart. $\mathbf{Exp}[Z] = \sum_{(u,v)\text{: friends}} \mathbf{Exp}[X_{uv}] = m/2$. In expectation, half the friendships are sundered apart.

(e) *In an ordering σ of $(1, 2, \ldots, n)$, an inversion is a pair $i < j$ such that $\sigma(i) > \sigma(j)$. How many inversions, in expectation, are there in a random permutation?*

Let σ be a random permutation. Define the *indicator random variable* X_{ij} for $i < j$, which takes the value 1 if $\sigma(i) > \sigma(j)$, and 0 otherwise. Note that $\mathbf{Pr}[X_{ij} = 1] = \frac{1}{2}$; there are equally many orderings with $\sigma(i) > \sigma(j)$ as $\sigma(i) < \sigma(j)$. Now note that $Z = \sum_{i=1}^{n} \sum_{j>i} X_{ij}$ is the number of inversions in σ. Thus, $\mathbf{Exp}[Z] = \sum_{i=1}^{n} \sum_{j>n} \mathbf{Exp}[X_{ij}] = \frac{1}{2} \cdot \frac{n(n-1)}{2}$.

7. **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 3. 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 4. 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, $\mathbf{Exp}[Z] = (3.5)^n$.

8. **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 5. $\mathbf{Var}[X] = \mathbf{Exp}[X^2] - (\mathbf{Exp}[X])^2$.

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:

Theorem 6. 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

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

Theorem 7. 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 8. For any two ***independent*** random variables X and Y, let $Z := X + Y$. Then,

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

Proof.

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

In the last equality, due to independence, we get that $2\left(\mathbf{Exp}[XY] - \mathbf{Exp}[X]\mathbf{Exp}[Y]\right) = 0$. □

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

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

9. **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 10. (Markov's Inequality)

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

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

Proof. By definition of expectation, we have

$$\mathbf{Exp}[X] = \sum_{x \in \mathsf{range}(X)} x \cdot \mathbf{Pr}[X = x] = \sum_{0 \leq x < t} x \cdot \mathbf{Pr}[X = x] + \sum_{x \geq t} x \cdot \mathbf{Pr}[X = x]$$

The first summation $\sum_{0 \leq x < t} x \cdot \mathbf{Pr}[X = x] \geq 0$. All terms are non-negative. The second summation is $\sum_{x \geq t} x \cdot \mathbf{Pr}[X = x] \geq t \cdot \sum_{x \geq t} \mathbf{Pr}[X = x] = t \cdot \mathbf{Pr}[X \geq t]$.

Putting it all together, we get

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

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

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

Proof. We first note that

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

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

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

□

Theorem 12. 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 9), 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*?*