Lecture 24: Large Deviations: Chebyshev and Chernoff Bound, Wrap up.

Readings: Chapter 20.

1 Announcements

- Schedule: Lecture today, no rec tomorrow (instead review sessions in 32-124), final exam Friday
- Final Exam, and Review Materials (see announcement)

2 Review: Variance

Let's start with the definition of variance from the last lecture.

Definition 1. ***Variance*** *of R is*

$$\text{Var}[R] = \text{Ex}\left[(R - \text{Ex}[R])^2\right]$$

Standard deviation *of R, denoted $\sigma(R)$, is the (positive) square root of the variance.*

Another way to compute the variance:

Theorem 1.

$$\text{Var}[R] = \text{Ex}\left[R^2\right] - \text{Ex}[R]^2$$

Proof. (Skip; proved in recitation.)

$$\begin{aligned}
\text{Var}[R] = \text{Ex}\left[(R - \text{Ex}[R])^2\right] &= \text{Ex}\left[R^2 - 2 \cdot \text{Ex}[R] \cdot R + \text{Ex}[R]^2\right] \\
&= \text{Ex}\left[R^2\right] - 2 \cdot \text{Ex}[R] \cdot \text{Ex}[R] + \text{Ex}[R]^2 \\
&= \text{Ex}\left[R^2\right] - \text{Ex}[R]^2
\end{aligned}$$

□

Theorem 2. *If $R_1, \ldots, R_n$ are* pairwise independent *random variables, then*

$$\text{Var}[R_1 + \ldots + R_n] = \text{Var}[R_1] + \ldots + \text{Var}[R_n]$$

Proof. (Skip; proved in recitation.)

$$
\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^{n} R_i\right] &= \mathrm{Ex}\left[\left(\sum_{i=1}^{n} R_i\right)^2\right] - \left(\mathrm{Ex}\left[\sum_{i=1}^{n} R_i\right]\right)^2 \\
&= \mathrm{Ex}\left[\sum_{i=1}^{n} R_i^2 + 2\cdot\sum_{i\neq j} R_i R_j\right] - \sum_{i=1}^{n} \mathrm{Ex}\left[R_i\right]^2 - 2\sum_{i\neq j} \mathrm{Ex}\left[R_i\right]\mathrm{Ex}\left[R_j\right] \\
&= \sum_{i=1}^{n}\left(\mathrm{Ex}\left[R_i^2\right] - \mathrm{Ex}\left[R_i\right]^2\right) + \sum_{i\neq j}\left(\mathrm{Ex}\left[R_i R_j\right] - \mathrm{Ex}\left[R_i\right]\mathrm{Ex}\left[R_j\right]\right)
\end{aligned}
$$

Since for every $i \neq j$, the r.v.s R_i and R_j are independent, $\mathrm{Ex}[R_i R_j] = \mathrm{Ex}[R_i]\,\mathrm{Ex}[R_j]$, so the second term above vanishes. The first term is just the sum of the variances of all the R_i, so there you go. □

Warning: $\sigma(R_1 + R_2)$ is *not* necessarily $\sigma(R_1) + \sigma(R_2)$ even when R_1 and R_2 are independent. But the theorem above tells us that $\sigma(R_1 + R_2)^2 = \sigma(R_1)^2 + \sigma(R_2)^2$, when they are independent.

3 Large Deviation Bounds

Theorem 3 (Markov's Inequality)**.** *Let R be a **non-negative** random variable. Then,*

$$\Pr[R \geq x] \leq \frac{\mathrm{Ex}[R]}{x}$$

Example: Let R be the weight of a random person. Say $\mathrm{Ex}[R] = 100$. What is the probability that $R \geq 200$?

Answer: We don't have enough information to compute the exact probability, but Markov tells us that this is at most $100/200 = 1/2$.

There is a definite, non-probabilistic, interpretation of this statement: at most half the population weighs at least 200 lbs. There is nothing probabilistic about that.

Proof of Markov's Inequality.

$$\mathrm{Ex}[R] = \mathrm{Ex}[R \mid R \geq x]\Pr[R \geq x] + \mathrm{Ex}[R \mid R < x]\Pr[R < x]$$

by the law of total probabilities applied to expectation. the first expectation term on the RHS is at least x and the second expectation term on the RHS is at least 0 (this is where we are using non-negativity of R.) So,

$$\mathrm{Ex}[R] \geq x \cdot \Pr[R \geq x] + 0 \cdot \Pr[R < x] \geq x \cdot \Pr[R \geq x]$$

Rearranging the terms, we get

$$\Pr[R \geq x] \leq \frac{\text{Ex}[R]}{x}$$

□

An alternate form of Markov:

Theorem 4 (Markov's Inequality, alternate form)**.** *Let R be a* ***non-negative*** *random variable. Then*

$$\Pr[R \geq c \cdot \text{Ex}[R]] \leq \frac{1}{c}$$

3.1 Useful strategy: adjusting bounds

Say R is test scores, always between 30% and 100%. Say average grade is 75%. Can we use Markov to bound the probability of getting at least 90%?

$$\Pr[R \geq 90] \leq \frac{\text{Ex}[R]}{90} = 75/90 \approx .833$$

Can get a better bound by noticing that $R - 30$ is a nonnegative random variable! $\Pr[R \geq 90] = \Pr[R - 30 \geq 60]$ (why?[1]), which by Markov *applied to random variable* $R-30$ is $\leq \text{Ex}[R-30]/60 = 45/60 \approx .75$.

What about probability that $R \leq 65$? Markov is usually for $R \geq k$, not $R \leq k$. But since we know an upper bound for R, we can instead look at $100 - R$, which is nonnegative! Then $\Pr[R \leq 65] = \Pr[100 - R \geq 35] \leq \text{Ex}[100 - R]/35 = 25/35 \approx 0.714$. We can use Markov since we know an upper bound for R.

In general, if we know $S \geq \ell$, try applying Markov to $S - \ell$. If we know $S \leq u$, try applying Markov to $u - S$ to bound the probability that S is *at most* something.

This includes some cases where the random variable might be negative: If we know $S \geq -4$, we can't apply Markov to S because S isn't nonnegative, but $S + 4$ is nonnegative, so Markov can be used.

3.2 Why does Markov need non-negativity anyway?

Here is a counterexample: consider R which takes on the value -1 if an unbiased coin comes up heads and $+1$ if it comes up tails. $\text{Ex}[R] = 0$. $\Pr[R \geq 1/2] = 1/2$ but (**incorrectly!**) applying Markov would have us conclude this is at most 0.

Looking at the proof of Markov above tells us where things go wrong if R is not non-negative. We used that $\text{Ex}[R \mid R < x]$ is at least 0 appealing to the non-negativity of R.

[1] $[R \geq 90]$ and $[R - 30 \geq 60]$ are exactly the same event

3.3 Markov is (often) not tight

Let's look at the cellphone check problem, from L22/23, starting from the lazy suzan version. Recall: n people sit around a table, place their cellphones on a lazy suzan and give it a spin. If R is the number of people who got their cellphones back, then we saw that $\text{Ex}[R] = 1$. What is the probability that all n get their cellphone back?

Markov has an answer. It is $\leq \text{Ex}[R]/n = 1/n$. What's the true answer? Also $1/n$.

Let's look at the original version of the cellphone check problem, where the n phones are permuted and returned. What is the probability that all n get their cellphone back?

Markov has the same answer! It is $\leq \text{Ex}[R]/n = 1/n$. What's the true answer? It is $1/(n!)$. $n! \gg n$, so Markov is way off in the estimate here. The *upper bound* that Markov gives us is *correct* but is *loose*. The true probability is much smaller.

What if we want tighter bounds? For that, we need to know something more about the probability distribution than just its mean.

3.4 A Recurring Example

Example: Let's look at the number of heads in a toss of n coins. Here,

$$R = R_1 + \ldots + R_n$$

where R_i is the indicator random variable which is 1 if and only if the i-th coin toss came up heads.

$$\text{Ex}[R_i] = 1/2 \text{ and } \text{Var}[R_i] = \text{Ex}\left[R_i^2\right] - \text{Ex}[R_i]^2 = 1/2 - 1/4 = 1/4$$

Now,

$$\text{Ex}[R] = \sum_{i=1}^{n} \text{Ex}[R_i] = n/2$$

and

$$\text{Var}[R] = \sum_{i=1}^{n} \text{Var}[R_i] = n/4$$

$$\sigma(R) = \sqrt{n/4} = \sqrt{n}/2$$

(We will see later in the lecture that this number $\sqrt{n}$ has a special meaning: there is a good chance that you won't see the number of heads in n coin tosses falling outside the range $[\frac{n}{2} - c\sqrt{n}, \frac{n}{2} + c\sqrt{n}]$ for large enough constants $c > 0$. The number of heads is "concentrated around $n/2$".)

Markov tells us that

$$\Pr[R \geq 3n/4] \leq \text{Ex}[R]/(3n/4) = (n/2)/(3n/4) = 2/3$$

We'll do much better later.

4 Chebyshev

Theorem 5 (Chebyshev's Inequality)**.** *For every $x > 0$ and for every r.v. R (not necessarily non-negative),*

$$\Pr[|R - \text{Ex}[R]| \geq x] \leq \frac{\text{Var}[R]}{x^2} = \left(\frac{\sigma(R)}{x}\right)^2$$

where $\sigma(R)$ is the standard deviation of R.

This bears repeating: R can be any random variable! It doesn't have to be nonnegative anymore.

Proof. Use Markov! With the (non-negative) random variable $(R - \text{Ex}[R])^2$. Now, and make sure you understand this step,

$$\Pr[|R - \text{Ex}[R]| \geq x] = \Pr\left[(R - \text{Ex}[R])^2 \geq x^2\right]$$

Now apply Markov and get

$$\Pr[|R - \text{Ex}[R]| \geq x] = \Pr\left[(R - \text{Ex}[R])^2 \geq x^2\right] \leq \frac{\text{Ex}[(R - \text{Ex}[R])^2]}{x^2} = \frac{\text{Var}[R]}{x^2}$$

□

Theorem 6. *For every $x > 0$ and for every r.v. R (not necessarily non-negative),*

$$\Pr[|R - \text{Ex}[R]| \geq c \cdot \sigma(R)] \leq \frac{1}{c^2}$$

where $\sigma(R)$ is the standard deviation of R.

Example 1: Let's go back to the test scores whose variance is, say 25 (so the standard deviation is 5).

$$\Pr[\mathsf{score} \leq 65] \leq \Pr[|\mathsf{score} - 75| \geq 10]$$

Why? The latter probability measures the union of two events — that $\mathsf{score} \leq 65$ and that $\mathsf{score} \geq 85$.

Apply Chebyshev:

$$\Pr[|\mathsf{score} - 75| \geq 10] \leq \frac{\text{Var}[\mathsf{score}]}{10^2} = \frac{25}{100} = .25$$

Equivalently, we're asking about the probability of being at least $c = 2$ standard deviations away from the mean, which Chebyshev shows has probability at most $1/c^2 = 1/4$. This is a much better bound than we got using Markov alone!

Example 2: Back to number of heads in n coin flips. Chebyshev tells us that

$$\Pr[R \geq 3n/4] \leq \Pr[|R - n/2| \geq n/4] \leq \frac{\text{Var}[R]}{(n/4)^2} = \frac{(n/4)}{(n/4)^2} = \frac{4}{n}$$

which is a far better bound.

5 Chernoff

It turns out there is something even better that one can do. Recall that Chebyshev only uses the *pairwise* independence of the coin tosses. Using the *mutual* independence of all the coin tosses gives us a better bound via the Chernoff bound.

Theorem 7 (Chernoff)**.** *Let Let $T_1, \ldots, T_n$ be mutually independent random variables such that $0 \leq T_i \leq 1$ for all i. Let $T = T_1 + T_2 + \ldots + T_n$. Then, for all $c \geq 1$,*

$$\Pr[T \geq c \cdot \mathrm{Ex}[T]] \leq e^{-(c \ln c - c + 1) \cdot \mathrm{Ex}[T]}$$

The proof, like that of Chebyshev, uses Markov on a different random variable, namely c^T. For the real proof, I will refer you to the book, section 20.5.6.

Let's apply Chernoff to the coin tosses. We get, letting $c = 3/2$,

$$\Pr[R \geq 3n/4] = \Pr[R \geq 3/2 \cdot n/2] \leq e^{-0.1 \cdot n/2} = e^{-n/20}$$

which is an exponentially better bound than Chebyshev!

Letting $c = 1 + (4/\sqrt{n})$, we can prove

$$\Pr\left[R \geq \frac{n}{2} + 2\sqrt{n}\right] \leq 0.02$$

for large n.

Note that $\sqrt{n}$ is *much* smaller than n, so this distribution clumps tighter and tighter around the mean (proportionally) as n increases. This is one sense in which coin flips are very concentrated around $n/2$.

6 The End!

Thanks for a fun semester. Good luck with finals and enjoy your summer!