More Tail Bounds

CSE 312 Summer 25
Lecture 20

Announcements

Homework 5 Solutions available at the front + outside Allen 206

Homework 6 will be due on Wednesday

Homework 7 will be out on Wednesday and due Monday at 11:59pm with a late deadline on Tuesday at 11:59pm

Quiz 7 on Friday

Midterm Retake Opportunity

Monday August 18th at 12pm in DEM 104.

You will have 60 minutes.

Same point value and similar difficulty.

The same reference sheet will be provided.

The score will be capped at 75% or 61.5/82 points.

We will take the max of the two scores.

Midterm score update will only be incorporated after grade breaks are set (like extra credit).

Looking ahead at next week

Monday: Midterm Retake + HW7 Due

Wednesday: Final Review

Thursday: Final Part 1 (hopefully in DEM 104)

Friday: Final Part 2

1:1 TA's

Anna

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?

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

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

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

Chebyshev's Inequality

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

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

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

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

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 $0 \leq \delta \leq 1$

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

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

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

Right Tail

$\mu = 600$

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

$P(X \geq 700)$

$\frac{1}{6}$

$700 = (1+\delta)\mu$

$700 = (1+\delta)600$

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

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

Left Tail

$P(X \le 500)$

$P(X \le (1-\delta) \cdot \mu)$

$\mu = 600$

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

$500 = (1-\delta) \cdot 600$

$\delta = \frac{1}{6}$

Chernoff Bound (left tail)

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 $0 \le \delta \le 1$

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

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!

One More Bound

One More Bound

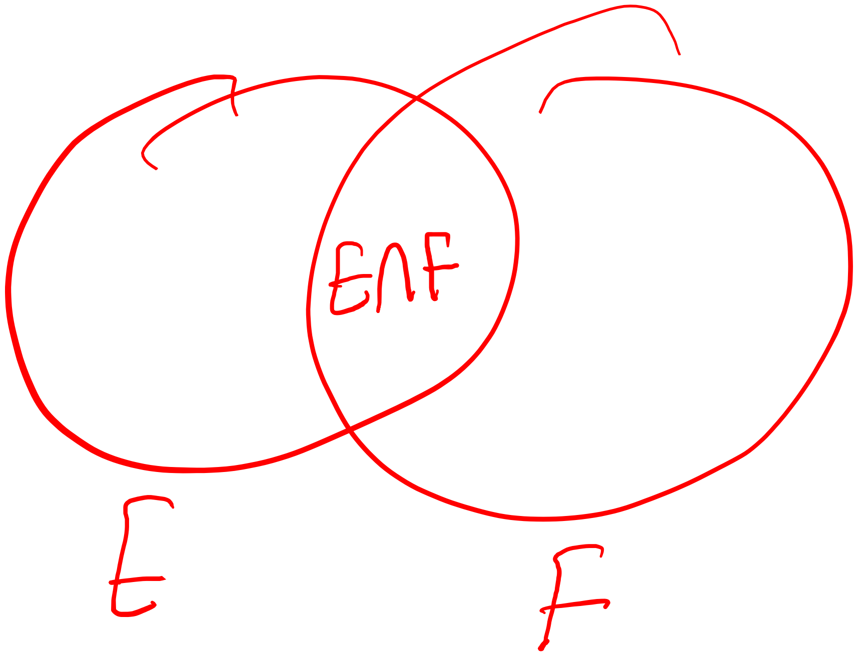
The Union bound

Union Bound

For any events E, F
$$\mathbb{P}(E \cup F) \leq \mathbb{P}(E) + \mathbb{P}(F)$$

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.

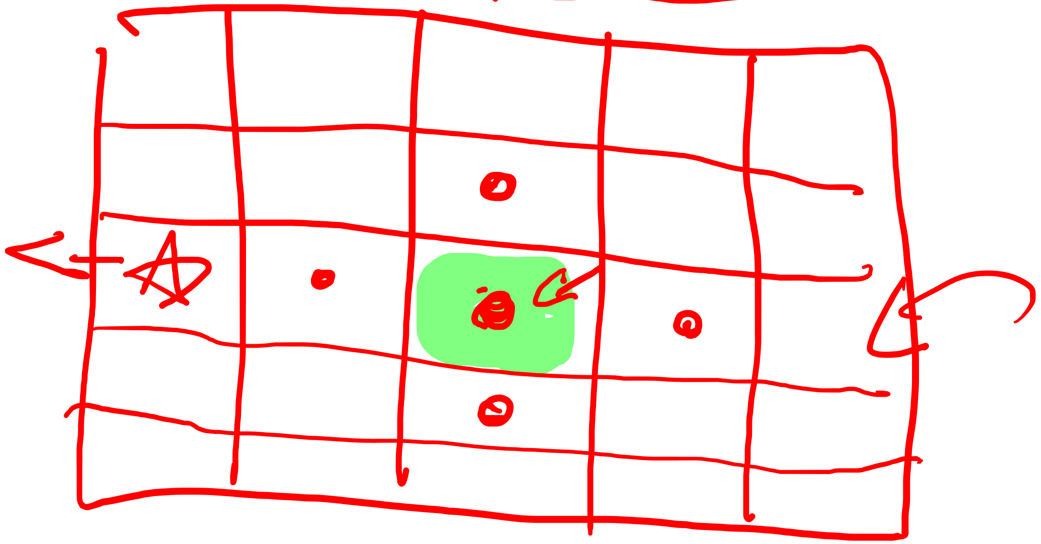
Frogs

500

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.

X = # frogs on green square



100

$\frac{1}{5}$

$X = \sum^{100} X_i$

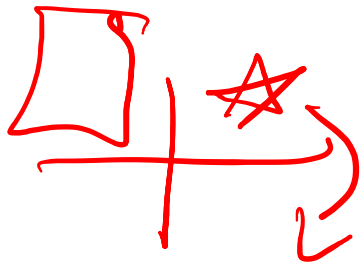
Frogs

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.

But will it be accurate?

But we've lost our data haven't we? People answered a different question. Can we still estimate how many people cheated?

Suppose you asked 100 people the "heads/tails" question, and 60 people said "heads." What do you predict would be the number of people who cheated on a partner?

Can you generalize your idea for n people polled, and X the number of people that said "heads"?

But will it be accurate?

But we've lost our data haven't we? People answered a different question. Can we still estimate how many people cheated?

Suppose you poll n people, and let X be the number of people who said "heads" We'll find an estimate Y of the number of people who cheated in the sample, and let p be the true probability of cheating in the population. What should Y be? Can we draw a margin of error around Y?

$$\mathbb{P}(X_i = 1) = \frac{1}{2} + \frac{1}{2} \cdot p$$

$$\mathbb{E}[X] = \frac{n}{2} + \frac{1}{2}\mathbb{E}[Y]$$

We'll define Y to be: $Y = 2\left(X - \frac{n}{2}\right)$.

This is a **definition**, based on how the $\mathbb{E}[Y]$ should relate to the $\mathbb{E}[X]$.

But will it be accurate?

$Y = 2\left(X - \frac{n}{2}\right)$

$\text{Var}(X) = \text{Var}(\sum X_i) = \sum \text{Var}(X_i)$

$\text{Var}(X_i)$? It's an indicator with parameter $p + (1-p) \cdot \frac{1}{2} = \frac{1}{2} + \frac{p}{2}$

So $\text{Var}(X_i) = \left(\frac{1}{2} + \frac{p}{2}\right)\left(\frac{1}{2} - \frac{p}{2}\right)$

$\text{Var}(Y) = 4\text{Var}(X) = 4n\text{Var}(X_i) = 4n\left(\frac{1}{2} + \frac{p}{2}\right)\left(\frac{1}{2} - \frac{p}{2}\right) \leq \frac{4n}{4} = n$

The variance is 4 times as much as it would have been for a non-anonymous poll.

Can we use Chernoff?

(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 $0 \le \delta \le 1$

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

What happens with $\mathrm{n} = 1000$ people?

What range will we be within at least 95% of the time?

A different inequality

If we try to use Chernoff, we'll hit a frustrating block.

Since μ depends on p, p appears in the formula for δ. And we wouldn't get an absolute guarantee unless we could plug in a p.

And it'll turn out that as $p \to 0$ that $\delta \to \infty$ so we don't say anything then.

Luckily, there's always another bound…

☹ Can't bound δ without bounding p

The right tail is the looser bound, so ensuring the right tail is less than 2.5% gives us the needed guarantee.

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right) = \exp\left(-\frac{\delta^2 1000p}{3}\right) \leq .025$$

$$-\frac{\delta^2 1000p}{3} \leq \ln(.025)$$

$$-\delta^2 \leq \frac{3 \cdot \ln(.025)}{1000p}$$

$$\delta \geq \sqrt{\frac{-3\ln(.025)}{1000p}}$$

As $p \to 0$, $\delta \to \infty$ – we're not actually making a claim anymore.

Hoeffding's Inequality

Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be *independent* RVs, each with range $[0,1]$.
Let $\bar{X} = \sum X_i / n$, and $\mu = \mathbb{E}[\bar{X}]$. For any $t \geq 0$

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp(-2nt^2)$$

$|X - \mathbb{E}[X]| \geq t$ if and only if $|Y - \mathbb{E}[Y]| \geq 2t$. Why?

$$Y = 2\left(X - \frac{n}{2}\right) \text{ or } X = \frac{Y+n}{2}$$

$$|X - \mathbb{E}[X]|$$

$$= \left|\frac{Y+n}{2} - \mathbb{E}\left[\frac{Y+n}{2}\right]\right|$$

$$= \left|\frac{Y+n}{2} - \mathbb{E}\left[\frac{Y}{2}\right] - \frac{n}{2}\right|$$

$$= \left|\frac{Y}{2} - \mathbb{E}\left[\frac{Y}{2}\right]\right|$$

$$= \frac{1}{2}|Y - \mathbb{E}[Y]|$$

So $|X - \mathbb{E}[X]| \geq t$ if and only if $\frac{1}{2}|Y - \mathbb{E}[Y]| \geq t$ iff $|Y - \mathbb{E}[Y]| \geq 2t$.

Hoeffding's Inequality

Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be ***independent*** RVs, each with range $[0,1]$.
Let $\bar{X} = \sum X_i / n$, **and** $\mu = \mathbb{E}[\bar{X}]$. **For any** $t \geq 0$

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp(-2nt^2)$$

How close will we be with n=1000 with probability at least .95?

$|X - \mathbb{E}[X]| \geq t$ if and only if $|Y - \mathbb{E}[Y]| \geq 2t$.

Margin of Error

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq t) = \mathbb{P}(|X - \mathbb{E}[X]| \geq t/2) \leq 2\exp(-2nt^2) \leq .05$$

For $n = 1000$, we get:

$$2\exp\left(-2n\left(\frac{t}{2}\right)^2\right) \leq .05 \Rightarrow -\frac{2000t^2}{4} \leq \ln(.025) \Rightarrow t \leq .086.$$

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq .086) \leq .05$$

So our margin of error is about 8.6%.

To get a margin-of-error of 5% need $2\exp\left(-2n\left(\frac{.05}{2}\right)^2\right) \leq .05$

$$n \geq 2952$$

How much do we lose?

We lose a factor of two in the length of the margin (equivalently, we'd need to talk to 4 times as many people to have the same confidence.

You can also control this tradeoff.

Want more accuracy? Make it roll a die: report 1 if cheated (truth o/w)

Want more security? Make it Bernoulli with probability $p \gg \frac{1}{2}$ or cheated have the same report (e.g. report "die roll 1 [and didn't cheat]" or "die roll 2-6 [or did cheat]"

In The Real World

Injecting randomness to preserve privacy is a real thing.

Instead of having everyone flip a coin, “random noise” can be inserted after all the data has been collected.

Differential privacy is being used to protect the 2020 Census data.

The overall count of people in each state is exact (well, exactly the data they collected). But the data per block or per city will be randomized to protect against revealing who lives where.

This video nicely explains what’s involved. Notice that the accuracy guarantees come in the same “inside-margin-of-error-with-probability” guarantees we’ve been giving for our randomness (just much stronger).