Announcements

Make sure you have version 2 of HW8.

Real world 1 grades back
- regrade requests will open tomorrow (and stay open ~1 week)
- If you had a regrade on HW6, those are open now.

Logistical info for final on Ed
- release Monday morning at 9 am est
- out for 48 hours, can use full time
- can work in groups of 3, but still indiv. write ups

Application: Tail Bounds

Markov's Inequality

Let X be a random variable supported (only) on non-negative numbers. For any $t > 0$

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Chebyshev's Inequality

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

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

(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 $0 \leq \delta \leq 1$

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

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

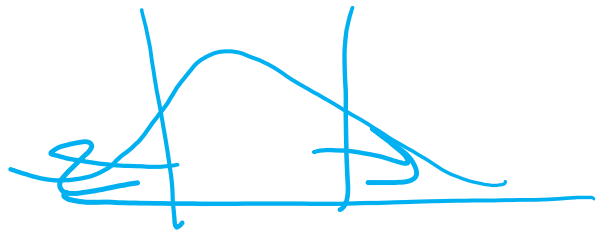
This is just a binomial!

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

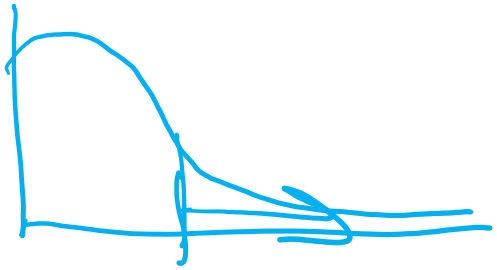
What happens when n gets big?

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

Next Time

One more bound (the union bound)

Not a concentration bound -- one more tool for handling non-independence.

We'll see it in the context of some applications!

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

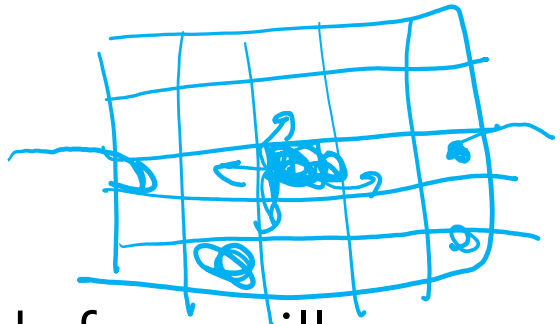
And $\mathbb{P}(E \cap F) \geq 0$.

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



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 $25 \cdot 0.015 \leq 0.375$.

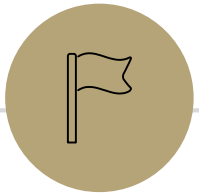
Tail Bounds – Takeaways

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 protocal 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 spouses.

We know about polling accuracy!

Do a poll, call up a random sample of married adults and ask them "have you ever cheated on your spouse?"

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 on their spouse.

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 on your spouse, please tell me "heads"

If the coin is tails and you have not ever cheated on your spouse, please tell me "tails"

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!"

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

Is this a substantial change?

No. For real world values (~15%) of $\mathbb{P}(C)$, the probability estimate would increase (to ~23%). 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 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?

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

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

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

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

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

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

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 $0 \leq \delta \leq 1$

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

What happens with $\mathbf{n = 1000}$ 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 $t \geq 0$

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

$|X - \mathbb{E}[X]| \geq t$ 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 $t \geq 0$

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

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

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

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 .

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).