P & C Recitation 4 – Chebyshev's Inequality

Misha Lavrov

February 10, 2012

Motivation

Variance and standard deviation supposedly measure the "spread" of the distribution of a random variable. What does that *mean*, exactly?

If it measures how close a variable tends to be to its mean, then we should be able to use it to, say, bound tail probabilities. So let's say that a variable X has mean μ, standard deviation σ, and with a probability of p, $|X - \mu| \geq d$ for some distance d. With this constraint, what is the smallest we can possibly make the standard deviation?

With probability p, X is outside the range $(\mu - d, \mu + d)$. But to minimize the standard deviation, we want X to be as close to the mean as possible. So say that $X = \mu + d$ with probability $p/2$ and $\mu - d$ with probability $p/2$. The rest of the time, X is inside the range $(\mu - d, \mu + d)$. So we can set it to μ in those cases.

What is the variance of X? $|X - \mu| = d$ with probability p, and 0 otherwise. So $(X - \mu)^2 = d^2$ with probability p, and $\sigma^2 = \mathrm{E}((X - \mu)^2) = d^2 p$.

This was the smallest possible case, so in general $\sigma^2 \geq d^2 p$. Solving for p, we get

$$\mathrm{P}(|X - \mu| \geq d) \leq \frac{\sigma^2}{d^2}.$$

And if we let $d = k\sigma$, we get what is known as Chebyshev's inequality:

$$\mathrm{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Careful Proof

Let's try to prove this in a more formal way. We will begin by proving a different inequality called Markov's inequality; this would be bad style except you've already been using Markov's inequality all the time so let's just try to put together what you already know.

Markov's inequality states: if Y is a nonnegative random variable, then

$$\mathrm{P}(Y \geq k) \leq \frac{\mathrm{E}(Y)}{k}.$$

You've made such arguments informally before, e.g. the file size problem on homework. Formally, we do the following:

$$\mathrm{E}(Y) = \mathrm{E}(Y \mid Y \geq k)\,\mathrm{P}(Y \geq k) + \mathrm{E}(Y \mid Y < k)\,\mathrm{P}(Y < k) \geq k\,\mathrm{P}(Y \geq k) + 0\,\mathrm{P}(Y \leq k).$$

Solve for $\mathrm{P}(Y \geq k)$ to get Markov's inequality.

How do we apply Markov to Chebyshev? Let's write

$$\mathrm{P}(|X - \mu| \geq k\sigma) = \mathrm{P}((X-\mu)^2 \geq k^2\sigma^2) = \mathrm{P}((X-\mu)^2 \geq k^2\,\mathrm{E}((X-\mu)^2)).$$

Let $Y = (X-\mu)^2$, then Markov's inequality states that $\mathrm{P}(Y \geq k^2\,\mathrm{E}(Y)) \leq 1/k^2$. This gives us the inequality we want.

Easy example

Suppose we flip a fair coin 100 times. Let X be the total number of heads. Can we pick a range within which X falls 99% of the time?

$$\mathrm{Var}(X_1 + \cdots + X_{100}) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_{100}) = 100 \cdot (1/2 - 1/4) = 25.$$

We want $k = 10$ so that $k^2 = 1/100$. So we stay within a range of 10σ from the mean... or 50. 99% of the time, the total number of heads is between 0 and 100. Success!

Alright, so it does get useful eventually. Suppose we flip the coin 10000 times, instead. Then $\mathrm{Var}(X) = 10000 \cdot 1/4 = 2500$, so $\sigma = 50$ and $10\sigma = 500$. Therefore we've shown that 99% of the tiem, between 4500 and 5500 coins will come up heads.

However, the computer will say that with 100 coins, the number of heads will be between 37 and 63 just over 99% of the time. With 10000 coins, the corresponding range is $[4871, 5129]$. This is because the binomial distribution is much nicer than the worst-case distribution we assumed in our proof.

Chebyshev's inequality is useful when we don't know much about a distribution: it's the weakest bound ever because it holds in every possible case.

Return to primality testing

Recall the Miller-Rabin primality algorithm: given an input n, we get a yes-or-no output such that

- $\mathrm{P}(\text{YES} \mid n \in \mathbb{P}) = 1$ and $\mathrm{P}(\text{NO} \mid n \in \mathbb{P}) = 0$.
- However, we only know $\mathrm{P}(\text{YES} \mid n \notin \mathbb{P}) = p \leq 1/2$.

Each trial requires a random number a, $0 \leq a \leq n-1$. What if we want to limit our use of random numbers?

Naive method: just try both random numbers. If n is composite, we fail with probability $p^2 \leq \frac{1}{4}$. We can do better! We use the numbers $a, a+b, a+2b, \ldots$ and get an arbitrarily low probability of failure.

We will assume in our analysis that n is composite: if n is prime, we know what the outcome will be. Let $X_i = 1$ if Miller-Rabin returns NO when we use the "random" number $a + bi$. The X_i are not all mutually independent: there is some overlap between the tests. However, we know that for any $i \neq j$, X_i and X_j are pairwise independent: $\mathrm{P}(X_i = 1, X_j = 1) = \mathrm{P}(X_i = 1)\,\mathrm{P}(X_j = 1)$. This is true because knowing just $a + bi$ tells us nothing about $a + bj$.

We let $X = X_1 + \cdots + X_t$ for some number t of trials. We have

$$\mathrm{E}(X) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_t) = t(1-p)$$

by linearity of expectation, and we needed nothing about independence for that. However, I claim that also

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_t) = t\,\mathrm{Var}(X_i) = tp(1-p).$$

Let's see this carefully:

$$\begin{aligned}
\mathrm{E}(X^2) &= \mathrm{E}((X_1 + \cdots + X_t)^2) \\
&= \mathrm{E}\left(\sum_{i=1}^{t} X_i^2 + 2\sum_{i<j} X_i X_j\right) \\
&= \sum_{i=1}^{t} \mathrm{E}(X_i^2) + 2\sum_{i<j} \mathrm{E}(X_i X_j) \\
&= \sum_{i=1}^{t} \mathrm{E}(X_i^2) + 2\sum_{i<j} \mathrm{E}(X_i)\,\mathrm{E}(X_j) \\
&= \sum_{i=1}^{t} (1-p) + 2\sum_{i<j} (1-p)^2 = t(1-p) + t(t-1)(1-p)^2.
\end{aligned}$$

But $\mathrm{E}(X)^2 = t^2(1-p)^2$, so subtracting, we get $(t(1-p)+t^2(1-p)^2-t(1-p)^2)-t^2(1-p)^2 = tp(1-p)$.

Now let's apply Chebyshev:

$$\mathrm{P}(X = 0) = \mathrm{P}(|X - \mathrm{E}(X)| \leq \mathrm{E}(X)) \leq \frac{\mathrm{Var}(X)}{\mathrm{E}(X)^2} = \frac{tp(1-p)}{t^2(1-p)^2} = \frac{p}{1-p} \cdot \frac{1}{t}.$$

So even though we only used 2 random variables, we can get better and better bounds on the probability of false positives, by increasing t.