Mathematics for Computer Science (30470023) Spring 2020

Lecture 4 — 2020.3.9

Instructor: Prof. Andrew C. Yao *Scribes: Zhiyuan Fan, Yiding Zhang*

1 Overview

In this lecture, we further talk about the expectation of random variables and presenting a useful formula for computing the expectation. Then we introduce the concept of variance as well as some basic techniques for computing it. We also introduce Chebyshev's inequality and Chernoff's bound that are related to the variance.

2 Expectation

We introduced the concept of expectation and its linearity in lectures before. In this lecture, we start with a useful formula to calculate the expectation.

Proposition 2.1. Suppose random variable X over U mapped to natural numbers, $X : U \to \mathbb{N}$. Then,

$$\mathbb{E}[X] = \sum_{i \geq 0} \Pr\{X > i\}.$$

Proof. From the definition of expection,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i \geq 1} i \Pr\{X = i\} \\
&= \sum_{i \geq 1} i(\Pr\{X > i-1)\} - \Pr\{X > i\}) \\
&= \sum_{i \geq 0} (i+1) \Pr\{X > i\} - \sum_{i \geq 1} i \Pr\{X > i\} \\
&= \sum_{i \geq 0} \Pr\{X > i\}.
\end{aligned}$$

□

Example 2.2 (Expected length of cycles). Suppose that u is a random permutation of n (e.g., $n = 7$ and $u = (1\ 3\ 5\ 7)(2\ 4\ 6)$). Let $X(u)$ be the length of cycle that the number 1 is on (and $X(u) = 4$ in the example above). Now we want to calculate $\mathbb{E}(X)$.

By the formula we know that $\mathbb{E}(X) = \sum_{i \geq 0} \Pr\{X > i\}$, and we can easily get

$$\Pr\{X > i\} = \begin{cases} \frac{n-i}{n} & 0 \leq i < n \\ 0 & i \geq n \end{cases}$$

from the fact that $\Pr\{X > 0\} = 1$ together with the recurrence relation from the chain rule that

$$\Pr\{X > i\} = \Pr\{X > i-1\} \cdot \Pr\{X > i \mid X > i-1\} = \frac{n-i}{n-(i-1)} \Pr\{X > i-1\}.$$

So we have

$$\mathbb{E}(X) = \sum_{i=0}^{n-1} \frac{n-i}{n} = 1 + \frac{1}{n} \sum_{i=1}^{n-1} i = \frac{n+1}{2}.$$

3 Conditional Expection

Since the random variables can be seen as a generalization of events. We can define conditional expection as well as conditional probablity.

Definition 3.1. Let X be a random variable and let T be a event. The conditional expection of X conditioned on T is defined by

$$\mathbb{E}[X \mid T] = \frac{\sum_{u \in T} X(u)p(u)}{\Pr\{T\}}.$$

Particularily, $\mathbb{E}[X \mid T] = 0$ when $\Pr\{T\} = 0$. ♢

For an event T on a probablity space $\mathbb{P} = (U, p)$, the conditional expection can be seen as the expection restricted on $\mathbb{P}' = (T, p')$ where $p'(u) = \frac{p(u)}{\Pr\{T\}}$ for all $u \in T$.

Theorem 3.2 (Distributric Law of Random Variables). Suppose U is a disjoint union of event $T_1, T_2, ..., T_n$. Then,

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X \mid T_i] \Pr\{T_i\}.$$

Proof. Trivial. □

4 Variance

Before introducing the concept of variance, we give an example to show that just knowing the expected value is not enough.

Example 4.1 (Lottery tickets). Suppose that we have two kinds of lottery tickets with price ¥50. The pay-off of lottery ticket #1 (denoted by X) is ¥40 with probability 45%, ¥100 with 30%, ¥200 with 10%, and ¥500 with 5%; the pay-off of lottery ticket #2 (denoted by Y) is ¥10^7 with probability 10^{-5}, and ¥10 with probability $1 - 10^{-5}$. If someone wants to buy one of them, then which one will he/she choose?

If we only consider the expected value of pay-off, we have $\mathbb{E}(X) = 93$ and $\mathbb{E}(Y) \approx 110$, and thus buying ticket 2 has higher expected pay-off. But probably people do not want to buy ticket #2

(we can suppose that everyone can only buy one ticket). Although ticket #2 has higher expected value, the spread is very large, and buyers may not be lucky enough to get ¥10^7. That is to say, the expected value tells us no information about the spread and the probable value of a random variable. To solve this problem, the concept of variance is designed to quantify the concept of "probable range".

Definition 4.2. Let X be a random variable. The variance of X is defined by

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

(Question: Why we do not use $\mathbb{E}[|X - \mathbb{E}[X]|]$ or something else?) ♢

Note that compared to the random variable X, $\mathrm{Var}(X)$ is squared in dimension. For example, if X is the running time in seconds, $\mathrm{Var}(X)$ is measured in units of square seconds. So we introduce the concept of standard deviation to avoid the square in dimension.

Definition 4.3. Let X be a random variable. The standard deviation of X is defined by

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}.$$

4.1 Techniques for computing the variance

Proposition 4.4.

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

Proof. According to the linearity of expection,

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}^2[X]] \\ &= \mathbb{E}[X^2] - \mathbb{E}^2[X].\end{aligned}$$

□

Proposition 4.5. Let X denote the sum of random variables $X_1, X_2, ..., X_n$. Then,

$$\mathrm{Var}(X) = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} (\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]).$$

Proof. According to the linearity of expection,

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}^2[X] \\ &= \sum_i \mathbb{E}[X_i^2] + \sum_{i \neq j} \mathbb{E}[X_i X_j] - \sum_i \mathbb{E}^2[X_i] - \sum_{i \neq j} \mathbb{E}[X_i]\mathbb{E}[X_j] \\ &= \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} (\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]).\end{aligned}$$

□

Note that the last term $\sum_{i\neq j}(\mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_j])$ is called the correlation term. In fact, $\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[XY]$ is usually called the *covariance* of X and Y, and it is often used to show the tendency in the linear relationship between two variables. The formal definition is shown below

Definition 4.6. For two random variables X, Y with finite second moments, their covariance $Cov(X, Y)$ is defined as

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

Here we present an example of calculating and utilizing the variance of a random variable.

Example 4.7 (0-1 tosses). Consider n independent 0-1 tosses with bias b for value 1. Let X be the random variable denoting the number of heads, and $X_i (1 \leq i \leq n)$ be the corresponding random variable for every single toss. Then we have $X = \sum_{i=1}^{n} X_i$ and $\mathbb{E}(X_i) = b$. We also have $X_i^2 = X_i$ since X_i is 0-1 valued. The variance $\mathrm{Var}(X)$ can be calculated as follows:

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{i=1}^{n} \mathbb{E}(X_i^2) + \sum_{i\neq j} \mathbb{E}(X_iX_j) - \mathbb{E}(X)^2 \\
&= \sum_{i=1}^{n} \mathbb{E}(X_i) + n(n-1)\mathbb{E}(X_1X_2) - \mathbb{E}(X)^2 \\
&= nb + n(n-1)b^2 - n^2b^2 \\
&= nb(1-b).
\end{aligned}$$

That is to say, if we conduct a poll among $n = 1000$ people to make a decision and the answer is yes/no with bias $b \approx \frac{1}{2}$, the uncertainty (i.e., the percentage of probable error) is

$$\frac{\sigma(x)}{\mathbb{E}(X)} = \frac{\sqrt{b(1-b)}\sqrt{n}}{bn} \approx \frac{1}{\sqrt{n}} \approx 3\%.$$

4.2 Bounds of random variables

Theorem 4.8 (Markov Inequality). Suppose that X is a non-negative random variable. For any $a > 0$, we have

$$\Pr\{X \geq a\} \leq \frac{\mathbb{E}(X)}{a}.$$

Proof. Here we only prove the discrete version of this inequality. Suppose that X is a non-negative random variable over $\mathcal{U}$ where $\mathbb{P} = (\mathcal{U}, p)$ is the probability space. By the definition of expectation, we have

$$\begin{aligned}\mathbb{E}[X] &= \sum_{u\in\mathcal{U}} X(u)p(u)\\ &= \sum_{u\in\mathcal{U}\wedge X(u)<a} X(u)p(u) + \sum_{u\in\mathcal{U}\wedge X(u)\geq a} X(u)p(u)\\ &\geq \sum_{u\in\mathcal{U}\wedge X(u)\geq a} X(u)p(u)\\ &\geq \sum_{u\in\mathcal{U}\wedge X(u)\geq a} a\cdot p(u)\\ &= a\cdot \Pr\{X\geq a\}\end{aligned}$$

□

So we can get $\Pr\{X \geq a\} \leq \frac{\mathbb{E}(X)}{a}$. The continuous version can be proved similarly by changing "$\sum$" into "$\int$".

By applying Markov inequality, we can get Chebyshev's inequality which gives us an upper bound of a random variable's tail distribution.

Theorem 4.9 (Chebyshev Inequality). Suppose X is a random variable, for any $c > 0$,

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

Proof. By applying Markov inequality,

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X-\mathbb{E}[X])^2]\\ &\geq c^2 Var(X)\Pr\{(X-\mathbb{E}[X])^2 \geq c^2 Var(X)\}\\ &= c^2 Var(X)\Pr\{|X-\mathbb{E}[X]|\geq c\sigma(X)\}.\end{aligned}$$

□

Therefore,

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

The Chebyshev Inequality give us a general bound of the tail distribution. However, as we can see later, the bound is usually not tight and we have some better bounds in some special cases.

Example 4.10 (Tossing a fair coin). Suppose that we toss a fair coin n times and $n = 10000$, then $\mathbb{E}(X) = 5000$, $\sigma(X) = \frac{\sqrt{n}}{2} = 50$. Chebyshev inequality tells us that

$$\Pr\{|X - 5000| > 500\} \leq \frac{1}{10^2} = 1\%.$$

However, the true probability is much more smaller. The next theorem will tell us that it is less than $2e^{-17}$.

Theorem 4.11 (Chernoff Bound). Suppose $X_1, X_2, ..., X_n$ are independent 01-random variables. Let X denote their sum and let $\mu = \mathbb{E}[X]$ denote the expection of X. Then we have

$$\Pr\{X > (1+\delta)\mu\} < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \text{ for } \delta > 0 \ ;$$
$$\Pr\{X < (1-\delta)\mu\} < \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} \text{ for } 0 < \delta < 1.$$

Since the above formula is often too heavy, we may use looser but more convenient bounds in practice,

$$\Pr\{X > (1+\delta)\mu\} < e^{-\frac{\delta^2\mu}{2}}$$
$$\Pr\{X < (1-\delta)\mu\} < e^{-\frac{\delta^2\mu}{2+\delta}}$$
$$\Pr\{|X-\mu| > \delta\mu\} < 2e^{-\frac{\delta^2\mu}{3}}.$$

♢