More Tail Bounds

CSE 312 Spring 24
Lecture 22

Announcements

Sections will go forward as normal this week.

OH are updated on the calendar (a few small tweaks later this week, those might go back to “normal”, checking with TAs).

The UAW and the university reached a tentative agreement late last night. We anticipate things being fully ‘back to normal’ within a few days.

Near the mean

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

$\bar{X} = \sum X_i / 1000$

$\mathbb{E}[\bar{X}]$

$\text{Var}(\bar{X})$

Chebyshev's Inequality

Let X be a random variable. For any $t > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\text{Var}(X)}{t^2}$$

Near the mean

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

$$\bar{X} = \sum X_i / 1000$$

$$\mathbb{E}[\bar{X}] = 1000 \cdot \frac{.6}{1000} = \frac{3}{5}$$

$$\text{Var}(\bar{X}) = 1000 \frac{.6 \cdot .4}{1000^2} = \frac{3}{12500}$$

Chebyshev's Inequality

Let X be a random variable. For any $t > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\text{Var}(X)}{t^2}$$

Near the mean

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

$\bar{X} = \sum X_i/1000$

$\mathbb{E}[\bar{X}] = 1000 \cdot \frac{.6}{1000} = \frac{3}{5}$

$\text{Var}(\bar{X}) = 1000 \frac{.6 \cdot .4}{1000^2} = \frac{3}{12500}$

Chebyshev's Inequality

Let X be a random variable. For any $t > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\text{Var}(X)}{t^2}$$

Near the mean

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

$$\bar{X} = \sum X_i / 1000$$

$$\mathbb{E}[\bar{X}] = 1000 \cdot \frac{.6}{1000} = \frac{3}{5}$$

$$\text{Var}(\bar{X}) = 1000 \frac{.6 \cdot .4}{1000^2} = \frac{3}{12500}$$

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq .1) \leq \frac{3/12500}{.1^2} = .024$$

Chebyshev's Inequality

Let X be a random variable. For any $t > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\text{Var}(X)}{t^2}$$

Chebyshev's – Repeated Experiments

How many coin flips (each head with probability p) are needed until you get n heads.

Let X be the number necessary. What is probability $X \geq 2n/p$?

Markov

Chebyshev

Chebyshev's – Repeated Experiments

How many coin flips (each head with probability p) are needed until you get n heads.

Let X be the number necessary. What is probability $X \geq 2n/p$?

Markov $\quad \mathbb{P}\left(X \geq \frac{2n}{p}\right) \leq \frac{n/p}{2n/p} = \frac{1}{2}$

Chebyshev $\quad \mathbb{P}\left(X \geq \frac{2n}{p}\right) \leq \mathbb{P}\left(\left|X - \frac{n}{p}\right| \geq \frac{n}{p}\right) \leq \frac{\text{Var}(X)}{n^2/p^2} = \frac{n(1-p)/p^2}{n^2/p^2} = \frac{1-p}{n}$

Takeaway

Chebyshev gets more powerful as the variance shrinks.

Repeated experiments are a great way to cause that to happen.

More Assumptions → Better Guarantee

(Multiplicative) Chernoff Bound

Let $X_1, X_2, \ldots, X_n$ be *independent* Bernoulli random variables.
Let $X = \sum X_i$, and $\mu = \mathbb{E}[X]$. For any $0 \le \delta \le 1$

$$\mathbb{P}(X \ge (1+\delta)\mu) \le \exp\left(-\frac{\delta^2\mu}{3}\right) \text{ and } \mathbb{P}(X \le (1-\delta)\mu) \le \exp\left(-\frac{\delta^2\mu}{2}\right)$$

Same Problem, New Solution

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

(Multiplicative) Chernoff Bound

Let $X_1, X_2, \dots, X_n$ be *independent* Bernoulli random variables. *Let* $X = \sum X_i$, and $\mu = \mathbb{E}[X]$. For any $0 \leq \delta \leq 1$

$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right)$ and $\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{2}\right)$

Right Tail

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

Want $\mathbb{P}\left(\frac{X}{1000} \geq .7\right)$

Chernoff Bound (right tail)

Let $X_1, X_2, \ldots, X_n$ be *independent* Bernoulli random variables. *Let* $X = \sum X_i$, and $\mu = \mathbb{E}[X]$. For any $0 \leq \delta \leq 1$

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{3}\right)$$

Right Tail

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

Want $\mathbb{P}\left(\frac{X}{1000} \geq .7\right) = \mathbb{P}(X \geq .7 \cdot 1000)$

$= \mathbb{P}(X \geq (1 + .1/.6) \cdot (.6 \cdot 1000))$

So $\delta = \frac{1}{6}$ and $\mu = .6 \cdot 1000$

$\mathbb{P}(X \geq 700) \leq \exp\left(-\frac{\frac{1}{6^2} \cdot .6 \cdot 1000}{3}\right)$

≤ 0.0039

Chernoff Bound (right tail)

Let $X_1, X_2, \ldots, X_n$ be *independent* Bernoulli random variables. *Let* $X = \sum X_i$, and $\mu = \mathbb{E}[X]$. For any $0 \leq \delta \leq 1$

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{3}\right)$$

Left Tail

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

Want $\mathbb{P}\left(\frac{X}{1000} \le .5\right) = \mathbb{P}(X \le .5 \cdot 1000)$

Chernoff Bound (left tail)

Let $X_1, X_2, \ldots, X_n$ be *independent* Bernoulli random variables. *Let* $X = \sum X_i$, and $\mu = \mathbb{E}[X]$. For any $0 \le \delta \le 1$

$$\mathbb{P}(X \le (1-\delta)\mu) \le \exp\left(-\frac{\delta^2 \mu}{2}\right)$$

Left Tail

Suppose you run a poll of 1000 people where in the true population 60% of the population supports you. What is the probability that the poll is not within 10-percentage-points of the true value?

Want $\mathbb{P}\left(\frac{X}{1000} \leq .5\right) = \mathbb{P}(X \leq .5 \cdot 1000)$

$= \mathbb{P}(X \leq (1 - .1/.6) \cdot (.6 \cdot 1000))$

So $\delta = \frac{1}{6}$ and $\mu = .6 \cdot 1000$

$\mathbb{P}(X \leq 500) \leq \exp\left(-\frac{\frac{1}{6^2} \cdot .6 \cdot 1000}{2}\right)$

≤ 0.0003

Chernoff Bound (left tail)

Let $X_1, X_2, \ldots, X_n$ be *independent* Bernoulli random variables. *Let* $X = \sum X_i$, and $\mu = \mathbb{E}[X]$. For any $0 \leq \delta \leq 1$

$$\mathbb{P}(X \leq (1 - \delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right)$$

Both Tails

Let E be the event that X is not between 500 and 700 (i.e. we're not within 10 percentage points of the true value)

$$\mathbb{P}(E) = \mathbb{P}(X < 500) + \mathbb{P}(X > 700)$$

$$\leq .0039 + .0003 = .0042$$

Less than 1%. That's a better bound than Chebyshev gave!

Wait a Minute

I asked Wikipedia about the "Chernoff Bound" and I saw something different?

This is the "easiest to use" version of the bound. If you need something more precise, there are other versions.

Why are the tails different??

The strongest/original versions of "Chernoff bounds" are symmetric ($1 + \delta$ and $1 - \delta$ **correspond**), but those bounds are ugly and hard to use.

When computer scientists made the "easy to use versions", they needed to use some inequalities. The numerators now have plain old δ's, instead of $1 +$ or $1 -$. As part of the simplification to this version, there were different inequalities used so you don't get exactly the same expression.

Wait a Minute

This is just a binomial!

Well if all the X_i have the same probability. It does work if they're independent but have different distributions. But there's bigger reasons to care…

The concentration inequality will let you control n easily, even as a variable. That's not easy with the binomial.

What happens when n gets big?

Evaluating $\binom{20000}{10000}.51^{10000} \cdot .49^{10000}$ is fraught with chances for floating point error and other issues. Chernoff is much better.

But Wait! There's More

For this class, please limit yourself to:
Markov, Chebyshev, and Chernoff, as stated in these slides…

But for your information. There's more.

Trying to apply Chebyshev, but only want a "one-sided" bound (and tired of losing that almost-factor-of-two)Try Cantelli's Inequality

In a position to use Chernoff, but want additive distance to the mean instead of multiplicative? They got one of those.

Have a sum of independent random variables that aren't indicators, but are bounded, you better believe Wikipedia's got one

Have a sum of random **matrices** instead of a sum of random numbers. Not only is that a thing you can do, but the eigenvalue of the matrix concentrates

There's a whole book of these!

Tail Bounds – Takeaways

Useful when an experiment is complicated and you just need the probability to be small (you don't need the exact value).

Choosing a minimum n for a poll – don't need exact probability of failure, just to make sure it's small.

Designing probabilistic algorithms – just need a guarantee that they'll be extremely accurate

Learning more about the situation (e.g. learning variance instead of just mean, knowing bounds on the support of the starting variables) usually lets you get more accurate bounds.

One More Bound

One More Bound

The Union bound

Union Bound

For any events E, F

$$\mathbb{P}(\boldsymbol{E} \cup \boldsymbol{F}) \leq \mathbb{P}(\boldsymbol{E}) + \mathbb{P}(\boldsymbol{F})$$

Proof? $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$

And $\mathbb{P}(E \cap F) \geq 0$.

Concentration Applications

A common pattern:

Figure out "what could possibly go wrong" – often these are dependent.

Use a concentration inequality for each of the things that could go wrong.

Union bound over everything that could go wrong.

Frogs

There are 20 frogs on each location in a 5x5 grid. Each frog will independently jump to the left, right, up, down, or stay where it is with equal probability. A frog at an edge of the grid magically warps to the corresponding edge (pac-man-style).

Bound the probability that at least one square ends up with at least 36 frogs.

These events are dependent – adjacent squares affect each other!

Frogs

For an arbitrary location:

There are 100 frogs who could end up there (those above, below, left, right, and at that location). Each with probability $.2$. Let X be the number that land at the location we're interested in.

$$\mathbb{P}(X \geq 36) = \mathbb{P}(X \geq (1+\delta)20) \leq \exp\left(-\frac{\left(\frac{4}{5}\right)^2 \cdot 20}{3}\right) \leq 0.015$$

There are 25 locations. Since all locations are symmetric, by the union bound the probability of at least one location having 36 or more frogs is at most $25 \cdot 0.015 \leq 0.375$.

Tail Bounds – Takeaways

Useful when an experiment is complicated and you just need the probability to be small (you don't need the exact value).

Choosing a minimum n for a poll – don't need exact probability of failure, just to make sure it's small.

Designing probabilistic algorithms – just need a guarantee that they'll be extremely accurate

Learning more about the situation (e.g. learning variance instead of just mean, knowing bounds on the support of the starting variables) usually lets you get more accurate bounds.

Applications

Privacy Preservation

A real-world example (adapted from *The Ethical Algorithm* by Kearns and Roth; based on protocol by Warner [1965]).

And gives a sense of how randomness is actually used to protect privacy.

Privacy Preservation with Randomness

You’re working with a social scientist. They want to get accurate data on the rate at which people cheat on their romantic partners.

We know about polling accuracy!

Do a poll, call up a random sample of adults and ask them “have you ever cheated on your romantic partner?”

Use a tail-bound to estimate the needed number n get a guaranteed good estimate, right?

You do that, and somehow, no one says they cheated.

What’s the problem?

People lie.

Or they might be concerned about you keeping this data.

Databases can be leaked (or infiltrated. Or subpoenaed).

You don’t want to hold this data, and the people you’re calling don’t want you to hold this data.

Doing Better With Randomness

You don't really need to know **who** was cheating. Just how many people were.

Here's a protocol:

Please flip a coin.

- If the coin is heads, or you have ever cheated, please tell me "heads"
- If the coin is tails and you have not ever cheated, please tell me "tails"

Will it be private?

If you are someone who has cheated, and you report heads can that be used against you? Not substantially – just say “no the coin came up heads!”

You discover your partner said heads, what’s the probability that they cheated?

Will it be private?

If you are someone who has cheated on your spouse, and you report heads can that be used against you? Not substantially – just say “no the coin came up heads!”

$$\mathbb{P}(C|H) = \frac{\mathbb{P}(H|C)\cdot\mathbb{P}(C)}{\mathbb{P}(H)} = \frac{1\cdot\mathbb{P}(C)}{\frac{1}{2}\mathbb{P}(\overline{C}) + 1\cdot\mathbb{P}(C)}$$

Is this a substantial change?

No. For real world values (~15%) of $\mathbb{P}(C)$, the probability estimate would increase (to ~26%). But that isn’t too damaging.