Lecture 13 (Part 2): Deviation from mean: Markov’s inequality, variance and its properties, Chebyshev’s inequality

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 20 of the Math for CS book.

So far we have learned how to model random experiments/phenomenon mathematically using probability theory: define what the outcomes are and the sample space $\Omega$, define an appropriate probability measure on $\Omega$, identify relevant events, define useful random variables, and compute expected value of random variables. But simply computing the expected value of a random variable does provide any guarantees about the “typical” values that a random variable takes. We want to be able to make predictions about the kind of values the random variable will take, and we want those predictions to hold with high probability.

Formally speaking, suppose we have a random variable $X$ defined on some probability space $(\Omega, P)$. We want to be able make statements of the form:

“$X$ will be in the range $[a, b]$ with probability at least $\alpha$”,

or

$P(a \leq X \leq b) \geq \alpha$.

Ideally, we would want $a$ and $b$ to be as close to each other as possible, i.e. the interval $[a, b]$ to be as small as possible, and simultaneously, the probability $\alpha$ to be as high as possible (for example, $\alpha \geq 0.99$).

For example, suppose you design a randomized algorithm for a problem whose running time $T$ is a random variable, i.e. it depends on some random choices the algorithm makes when it’s run on any given input. Also suppose that $E[T]$, the expected running time of the algorithm, is small, and so you want to be able to make a statement of the form “with high probability, the running time $T$ of the algorithm is “close” to the expected running time $E[T]$.” because that shows that most of the time your algorithm runs very fast (you will encounter such situations when you study and design algorithms in CS 344).

In this lecture, we will see how one can provide guarantees on the typical value that random variables take. In particular, we will show that for many distributions it’s highly unlikely that a random variable takes a value that’s “too far” from it’s expected value, i.e. it’s highly unlikely that the random variable deviates too much from it’s expected value!

1 Markov’s inequality

Suppose $X$ is a random variable defined on a probability space $(\Omega, P)$ such that $X$ always takes nonnegative values, i.e. for every $a \in \text{Range}(X)$, $a \geq 0$. We call $X$ a nonnegative random variable.
Theorem 1 (Markov’s inequality). Let $X$ be a nonnegative random variable with expectation $\mathbb{E}[X]$, and let $a \geq 0$ be a constant. Then

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$  

Equivalently, the probability that $X$ becomes $k\mathbb{E}[X]$ (for some $k \geq 1$) is at most $\frac{1}{k}$:

$$P(X \geq k\mathbb{E}[X]) \leq \frac{1}{k}.$$  

Proof. We will prove the first statement of the theorem. The second statement follows from the first by setting $a = k\mathbb{E}[X]$. Using the second definition of $\mathbb{E}[X]$:

$$\mathbb{E}[X] = \sum_{i=0}^{a-1} i P(X = b) = \sum_{i=0}^{a-1} i P(X = i) + \sum_{j=a}^{\infty} j P(X = j).$$

Clearly, $\sum_{i=0}^{a-1} i P(X = i) \geq 0$, and so

$$\mathbb{E}[X] \geq \sum_{j=a}^{\infty} j P(X = j) \geq \sum_{j=a}^{\infty} a P(X = j) = a \left( \sum_{j=a}^{\infty} P(X = j) \right) = a P(X \geq a),$$

where the last step follows from the fact that $P(X \geq a) = \sum_{j=a}^{\infty} P(X = j)$. Thus, this implies that

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$  

Question. Consider the experiment of tossing a biased coin (probability of heads is $1/3$) 100 times. What is the probability that the number of heads observed is less than 70?

Proof. This experiment can be modelled using a binomial random variable $X$ with $n = 100$ and $p = 1/3$ (why?), and so $\mathbb{E}[X] = np = \frac{100}{3} \approx 33.33$. Using Markov’s inequality,

$$P(X \geq 70) \leq \frac{\mathbb{E}[X]}{70} \leq \frac{33.33}{70} \leq 0.48.$$  

So, with probability at least 0.52, the number of heads is at most 70.

2 Variance

A useful quantity to understand that helps prove somewhat better guarantees for random variables is the average deviation of a random variable from its mean\(^1\).

Of course, if we define deviation from the expected value naively, i.e. as $X - \mathbb{E}[X]$, then the average deviation will be zero: $\mathbb{E}[(X - \mathbb{E}[X])] = \mathbb{E}[X] - \mathbb{E}[X] = 0$. The reason that this naive definition becomes zero is because it doesn’t consider the magnitude of deviation and instead considers the

\(^1\)The expected value is also called the mean of a random variable
signed value of deviation. For example, if there is a random variable $X$ with range \{-1000, 1000\} such that $P(X = -1000) = P(X = 1000) = 1/2$, then $\mathbb{E}[X] = 0$. Now if we define deviation as $X - \mathbb{E}[X]$, then when $X = -1000$, the deviation becomes $-1000$, and when $X = 1000$, the deviation becomes $1000$, and so the deviations cancel each other out when we average them.

Intuitively, what we would want as the average deviation in this case would be $\frac{-1000+1000}{2} = 1000$. This suggests we define deviation as $|X - \mathbb{E}[X]|$, where $|\cdot|$ is the absolute value function. In this case, the average/expected deviation would be $\mathbb{E}[|X - \mathbb{E}[X]|]$. While this is a sensible quantity to study, it turns out that it’s not very friendly to mathematical manipulation and analysis. Instead what’s preferred is to consider the square of deviation, i.e. $(X - \mathbb{E}[X])^2$, and then compute its average/expected value. This is called the variance of $X$.

**Definition 2 (Variance).** The variance of a random variable $X$, denoted by $\text{Var}(X)$, with expected value $\mathbb{E}[X]$ is the average squared deviation of $X$ from $\mathbb{E}[X]$:

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

There is an alternate expression that we can derive for the variance that makes computing variance easy in many cases:

**Lemma 3.** For a random variable $X$ with mean $\mathbb{E}[X]$,

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

**Proof.** Notice from the definition of variance that

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])] = \text{Cov}(X, X),$$

using the second definition of covariance. Using the first definition of covariance we know that

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Thus, $\text{Cov}(X, X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, and so

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

Notice that variance is the average of the square of the deviation of a random variable from its mean, and so to bring it down to the same scale as the mean, it’s often useful to consider the square root of the variance. The square root of the variance is called the standard deviation, and is denoted be $\sigma(X)$ for a random variable $X$, i.e.

$$\sigma(X) = \sqrt{\text{Var}(X)}.$$
The quantity $\mathbb{E}[X^2]$ is called the second moment of $X$. In general, one can study $\mathbb{E}[X^k]$ for $k \geq 1$, and this quantity is called the $k^{th}$ moment of $X$, and can be computed in a similar way:

$$\mathbb{E}[X^k] = \sum_{a \in \text{Range}(X)} a^k P(X = a).$$

**Question.** Let $X$ be the number rolled by a fair dice. Find $\text{Var}(X)$.

**Proof.** Recall that $\mathbb{E}[X] = 3.5$. We will use the formula

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

We already know the second term, so we only need to compute $\mathbb{E}[X^2]$. We will use the formula for the second moment to do this (see above):

$$\mathbb{E}[X^2] = \sum_{a \in \text{Range}(X)} a^2 P(X = a) = \sum_{i=1}^{6} i^2 P(X = i) = \sum_{i=1}^{6} i^2 \frac{1}{6} = \frac{6}{6} \sum_{i=1}^{6} i^2.$$

The formula for the sum of the squares of numbers from 1 to $n$ is

$$\frac{n(n + 1)(2n + 1)}{6}.$$  

Thus,

$$\mathbb{E}[X^2] = \frac{6(6+1)(12+1)}{6} = \frac{91}{6}.$$  

Thus,

$$\text{Var}(X) = \frac{91}{6} - (3.5)^2 \approx 2.91.$$  

\[ \square \]

### 2.1 Properties of variance

Here are some properties of variance:

1. If $X : \Omega \rightarrow \mathbb{R}$ is a constant random variable, i.e. $\forall \omega \in \Omega, X(\omega) = c$, then $\text{Var}(X) = 0$ (check this using the definition of variance). In other words, for any constant $c \in \mathbb{R}$, $\text{Var}(c) = 0$.

2. Let $a \in \mathbb{R}$ be a constant and $X$ be a random variable, then $\text{Var}(aX) = a^2 \text{Var}(X)$. This follows because $\text{Var}(aX) = \mathbb{E}[(aX)^2] - (\mathbb{E}[aX])^2 = \mathbb{E}[a^2X^2] - (a\mathbb{E}[X])^2 = a^2 \mathbb{E}[X^2] - a^2(\mathbb{E}[X])^2 = a^2(\mathbb{E}[X^2] - (\mathbb{E}[X])^2) = a^2 \text{Var}(X)$.

3. This one is very useful: if $X_1, \ldots, X_n$ are pairwise independent random variables, then

$$\text{Var}(X_1 + \ldots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n).$$

Note that we only need $X_1, \ldots, X_n$ to be pairwise independent\(^2\) and not mutually independent (the latter implies the former). To get an idea as to why this is true, let's look at the case when $n = 2$ (the general case then follows from induction). Let's look at $\text{Var}(X + Y)$ for

\(^2\)It's important to contrast this with linearity of expectation where $X_1, \ldots, X_n$ did not need any independence whatsoever.
two independent random variables $X$ and $Y$. Using the definition of variance and linearity of expectation,

$$\text{Var}(X + Y) = \mathbb{E}[(X + Y)^2] - (\mathbb{E}[X + Y])^2 = \mathbb{E}[X^2 + Y^2 + 2XY] - (\mathbb{E}[X] + \mathbb{E}[Y])^2$$

$$= \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY] - ((\mathbb{E}[X])^2 + (\mathbb{E}[Y])^2 + 2\mathbb{E}[X]\mathbb{E}[Y])$$

Rearranging terms, we get

$$\text{Var}(X + Y) = (\mathbb{E}[X^2] - (\mathbb{E}[X])^2) + (\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y])$$

$$\implies \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y).$$

Since $X$ and $Y$ are independent, $\text{Cov}(X, Y) = 0$, and so

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

4. In general, we have that, for constants $a_1, \ldots, a_n, b \in \mathbb{R}$, and pairwise independent random variables $X_1, \ldots, X_n$,

$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i + b\right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i).$$

**Question.** A fair dice is rolled 100 times. Let $X$ be the sum of the numbers observed. What is $\text{Var}(X)$?

**Proof.** Let $X_i$ be the number observed in the $i^{\text{th}}$ roll. Then $X = \sum_{i=1}^{100} X_i$. Using the properties of variance,

$$\text{Var}(X) = \sum_{i=1}^{100} \text{Var}(X_i).$$

We computed the variance of a single dice roll at the end of the previous subsection: $\text{Var}(X_i) = \frac{91}{6} - \left(\frac{3}{2}\right)^2$, and so

$$\text{Var}(X) = 100 \times \left(\frac{91}{6} - \left(\frac{3}{2}\right)^2\right) \approx 291.67$$

\[\square\]

### 2.2 Variance of some common distributions

We will now revisit some common distributions/random variables to see what their respective variance are:

#### 2.2.1 Bernoulli

Recall that a Bernoulli random variable $X$ is 1 with probability $p$ and 0 with probability $1 - p$, and $\mathbb{E}[X] = p$. Let’s compute $\text{Var}(X)$:

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

Note that $\mathbb{E}[X^2] = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = p$, and so

$$\text{Var}(X) = p - p^2 = p(1 - p).$$
2.2.2 Binomial

Recall that a Binomial random variable is basically a sum of \( n \) independent Bernoulli random variables with parameter \( p \), i.e. \( X = X_1 + \ldots + X_n \), where each \( X_i \) is Bernoulli with parameter \( p \), i.e. \( P(X = 1) = p \) and \( P(X = 0) = (1 - p) \). Using the properties of variance, we know that

\[
\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i).
\]

Since each \( X_i \) is Bernoulli with parameter \( p \), we have that \( \text{Var}(X_i) = p(1-p) \), and so

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

Thus, the variance of a binomial random variable with parameter \( p \) is \( np(1-p) \).

2.2.3 Geometric

Recall that a geometric random variable with parameter \( p \) is the basically the total number of coin flips in an experiment in which we keep tossing a biased coin (i.e., the probability of heads is \( p \)) till we see a heads. Using the law of total expectation, we proved that if \( X \) is a geometric random variable with parameter \( p \), then \( \mathbb{E}[X] = \frac{1}{p} \). Using a similar proof, one can show that

\[
\text{Var}(X) = \frac{1-p}{p^2}.
\]

To see how, we use the fact that \( \text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[X^2] - \frac{1}{p^2} \). So we basically have to compute \( \mathbb{E}[X^2] \) for a geometric random variable with parameter \( p \). Let \( A \) be the event that the first coin toss is heads. Then using the the law of total expectation:

\[
\mathbb{E}[X^2] = P(A)\mathbb{E}[X^2|A] + P(A^c)\mathbb{E}[X^2|A^c].
\]

It’s easy to see that \( \mathbb{E}[X^2|A] = 1 \) (why?). Also, \( P(A) = p \) and \( P(A^c) = 1 - p \). How about \( \mathbb{E}[X^2|A^c] \)? Once we know that the first coin toss is tails we basically have to start over, and if \( Y \) is the total number of tosses, \textit{starting with the second coin toss} (i.e., not counting/including the first coin toss), needed to see a heads, then

\[
\mathbb{E}[X^2|A^c] = \mathbb{E}[(1+Y)^2].
\]

The key observation is that \( Y \) is also a geometric random variable with parameter \( p \), and in fact, has the same probability mass function as \( X \), and so \( \mathbb{E}[(1+Y)^2] = \mathbb{E}[(1+X)^2] \). Putting everything together, we get

\[
\mathbb{E}[X^2] = p \cdot 1 + (1-p)\mathbb{E}[(1+X)^2] = p + (1-p) \left( \mathbb{E}[X^2 + 1 + 2X] \right)
\]

\[
= p + (1-p) \left( \mathbb{E}[X^2] + 1 + 2\mathbb{E}[X] \right) = p + (1-p)\mathbb{E}[X^2] + (1-p) + 2(1-p)\mathbb{E}[X].
\]

Using the fact that \( \mathbb{E}[X] = 1/p \), and rearranging terms we get

\[
p\mathbb{E}[X^2] = \frac{2-p}{p} \implies \mathbb{E}[X^2] = \frac{2-p}{p^2}.
\]

This implies that

\[
\text{Var}(X) = \frac{2-p}{p^2} - \frac{1}{p} = \frac{1-p}{p^2}.
\]
3 Chebyshev’s inequality

Chebyshev’s inequality gives a much stronger guarantee on how much a random variable can deviate from its expectation if we know the variance of the random variable.

**Theorem 4** (Chebyshev’s inequality). Let $X$ be a random variable with variance $\text{Var}(X)$ and expectation $\mathbb{E}[X]$. Then

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

Equivalently,

$$P(|X - \mathbb{E}[X]| \geq k \cdot \text{Var}(X)) \leq \frac{1}{k^2}.$$  

**Proof.** As in the case of Markov’s inequality, we will prove the first statement, because the second statement follows from the first one by setting $a = k \mathbb{E}[X]$. Let $Y$ be equal to the square of the deviation of $X$ from $\mathbb{E}[X]$:

$$Y = (X - \mathbb{E}[X])^2$$

Note that $Y$ is also a random variable (why?), and is in fact a nonnegative random variable (why?), and from the definition of variance, we know that

$$\mathbb{E}[Y] = \text{Var}(X).$$

Also, note that

$$P(|X - \mathbb{E}[X]| \geq a) = P((X - \mathbb{E}[X])^2 \geq a^2) = P(Y \geq a^2),$$

where the first equality is true since we can just square both sides of the inequality. Thus, if we can prove an upper bound on $P(Y \geq a^2)$, that will also prove an upper bound on $P(|X - \mathbb{E}[X]| \geq a)$. We can now use Markov’s inequality to bound $P(Y \geq a^2)$ (we can use Markov’s because $Y$ is a nonnegative random variable):

$$P(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a^2} = \frac{\text{Var}(X)}{a^2}.$$  

Thus,

$$P(|X - \mathbb{E}[X]| \geq a) = P(Y \geq a^2) \leq \frac{\text{Var}(X)}{a^2}.$$  

□

If the variance of $X$ is small, then Chebyshev’s inequality provides a much stronger guarantee as compared to Markov’s inequality:

- With Markov’s inequality you can make predictions of the form “$X$ is between $0$ and $k \mathbb{E}[X]$” with probability at least $1 - \frac{1}{k}$,

- whereas with Chebyshev’s inequality you can make predictions of the form “$X$ is between $\mathbb{E}[X] - k \cdot \text{Var}(X)$ and $\mathbb{E}[X] + k \cdot \text{Var}(X)$ with probability at least $1 - \frac{1}{k^2}$.”
Note that firstly, the probability of the predictions made by Chebyshev’s is much better (larger) than that of Markov’s: $k \geq 0$ as an integer, and then clearly, $1 - \frac{1}{k} < 1 - \frac{1}{k^2}$. For example, when $k = 10$, $1 - \frac{1}{10} = 0.90$, while $1 - \frac{1}{100} = 0.99!$

Secondly, the range of the prediction made by Chebyshev’s is smaller than that of Markov’s: while Markov’s says that $X$ can be anywhere between 0 and $k\mathbb{E}[X]$ (which is a pretty big range!), Chebyshev says that $X$ has to be between $\mathbb{E}[X] - k \cdot \text{Var}(X)$ and $\mathbb{E}[X] + k \cdot \text{Var}(X)$ which can be pretty small if $\text{Var}(X)$ is small.

Let’s revisit an earlier problem and try to solve it using Chebyshev’s inequality:

**Question.** Consider the experiment of tossing a biased coin (probability of heads is $1/3$) 100 times. What is the probability that the number of heads observed is less than 70?

**Proof.** The number of heads, say we call it $X$, is a binomial random variable with parameter $p$, and so $\mathbb{E}[X] = \frac{100}{3} \approx 33.33$, and $\text{Var}(X) = 100 \cdot \frac{1}{3} \left(1 - \frac{1}{3}\right) = \frac{200}{9} \approx 22.22$. We want to upper bound $P(X > 70)$ which is the same as $P(X - \mathbb{E}[X] > 70 - 33.33) = P(X - \mathbb{E}[X] > 36.67) \leq P(|X - \mathbb{E}[X]| > 36.67)$. Using Chebyshev’s inequality,

$$P(|X - \mathbb{E}[X]| > 36.67) \leq \frac{\text{Var}(X)}{(36.67)^2} = \frac{22.22}{(36.67)^2} \leq 0.0165.$$

This means that the probability the numbers of heads is more than 70 is at most 0.016, and so with probability greater than 0.98, $X$ will be less than 70! This is a much better guarantee than what we got using Markov’s inequality.