Mathematical Toolkit **Autumn 2016**

Lecture 12: November 8, 2016

Lecturer: Madhur Tulsiani

1 Computing expectations

For the next example, we consider an *infinite* sequence of independent coin tosses, with $\mathbb{P}[\text{heads}] = p$ for each coin.

Example 1.1 *Given, that* $\mathbb{P}[\textit{heads}] = p$*, what is* $\mathbb{E}[\#\textit{tosses till the first heads}]$*?*

We define Z as the number of tosses till the first heads. Let E be the event that the first toss is heads. Then we have,

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}[Z|E] \cdot \mathbb{P}[E] + \mathbb{E}[Z|\neg E] \cdot \mathbb{P}[\neg E] \\ &= 1 \cdot \mathbb{P}[E] + (1 + \mathbb{E}[Z]) \cdot (1 - p) \end{aligned}$$

Thus we have, $\mathbb{E}[Z] = \dfrac{1}{p}$.

The above is known as a *geometric random variable* with parameter p.

Remark 1.2 *As was pointed out in class, one cannot define a countable probability space for the infinite sequence of random variables corresponding to the coin tosses. However, if we just want define a space for* Z*, we can take* $\Omega = \mathbb{N}$ *and* $\mu(i) = (1-p)^{i-1} \cdot p$ *for* $i \geq 1$*.*

1.1 Coupon Collection

Consider the following problem: There are n kinds of items/coupons and at each time step we get one coupon chosen to be from one of the n types at random. All types are equally likely at each step and the choices at different time steps are independent. We define a random variable, T which is the time when we first have all the n types of coupons. Find $\mathbb{E}[T]$.

We can make the following claim:

$$T = \sum_{i=1}^{n} X_i \,,$$

where X_i is the time to get from the $i-1$ to the i types of coupons. Thus we have,

$$\mathbb{E}[T] = \sum_i \mathbb{E}[X_i]$$

Note that X_i is a geometric random variable with parameter $\frac{n-i+1}{n}$, since if we have $i-1$ type of coupons, X_i represents the time till we receive a coupon belonging to any one of the remaining $n-i+1$ types. Thus,

$$\mathbb{E}[X_i] \;=\; \frac{n}{n-i+1}\,.$$

Therefore,

$$\mathbb{E}[T] = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} = n \cdot H(n)$$

where $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$ is the n^{th} *harmonic number*. It is known (see Wikipedia for example) that $H_n = \ln n + \Theta(1)$. Thus, we have that $\mathbb{E}[T] = n \ln n + \Theta(n)$.

2 Inequalities

We will develop some inequalities which let us bound the probability of a random variable taking a value very far from its expectation.

2.1 Markov's Inequality

This is the most basic inequality we will use. This is useful if the only thing we know about a random variable is its expectation. It will also be useful to derive other inequalities later.

Lemma 2.1 (Markov's Inequality) *Let Z be non-negative variable. Then,*

$$\mathbb{P}[Z \geq t] \;\leq\; \frac{\mathbb{E}[Z]}{t}\,. \tag{1}$$

Proof: We start by writing

$$\mathbb{E}[Z] \;=\; \mathbb{E}\left[Z \cdot \left(\mathbb{1}_{\{Z \geq t\}} + \mathbb{1}_{\{Z < t\}}\right)\right]\,,$$

where $\mathbb{1}_{\{Z \geq t\}}$ denotes the function which is 1 when $Z \geq t$ and 0 otherwise (similarly for $\mathbb{1}_{\{Z<t\}}$). Using non-negativity of Z, we get

$$\mathbb{E}[Z] \;=\; \mathbb{E}\left[Z \cdot \left(\mathbb{1}_{\{Z \geq t\}} + \mathbb{1}_{\{Z < t\}}\right)\right] \;\geq\; \mathbb{E}\left[Z \cdot \mathbb{1}_{\{Z \geq t\}}\right] \;\geq\; \mathbb{E}\left[t \cdot \mathbb{1}_{\{Z \geq t\}}\right] \;=\; t \cdot \mathbb{P}[Z \geq t]\,,$$

which completes the proof. ∎

2.2 Chebyshev's Inequality

Recall that the variance of random variable X is defined as

$$\mathsf{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

Lemma 2.2 (Chebyshev's inequality) *Let Z be a random variable and let $\mu = \mathbb{E}[Z]$. Then,*

$$\mathbb{P}[|Z - \mu| \geq t] \;\leq\; \frac{\mathsf{Var}[Z]}{t^2} \;=\; \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{t^2}. \tag{2}$$

Proof: Consider the non-negative random variable $(Z - \mu)^2$. Applying Markov's inequality we have

$$\mathbb{P}[|Z - \mu| \geq t] \;=\; \mathbb{P}\left[(Z-\mu)^2 \geq t^2\right] \;\leq\; \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{t^2}.$$

■

3 Coin tosses revisited

An unbiased coin is tossed n times. Probability that head shows up in each toss is $\frac{1}{2}$. Let Z be a random variable for the number of heads that have showed up after n tosses. We also have random variables X for i^{th} coin toss, where $X_i = 1$ if head shows up in i^{th} toss and 0 otherwise.

So we have

$$Z = \sum_{i=1}^{n} X_i \quad \text{and} \quad \mathbb{E}[Z] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{n}{2}.$$

Let us now compare the kind of bounds we get using Markov's and Chebyshev's inequalities.

3.1 Application of Markov's inequality

Using Markov's inequality we have,

$$\mathbb{P}\left[Z \geq \frac{3n}{4}\right] \leq \frac{\mathbb{E}[Z]}{(3n/4)} \;\Rightarrow\; \mathbb{P}\left[Z \geq \frac{3n}{4}\right] \leq \frac{2}{3} \;\Rightarrow\; \mathbb{P}\left[Z - \frac{n}{2} \geq \frac{n}{4}\right] \leq \frac{2}{3}.$$

3.2 Application of Chebyshev's inequality

We want to show that Chebyshev's inequality gives a stronger bound on probability. For this we need to calculate the variance of Z. We do this calculation below in a way that applies in many other situations as well. We have

$$\mathsf{Var}\left[Z\right] \;=\; \mathbb{E}\left[Z^2\right] - \left(\mathbb{E}\left[Z\right]\right)^2 .$$

We observe that

$$\mathbb{E}\left[Z^2\right] \;=\; \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] \;=\; \mathbb{E}\left[\sum_{i,j} X_i X_j\right] \;=\; \sum_{i,j} \mathbb{E}\left[X_i X_j\right].$$

Similarly,

$$\left(\mathbb{E}\left[Z\right]\right)^2 \;=\; \left(\mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)\right]\right)^2 \;=\; \sum_{i,j} \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]$$

So we have

$$\begin{aligned}
\mathsf{Var}\left[Z\right] &= \sum_{i,j} \mathbb{E}\left[X_i X_j\right] - \sum_{i,j} \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right] \\
&= \sum_{i} \left(\mathbb{E}\left[X_i^2\right] - \left(\mathbb{E}\left[X_i\right]\right)^2\right) + \sum_{i \neq j} \left(\mathbb{E}\left[X_i, X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]\right) \\
&= \sum_{i} \mathsf{Var}\left[X_i\right] + \sum_{i \neq j} \mathsf{Cov}\left[X_i, X_j\right],
\end{aligned}$$

where $\mathsf{Cov}\left[X_i, X_j\right]$ denotes $\mathbb{E}\left[X_i \cdot X_j\right] - \mathbb{E}\left[X_i\right] \cdot \mathbb{E}\left[X_j\right]$. Since the coin tosses are independent, we have $\mathbb{E}\left[X_i X_j\right] = \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]$ and hence $\mathsf{Cov}\left[X_i, X_j\right] = 0$. This yields,

$$\mathsf{Var}\left[Z\right] \;=\; \sum_{i} \mathsf{Var}\left[X_i\right] \quad \text{for independent random variables } X_i . \tag{3}$$

Also $\mathsf{Var}\left[X_i\right] \;=\; \mathbb{E}\left[X_i^2\right] - \left(\mathbb{E}\left[X_i\right]\right)^2 \;=\; p - p^2$, where $p \;=\; \mathbb{P}\left[X_i = 1\right]$. Here $p \;=\; \frac{1}{2}$, so $\mathsf{Var}\left[X_i\right] = \frac{1}{4}$ and hence, $\mathsf{Var}\left[Z\right] = \frac{n}{4}$. Applying Chebyshev's inequality we have,

$$\mathbb{P}\left[\left|Z - \frac{n}{2}\right| \geq t\right] \;\leq\; \frac{n}{4t^2} .$$

Setting $t = n/2$ and $t = \sqrt{n}$, gives the following bounds

$$\mathbb{P}\left[\left|Z - \frac{n}{2}\right| \geq \frac{n}{4}\right] \;\leq\; \frac{4}{n} \quad \text{and} \quad \mathbb{P}\left[\left|Z - \frac{n}{2}\right| \geq \sqrt{n}\right] \;\leq\; \frac{1}{4}$$

Thus, Chebyshev's inequality gives a much stronger bound on a deviation of $n/4$ from the mean, and can also bound the probability of deviations as small as $\sqrt{n}$. In particular, it gives a non-trivial bound whenever the deviation is larger than $\sqrt{\mathsf{Var}\left[Z\right]}$, a quantity which is referred to as the *standard deviation* of the random variable Z.