Massachusetts Institute of Technology | Handout 59
6.042J/18.062J: Mathematics for Computer Science | December 5, 1995
Professors David Karger and Tom Leighton

Notes 24

Last week we spent a lot of time talking about the expected values of random variables. One thing we saw is that it is possible for a random variable to have expected value 1 even though it is never actually anywhere near 1 (e.g., if I flip two coins that are tied together, what is the expected number of heads?). This week, we are going to study *deviations from expectation.* We will study the likelihood that a random variable has a value distant from its expectation and give methods for proving that probability is small.

As a first example, we will consider the coupon collecting problem and show a "high probability result" saying the deviation from the expected mean is almost certainly small. This analysis uses a lot of information about the particular experiment being conducted. There are other situations when we don't have so much information, so we will give a bunch of tools that work even if only a little bit is known, and are therefore very general and easy to apply.

0.1 Deviation from the Mean in Coupon Collecting

What is the probability of the event A that we don't have all the cards after $n \ln n + cn$ time steps? Let A_i be the event that we do not have the i^{th} card. Then

$$\begin{aligned}
\Pr[A_i] &= (1-1/n)^{n\ln n + cn} \\
&\leq e^{-\ln n - c} \\
&= e^{-c}/n
\end{aligned}$$

But then, since $A = \bigcup_{i=1}^{n} A_i$, the union bound says that

$$\begin{aligned}
\Pr[A] &\leq \sum_i \Pr[A_i] \\
&\leq n(e^{-c}/n) \\
&= e^{-c}
\end{aligned}$$

1 Markov's Theorem

Suppose the average person's weight is 140 lb. Can all people weigh 200 lb? Why not? Markov's theorem formalizes this idea:

Theorem 1.1 *If $R : S \to R^+$ is a random variable, then for all k,*

$$Pr(R \geq K) \leq \frac{E[R]}{K}.$$

In other words, if $E[R]$ is small, then R is not likely to be large. Another way to state the same theorem is the following:

Corollary 1.2 *If $R : S \to R^+$, then for all positive real number c,*

$$\Pr(R \geq cE[R]) \leq \frac{1}{c}.$$

Proof. (of corollary) Just write K=cE[R] in the theorem. ■

As an example, if we know that the average person weights 140 pounds, then we know that at most half the people in the world weight over 280 pounds (maybe lots fewer, but this is all Markov's theorem tells us).

Proof. (of theorem)

$$\begin{aligned}
E[R] &= \sum_x x \Pr(R = x) \\
&\geq \sum_{x \geq K} x \Pr(R \geq x) \\
&\geq \sum_{x \geq K} K \Pr(R = x) \\
&= K \sum_{x \geq K} \Pr(R = x) \\
&= K \Pr(R \geq K)
\end{aligned}$$

■

Another example: if I give each person a random homework, what is the probability k people get their own homework back? At most $1/k$, since the expected number getting their own is 1.

Note that Markov's theorem doesn't hold if R can be negative. For example, consider a variable that is ± 1 with probabilities $1/2$ each. It's expectation is 0! So Markov's theorem is clearly violated.

There is a converse to Markov's Theorem for bounding how far *below* expectation you can go:

Theorem 1.3 *If* $\Pr[R > u] = 0$, *then*

$$\Pr[R > k] \geq \frac{E(R) - k}{u - k}$$

Proof is like before.

As an example, if the average score on a test is 75%, then

$$\Pr[R > 50] \geq \frac{75 - 50}{100 - 50} = 1/2$$

Markov's theorem relies only on expectation, and is therefore quite weak. For example: it claims that the probability of getting n heads in n coin flips is at most 1/2, which is true but way off the mark. The greatest use of Markov's theorem is as a tool in proving stronger bounding theorems.

2 Variance and Chebyshev's Theorem

We know that the probability of n heads is much less than the 50% described by Markov' Theorem. How can we formalize this intuition?

Let's consider a random variable R with mean μ. What we are interested in is the deviation of the random variable from its mean, namely $R - \mu$. This is a random variable, so there are are obvious ways to study it. For example, the most obvious measure of this quantity is

$$E[R - \mu].$$

Unfortunately, $E[R - \mu] = E[R] - \mu = 0$, so the positive and negative deviations cancel out and give us no information. So perhaps the thing to do is measure the average *magnitude* of the deviation without worrying about its direction. This suggests computing $E[|X - \mu|]$. However, absolute values are painful in analysis (not differentiable, etc.) so another approach is more effective:

Definition 2.1 *The variance of* R *is*

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

Variance is a useful single-number measure of a distribution that lets us get tighter deviation bounds than Markov' theorem for expectations.

2.1 Standard Deviation

People often like to "renormalize" the variance:

Definition 2.2 *The* standard deviation *of R is*

$$\sigma_R = \sqrt{\mathrm{Var}[R]}.$$

Thus the variance is the square of the standard deviation. It follows that the standard deviation has the same units (days, people, etc.) as the original random variable and its expectation. On an *intuitive* level, the standard deviation measures $E[|X-\mu|]$ (careful—they aren't really equal!).

You will often see a statement like "the temperature tomorrow will be $30° \pm 2°$. That "plus-minus" term is often a reference to the standard deviation of the measurement under some probability model.

Following standard practice, we will use σ to denote the standard deviation of a random variable. The variance is then σ^2.

2.2 An Example

The larger the variance is, the more likely R is to be far from E[R]. Let's consider playing roulette on a fair wheel (one without the 0 that the house always wins on). There are two possible ways to place a bet. You can bet one of two colors (red/black). If you win (50% chance), you double your money. Alternatively, you can bet one of the 100 numbers on the wheel. If you win (1% chance) you get paid 100 to 1 (meaning your profit is 99 times your bet). Let's consider these betting 1$ according to these strategies:

Bet red With probability 1/2. you win 1$. Otherwise, you lose 1$.

Game B. With probability 1/100. you win 99$. Otherwise, you lose 1$.

Which would you play? Well, they both have the same expected profit of 0 dollars. However, the second game is "riskier." Let's compute variances:

- Game A:

$$A - E[A] = \begin{cases} 1 & \text{with probability } 1/2, \\ -1 & \text{with prob. } 1/2. \end{cases}$$

$$(A - E[A])^2 = \begin{cases} 1 & \text{with probability } 1/2 \\ 1 & \text{with prob. } 1/2 \end{cases}$$

$$\begin{aligned} \text{Var}[A] &= 1 \cdot 1/2 + 1 \cdot 1/2 \\ &= 1 \end{aligned}$$

- Game B:

$$B - E[B] = \begin{cases} 99 & \text{with probability } 1/100, \\ -1 & \text{with prob. } 99/100 \end{cases}$$

$$(B - E[B])^2 = \begin{cases} 9801 & \text{with probability } 1/100 \\ 1 & \text{with prob. } 99/100 \end{cases}$$

$$\begin{aligned} \text{Var}[B] &= \frac{1}{100}(9801) + \frac{99}{100}(1) \\ &= 99 \end{aligned}$$

In short, playing a game with high variance is a serious risk: we can win a lot, but we can loose a lot, too. Playing another game with the same expectation of winnings but less variance we can hope for the same average gains risking less. In investments, many people prefer bonds to stocks because, although the expected earnings in bonds are lower, the variance is much lower too. They are willing to pay for risk-reduction.

2.3 The Chebyshev Bound

A variable with small variance is unlikely to be far from its mean. With Markov's Theorem, we deduced that given a mean IQ of 100, only 1/2 the people can have IQ bigger than 200. But suppose we know that the standard deviation in IQ is 10. Then intuitively, it should quite rare to have an IQ bigger than 110. We can show that in fact only 1/100 people have IQ bigger than 200.

Chebyshev's theorem formalizes this observation:

Theorem 2.3 *For any random variable R,*

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

Corollary 2.4 $\Pr(|R - E[R]| > K) \leq \sigma^2/K^2$.

Example 1 In IQ, $\sigma_R = 10$ while $\mu_R = 100$. Therefore,

$$\begin{aligned}\Pr[R \geq 200] &= \Pr[(R-100) \geq 100] \\ &= \Pr[|R - \mu_R| \geq 10\sigma_R] \\ &\leq 1/100\end{aligned}$$

Proof. Consider the random variable $U = (R - E[R])^2$. Note that $E[U] = \mathrm{Var}[R] = \sigma_R^2$. Also,

$$\begin{aligned}\Pr(|R - E[R]| \geq c\sigma_R) &= Pr(|R - E[R]|^2 \geq c^2\sigma_R^2) \\ &= \Pr[U \geq c^2\sigma^2]\end{aligned}$$

Now apply Markov's theorem to the random variable U with $K = c^2\sigma_R^2$ on the right-hand side to get

$$\begin{aligned}\Pr[U \geq c^2\sigma_R^2) &\leq \frac{E[U]}{c^2\sigma_R^2} \\ &= \frac{\sigma_R^2}{c^2\sigma_R^2} \\ &= \frac{1}{c^2}.\end{aligned}$$

∎

Example 2 At most 1/4 of the students score more than 2 standard deviations away from the mean, whatever the set of scores is. Indeed, if R is the score of a random student, then

$$Pr(|R - E(R)| \geq 2\sigma_R) \leq \frac{1}{4},$$

which means that at most 1/4 of all the scores can be that far from the average.

Chebyshev's theorem holds for any distribution. For some distributions, you can prove a much better bound. For example, if R has a binomial distribution, we will see next time that

$$\Pr[R - E[R] \geq c\sigma_R] = O(e^{-c^2/2})$$

a much stronger bound than $1/c^2$.

2.4 A Different Variance Formula

Theorem 2.5 *For all R, we have*

$$Var(R) = E[R^2] - E[R]^2.$$

Proof.

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

■

This formula is often more useful to compute variances than the previous one. For example, if R is a regular die roll,

$$\begin{aligned} \mathrm{Var}[R] &= \frac{1}{6}(1 + 4 + 9 + \cdots + 36) - (3.5)^2 \\ &= 15\frac{1}{6} - 12\frac{1}{4} = 2\frac{11}{12} \end{aligned}$$

This equation is also useful computationally. It means that you can compute variance in a single pass through a stream of data, since you can keep a running total for X and X^2 and at the end, divide by the number of sample points. The other equation would require two passes, one to get the mean and the next the variance.

Corollary 2.6 *For all R,* $E(R^2) \geq E(R)^2$.

Proof. This follows of the previous theorem since Var(R) is always nonnegative, being the expectation of a square. ■

2.5 Manipulating Variances

Theorem 2.7 *If Var(R)=0, then R is a constant.*

Proof. Suppose that

$$\mathrm{Var}[R] = E(R - E[R])^2] = 0.$$

Then $(R - E[R])^2$ must be always 0, thus R must always be equal to E[R]. ■

Theorem 2.8 *For any constants a and b,* $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$.

A Bernoulli Random variable with probability p has variance

$$p \cdot 1^2 + (1-p) \cdot 0^2 - p^2 = p(1-p)$$

What about the binomial distribution? This is messy to do directly, but easy if we figure out how to sum variances:

Theorem 2.9 *If X and Y are* independent*, then*

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$$

Proof. The assumption that X and Y are independent means that $E[XY] = E[X]E[Y]$. Therefore,

$$\begin{aligned}
\mathrm{Var}[X+Y] &= E[(X+Y)^2] - E[X+Y]^2 \\
&= E[X^2 + 2XY + Y^2] - (\mu_X + \mu_Y)^2 \\
&= E[X^2] + 2E[XY] + E[Y^2] - \mu_X^2 - 2\mu_X\mu_Y - 2\mu_Y^2 \\
&= E[X^2] - \mu_X^2 + E[Y^2] - \mu_Y^2 + 2E[XY] - 2\mu_X\mu_Y \\
&= E[X^2] - \mu_X^2 + E[Y^2] - \mu_Y^2
\end{aligned}$$

(since $E[XY] = \mu_X\mu_Y$ by independence)

$$= \mathrm{Var}[X] + \mathrm{Var}[Y]$$

∎

Corollary 2.10 *If $X_1, \ldots, X_n$ are* pairwise *independent random variables, then*

$$\mathrm{Var}[\sum X_i] = \sum \mathrm{Var}[X_i].$$

note that mutual independence is *not* required, because the only fact we used that relied on independence was that $E[X_iX_j] = E[X_i]E[X_j]$. This follows from pairwise independence, since it means X_i and X_j are independent.

Therefore, the variance of the binomial distribution on n Bernoulli trials with probability p is $np(1-p)$. It follows that the standard deviation is $\sigma = \sqrt{np(1-p)}$. As you determined in your homework, this is about the amount of of the expected absolute deviation from the mean in a binomial distribution.

Example 3 It turns out that the variance of the negative binomial distribution is $(1-p)/p^2$. If we go back to our coupon collecting problem, we see that the variance in the time to collect all the cards is

$$\begin{aligned}\sum_{k=1}^{n}(1-\frac{k}{n})\left(\frac{n}{k}\right)^2 &\leq \sum_{k=1}^{n}\left(\frac{n}{k}\right)^2 \\ &\leq n^2\sum_{k=1}^{n}\frac{1}{k^2} \\ &\leq n^2(1+\int_1^i nfty\frac{1}{k^2}\,dk \\ &= 2n^2\end{aligned}$$

It follows that the standard deviation $\sigma \leq n\sqrt{2}$, so

$$\Pr[|W_n - E[W_n]| > cn] < \frac{2}{c^2}$$

At the start of lecture, we showed a much better bound of e^{-c}. However, our analysis there only considered the case of W_n deviating *above* the mean, while the analysis here also gives for free the probability of W_n deviating *below* the mean.

3 A Law of Large Numbers

We now give a more accurate explanation of how a sequence of experiments "converges towards the mean."

Theorem 3.1 (Law of Large numbers) *Consider any distribution with mean μ and finite variance, and let $X_1, X_2, \ldots$ be a sequence of random variables drawn independently from the distribution. Then for any ϵ,*

$$\lim_{n\to\infty}\Pr\left(\left|\frac{X_1+X_2+\cdots+X_n}{n}\mu\right| > \epsilon\right) = 0$$

Proof. First note that the mean of the expression above is, by linearity of expectation, $n\mu/n = \mu$. From our rule for sums and our rule for product by a constant, we know that

$$\begin{aligned}\mathrm{Var}[\frac{X_1+X_2+\cdots+X_n}{n}] &= \frac{1}{n^2}(n\sigma) \\ &= \frac{\sigma}{n}\end{aligned}$$

Therefore, by Chebyshev's theorem, the probability of divergence by ϵ from the mean μ is

$$\frac{\sigma}{\epsilon^2 n}$$

which goes to 0 as $n \to \infty$. ■

This theorem shows that when you average more and more samples, the result gets arbitrarily tightly concentrated around its expectation. Does this mean that a dramatic deviation from the expectation is impossible? No, just very unlikely.