Concentration Inequalities and Law of Large Numbers

CS70: Discrete Mathematics and Probability Theory

UC Berkeley – Summer 2026

Lecture 22

Ref: Note 19

Today

Concentration Inequalities: Probability of deviation from mean

Markov's Inequality: Used to bound $\mathbb{P}[X \geq c]$

- X must be a non-negative random variable (i.e., $\mathbb{P}[X < 0] = 0$)
- All you need to know about X is $\mathbb{E}[X]$
- Best bound possible if all you know is $\mathbb{E}[X]$, but often not good enough

Chebyshev's Inequality: Used to bound $\mathbb{P}[|X - \mu| \geq c]$

- Uses $\text{Var}(X)$ as indication of "spread" of distribution
- Best possible bound if all you know is $\mathbb{E}[X]$ and $\text{Var}(X)$
- Gives good results for estimating probabilities and polling
- Confidence intervals

Law of Large Numbers

Motivation: Polling

We want to know: What fraction of people think bears are great?

"People" left vague... all humans? Americans? Registered voters? Likely voters?

$\longleftarrow$ That's one vote...

Solution 1: Ask every person – practical?

Solution 2: Estimate by random sampling

Sample *uniformly*

Sample space: Ω is people   Event $B \subseteq \Omega$ people who think bears are great

Then $\mathbb{P}[B] = \frac{|B|}{|\Omega|}$ The fraction of people who think bears are great!

Can we estimate $\mathbb{P}[B]$?

Enough about bears...

We have iid random variables $X_i \sim \text{Bernoulli}(p)$

Can measure X_is and use to estimate unknown p

Motivation: Coin Flipping Context

You have a (possibly) biased coin probability of heads p.
 $\Rightarrow$ You don't know p, but you want to determine it.

Method: Toss the coin n times. Let $\hat{p}$ be the fraction heads over total flips.
 $\Rightarrow$ Sensible intuition: $\hat{p}$ should be close to p.

We want bounded error, say $|p - \hat{p}| < \varepsilon$ for some small ε

Can we guarantee this?

What we *can* get: $|p - \hat{p}| < \varepsilon$ with confidence $1 - \delta$

Example: $|p - \hat{p}| < \varepsilon$ with 95% confidence ($\delta = 0.05$)

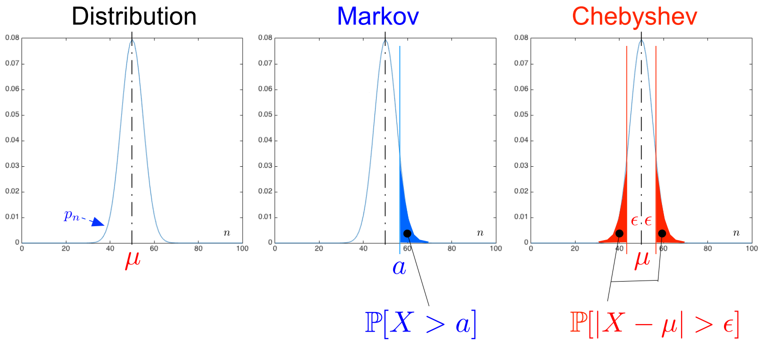
Or in terms of probabilities, we want $\mathbb{P}[|p - \hat{p}| < \varepsilon] \geq 1 - \delta$

A peak ahead: Using $n \geq \frac{1}{4\varepsilon^2\delta}$ coin flips is sufficient

Example: if $\delta = 0.05$ and $\varepsilon = 0.04$, sample size $n = \frac{1}{4(0.04)^2 0.05} = 3125$
Think "this poll has a margin of error of 4%"

Inequalities: An Overview

Idea: We want to bound probability of being "too far off" – far from mean



Markov: Bound probability that value of RV is much bigger than mean

Chebyshev: Bound probability that distance to mean is big

This kind of bound is sometimes called a tail bound (tail of the distribution)

Statistics Thought Experiment Number 1

Setting: 100 students take a midterm with average score $\mu = 45$
Scores are $s_1, s_2, \ldots, s_{100}$

Question: Can all 100 students score greater than 45?

If each $s_i > 45$, then we'd have

$$\mu = \frac{1}{100}\sum_{i=1}^{100} s_i > \frac{1}{100}\sum_{i=1}^{100} 45 = \frac{1}{100}(100 \cdot 45) = 45.$$

But $\mu > 45$ is a contradiction, so this is impossible.

Informally: everyone can't be above average (the "Lake Wobegon Theorem")

Everyone *could* be average though (i.e., $s_i = 45$ for all i)

... and if one student is *above* average, one must be *below* average

Thought Experiment Number 2

Setting: 100 students take a midterm with average score $\mu = 45$

Question: Can *more than* 50 students score at least $2\mu = 90$?

If each $s_i \geq 90$ for $i = 1, 2, \ldots, 51$ (might as well be first ones...), then we'd have

$$\mu = \frac{1}{100}\sum_{i=1}^{100} s_i \geq \frac{1}{100}\sum_{i=1}^{51} s_i \geq \frac{1}{100}\sum_{i=1}^{51} 90 = \frac{51 \cdot 90}{100} = \frac{102 \cdot 45}{100} > 45.$$

But $\mu > 45$ is a contradiction, so this is impossible.

Important condition: No negative scores... (on an exam? yikes!)

Informally: *More than* half of the students can't be at least twice the average.

Exactly half could score 90 though – as long as the other half scored zeros.

Probability Thought Experiment

Setting: Non-negative discrete random variable X with $\mathbb{E}[X] = 45$

Question: Can we have $\mathbb{P}[X \geq 90] > \frac{1}{2}$?

Consider conditional probabilities:

If $0 \leq X < 90$, then $X \geq 0$, so $\mathbb{E}[X \mid 0 \leq X < 90] \geq$

If $X \geq 90$, then $X \geq 90$, so $\mathbb{E}[X \mid X \geq 90] \geq$

Using Law of Total Expectation:

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X \mid 0 \leq X < 90] \cdot \mathbb{P}[0 \leq X < 90] + \mathbb{E}[X \mid X \geq 90] \cdot \mathbb{P}[X \geq 90] \\
&\geq 0 \cdot \mathbb{P}[0 \leq X < 90] + 90 \cdot \mathbb{P}[X \geq 90] \\
&> 90 \cdot \tfrac{1}{2} = 45
\end{aligned}$$

But $\mathbb{E}[X] > 45$ is a contradiction, so $\mathbb{P}[X \geq 90] > \frac{1}{2}$ is impossible.

Since that's impossible, we know $\mathbb{P}[X \geq 90] \leq \frac{1}{2}$

Important: We know nothing about distribution of X other than that it's non-negative.

Markov's Inequality

Generalizing the preceding thought experiments...

Theorem (Markov's Inequality): For *non-negative* random variable X with finite expected value, and any $c > 0$,

$$\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c} .$$

Proof: Using Law of Total Expectation:

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X \mid 0 \leq X < c] \cdot \mathbb{P}[0 \leq X < c] + \mathbb{E}[X \mid X \geq c] \cdot \mathbb{P}[X \geq c] \\ &\geq 0 \cdot \mathbb{P}[0 \leq X < c] + c \cdot \mathbb{P}[X \geq c] \\ &= c \cdot \mathbb{P}[X \geq c] \end{aligned}$$

It follows that $\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$. $\square$

Example: $\mathbb{E}[X] = 45$ and $c = 90$:

$$\mathbb{P}[X \geq 90] = \mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c} = \frac{45}{90} = \frac{1}{2} \qquad \text{(i.e., } \mathbb{P}[X \geq 90] \leq \tfrac{1}{2}\text{)}$$

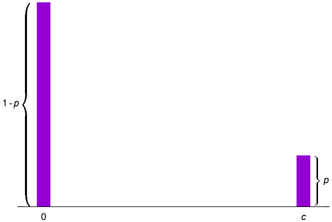
How Tight is the Markov Bound

Context: Sometimes (examples soon!) Markov is not a tight or useful bound.

Question: Can we use a smaller multiple of $\mathbb{E}[X]$ as a tighter bound?

Consider a random variable that can only takes values 0 or c, with

$$\mathbb{P}[X = c] = p$$
$$\mathbb{P}[X = 0] = 1 - p$$



Recall exams: “Half could score exactly 90 – if the other half score zero”

Expected value of this distribution: $\mathbb{E}[X] = 0 \cdot (1 - p) + c \cdot p = c \cdot p$

$\Rightarrow$ Markov bound is $\frac{\mathbb{E}[X]}{c} = \frac{c \cdot p}{c} = p \;\Rightarrow\; \mathbb{P}[X \geq c] \leq p$

Exact value of $\mathbb{P}[X \geq c] = p$

$\Rightarrow$ So Markov bound is tight for this distribution.

Markov's Inequality: Example 1

Markov's Inequality: For non-negative X, $\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$.

Setting: A randomized algorithm has running time T (a random variable, measured in minutes) with $\mathbb{E}[T] = 10$.

Question: What is the probability that it runs for over an hour?

What do we need to ask to see if we can use Markov's Inequality?

So plug in $c = 60$ to Markov's inequality, we get $\mathbb{P}[T \geq 60] \leq \frac{\mathbb{E}[T]}{c} =$.

What's great about this: Only need $\mathbb{E}[T]$ – no details about distribution...

What's bad about this: No additional info about T is used...

And a warning: How do you get $\mathbb{E}[T]$? It's tempting to run the program a few times (or a lot...) and take the average.

- $\Rightarrow$ *However*, this only estimates $\mathbb{E}[T]$ and can be inaccurate.
- $\Rightarrow$ Estimating $\mathbb{E}[T]$ is like estimating the probability of biased coin.

Markov's Inequality: Example 2

Markov's Inequality: For non-negative X, $\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$.

Setting: You have a biased coin and you want to estimate probability of heads (...or fraction of people who think bears are great).

Question: What is the probability of heads for each toss?

Let $X_n =$"number of heads in n tosses" ; $X \sim$ Binomial(n,p)

To estimate p, divide X by number of tosses (n): $A_n = \frac{1}{n}X_n$

$\Rightarrow \quad \mathbb{E}[A_n] = \mathbb{E}[\frac{1}{n}X_n] = \frac{1}{n}(np) = p$

Use Markov to bound probability of multiples of p: $\mathbb{P}[A_n \geq kp] \leq \frac{\mathbb{E}[A_n]}{kp} = \frac{p}{kp} = \frac{1}{k}$

Correct? Yes. Useful? Generally no.

- $\Rightarrow$ What's the probability our estimate A_n is too high? $\mathbb{P}[A_n \geq p] \leq 1$
- $\Rightarrow$ What's the probability our estimate is over twice p? $\mathbb{P}[A_n \geq 2p] \leq \frac{1}{2}$
- $\Rightarrow$ When coin is "close to fair" ($p \approx \frac{1}{2}$)? Pretty useless...

Markov's Inequality: Example 3

Markov's Inequality: For non-negative X, $\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$.

Setting: You toss a fair coin n times, with H being number of heads.

Question: What is $\mathbb{P}[H \geq \frac{3}{4}n]$?

Is H non-negative? Yes.

It's a fair coin (i.e., Binomial($n, \frac{1}{2}$)), so by Markov's inequality:

$$Pr[H \geq \frac{3}{4}n] \leq \qquad\qquad\qquad .$$

This is not a good bound... think about:

- $\Rightarrow$ Doesn't even establish this is unlikely (probability $< 1/2$)
- $\Rightarrow$ Doesn't depend on n – shouldn't it get more unlikely for larger n?

In the last two examples, Markov's Inequality didn't give a useful bound.

- $\Rightarrow$ So what's the use of Markov?
- $\Rightarrow$ Used to prove a different bound, which is much better!

Chebyshev's Inequality

Markov uses expectation and no other information – can we use variance?

Theorem (Chebyshev's Inequality): Let X be a random variable with mean $\mu = \mathbb{E}[X]$, then for any $c > 0$

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

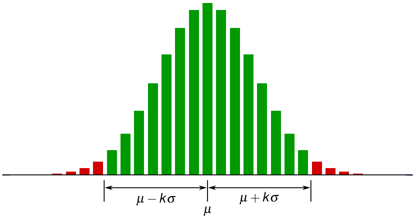
Proof: Define random variable $Y = (X - \mu)^2$, so $\mathbb{E}[Y] = \text{Var}(X)$. We use Markov's inequality with this (non-negative) variable to get

$$\mathbb{P}[|X - \mu| \geq c] = \underbrace{Pr[Y \geq c^2] \leq \frac{\mathbb{E}[Y]}{c^2}}_{\text{Markov}} = \frac{\text{Var}(X)}{c^2}. \qquad \square$$

Informal: Smaller variance makes it less likely that you're far from the mean.

Chebyshev's Inequality: Continued

Chebyshev is a "Two-Sided Bound" – above and below mean.



Sensible to express distance from mean as "number of standard deviations"

Using σ for standard deviation (so $\mathrm{Var}(X) = \sigma^2$) and $c = k\sigma$, Chebyshev is:

$$\mathbb{P}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

Probability drops as inverse square for every "standard deviation" of distance.

Chebyshev's Inequality: Example 1

Chebyshev's Inequality: For X with $\mu = \mathbb{E}[X]$, $\mathbb{P}[|X-\mu| \geq c] \leq \frac{\mathrm{Var}(X)}{c^2}$.

Setting: You toss a fair coin n times, with H being number of heads.

Question: What is $\mathbb{P}[H \geq \frac{3}{4}n]$? ... two-sided actually bounds $\mathbb{P}[|H - \frac{n}{2}| \geq \frac{n}{4}]$

We know: $H \sim \mathrm{Binomial}(n, \frac{1}{2})$

$\Rightarrow$ What is $\mathrm{Var}(H)$?

With $c = \frac{n}{4}$, Chebyshev gives

$$\mathbb{P}\left[\left|H - \frac{1}{2}n\right| \geq \frac{n}{4}\right] \leq \frac{n/4}{(n/4)^2} = \frac{4}{n}.$$

Much better!

$\Rightarrow$ For $n = 100$, the probability bound is 0.04 – unlikely! (Markov gave 2/3)

$\Rightarrow$ ... and really half this (careful! not all distributions are symmetric)

$\Rightarrow$ Reflects that this is increasingly unlikely as n gets bigger.

Both problems with using Markov's Inequality have been fixed!

Chebyshev's Inequality: Example 1 continued

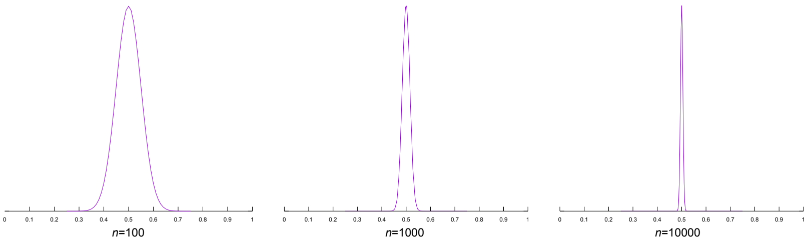
Chebyshev's Inequality: For X with $\mu = \mathbb{E}[X]$, $\mathbb{P}[|X - \mu| \geq c] \leq \frac{\mathrm{Var}(X)}{c^2}$.

Setting: You toss a fair coin n times, with H being number of heads.

Last slide: Calculated a bound on $\mathbb{P}[|H - \frac{n}{2}| \geq \frac{n}{4}$.

What about tighter around the mean? $\mathbb{P}[|H - \frac{1}{2}n| \geq \frac{n}{200}] \leq \frac{n/4}{(n/200)^2} = \frac{10000}{n}$
Pretty useless for small n, but still $\to 0$ as $n \to \infty$.

Distribution shape of $\frac{1}{n}H$ as n grows:



Higher $n \longrightarrow$ tighter distribution. In the limit?

Weak Law of Large Numbers (LLN)

Theorem (Weak Law of Large Numbers): Let $X_1, X_2, \ldots$ be iid random variables with $\mathbb{E}[X_i] = \mu$. Define $S_n = X_1 + X_2 + \cdots + X_n$ for all $n \geq 1$. Then

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{1}{n}S_n - \mu\right| \geq \varepsilon\right] = 0 \; .$$

Important: *Any* distribution with finite mean – not just "coin flips."

Proof: Let $A_n = \frac{1}{n}S_n$, and use Chebyshev and independence of X_i's to bound

$$\mathbb{P}[|A_n - \mu| \geq \varepsilon] \leq \frac{\text{Var}(A_n)}{\varepsilon^2} = \frac{\frac{1}{n^2}\text{Var}(X_1 + \cdots + X_n)}{\varepsilon^2} = \frac{\frac{1}{n}\cdot\text{Var}(X_1)}{\varepsilon^2} = \frac{Var(X_1)}{n\varepsilon^2} \; .$$

The limit of this last formula goes to 0 as $n \to \infty$. $\square$

The notes just call this the "Law of Large Numbers," but this is the weak form.

The "strong" form puts the limit on the average instead of probability:

$\mathbb{P}[\lim_{n\to\infty} \frac{1}{n}S_n = \mu] = 1$ – we won't use this.

Estimating Expected Value – General Bound

Setting: You have a stream of i.i.d. random variables $X_1, X_2, \dots$ (samples)

Goal: You want to estimate $\mu = \mathbb{E}[X_1]$ (same as $\mathbb{E}[X_i]$ for any i)

From LLN proof: Take n samples and use the average to estimate $\mathbb{E}[X_1]$.

$\Rightarrow$ Specifically, let $A_n = \frac{X_1 + \cdots + X_n}{n}$ – this is our estimate for μ

$\Rightarrow$ Proof (previous slide) derived $\mathbb{P}[|A_n - \mu| \geq \varepsilon] \leq \frac{\mathrm{Var}(X_1)}{n\varepsilon^2}$

Given ε and δ, can we pick n so that $\mathbb{P}[|A_n - \mu| \geq \varepsilon] \leq \delta$?

We want $\frac{\mathrm{Var}(X_1)}{n\varepsilon^2} \leq \delta$, which is met when $n \geq \frac{\mathrm{Var}(X_1)}{\delta\varepsilon^2}$

Summary: For any $\delta, \varepsilon > 0$, if $n \geq \frac{\mathrm{Var}(X_1)}{\delta\varepsilon^2}$, then

$$\mathbb{P}[|A_n - \mu| \geq \varepsilon] \leq \delta \ .$$

Problem: We don't know the distribution of the X_1's, so don't know $\mathrm{Var}(X_1)$.

$\Rightarrow$ There are some cases we can handle though!

Estimating the Probability of a Biased Coin

Setting: You have a stream of i.i.d. coin flips $X_1, X_2, \ldots$ (1 means "heads")

$\Rightarrow$ So each $X_i \sim \text{Bernoulli}(p)$ for some p

$\Rightarrow$ Or an "opinion poll" – picking random people to ask if bears are great

Goal: You want to estimate p ; in this case, $\mu = \mathbb{E}[X_1] = p$

As on previous slide: Use $A_n = \frac{X_1 + \cdots + X_n}{n}$ to estimate $\mu = p$

$\Rightarrow$ In our setting: flip n times, count heads, divide by n

Previous slide: Given ε and δ, pick $n \geq \frac{\text{Var}(X_1)}{\delta\varepsilon^2}$ to guarantee $\mathbb{P}[|A_n - p| \geq \varepsilon] \leq \delta$

So what about $\text{Var}(X_1)$? $X_1 \sim \text{Bernoulli}(p)$ so $\text{Var}(X_1) = p(1-p)$

$\Rightarrow$ Maximize $\text{Var}(X_1)$: max when $p = \frac{1}{2}$, giving max $\text{Var}(X_1) = \frac{1}{4}$

Conclusion: We have $\mathbb{P}[|A_n - p| \geq \varepsilon] \leq \delta$ when $n \geq \frac{1}{4\delta\varepsilon^2}$

What percentage of people think bears are great?

$\Rightarrow$ 95% chance that we get right percentage $\pm 3\%$: $\delta = 0.05$ and $\varepsilon = 0.03$

$\Rightarrow$ Ask a random sample of $n = \left\lceil \frac{1}{4 \cdot 0.05 \cdot (0.03)^2} \right\rceil = 5,556$ people

Polling: Some Numbers

Recap: To get $\mathbb{P}[|A_n - p| \geq \varepsilon] \leq \delta$, use $n \geq \frac{1}{4\delta\varepsilon^2}$

Confidence (δ)	Accuracy (ε)	Min n	
0.05	0.03	5,556	
0.03	0.05	3,334	*Confidence is cheaper than accuracy*
0.02	0.03	13,889	*... but high confidence still costs*
0.05	0.01	50,000	*very high accuracy*
0.01	0.05	10,000	*vs very high confidence*
0.01	0.01	250,000	*both very high*

Note: Published political polls usually only mention ε ("margin of error")

$\Rightarrow$ Typical standard is $\delta = 0.05$ (i.e., 95% confidence)

Final comments:

$\Rightarrow$ Population size doesn't matter – same n whether Canada (40M) or China (1.4B)
"3% of Canada" is different from "3% of China" however!

$\Rightarrow$ This is a sufficient condition for n – can we do better? (Yes – later)

$\Rightarrow$ Big difficulty: Getting a uniform sample of a population

$\Rightarrow$ Do people tell the truth?

Estimating a General Expectation: Relative Error

Setting: You have a stream of i.i.d. random variables $X_1, X_2, \ldots$ (samples)

Goal: You want to estimate $\mu = \mathbb{E}[X_1]$ (same as $\mathbb{E}[X_i]$ for any i)
What we've done: With $n \geq \frac{\text{Var}(X_1)}{\delta\varepsilon^2}$ we get $\mathbb{P}[|A_n - \mu| \geq \varepsilon] \leq \delta$

Is absolute error the right goal?

$\Rightarrow$ Maybe "within 25%" is better

$\Rightarrow$ For small $\mathbb{E}[X_1]$ not much difference

$\Rightarrow$ We already know everything we need!

Goal: Guarantee $\mathbb{P}[|A_n - \mu| \geq \varepsilon\mu] \leq \delta$ (e.g, no more than 25% from mean)

Previous result, but with $\varepsilon\mu$ for error:

With $n \geq \frac{\text{Var}(X_1)}{\delta(\varepsilon\mu)^2}$, we guarantee $\mathbb{P}[|A_n - \mu| \geq \varepsilon\mu] \leq \delta$

If σ is standard deviation of X_1, then $\text{Var}(X_1) = \sigma^2$

Use $n \geq \frac{\sigma^2}{\mu^2} \cdot \frac{1}{\delta\varepsilon^2}$ (this is what the notes give)

Just a notation change, but can change how you think about a problem!

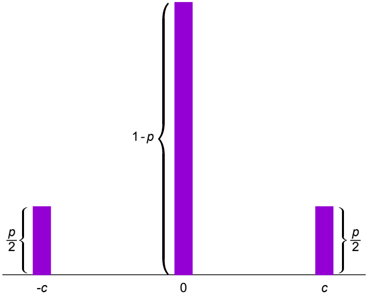
Example use: Running time of a program (takes 25% longer than average?)

Tightness of Chebyshev

Recall: Markov's inequality is tight in terms of expected value

Question: Is Chebyshev's Inequality tight in terms of variance?

$$\mathbb{P}[X=k] = \begin{cases} \frac{p}{2}, & \text{if } k=-c \\ 1-p & \text{if } k=0 \\ \frac{p}{2}, & \text{if } k=+c \end{cases}$$



Basic measures of this distribution:

$\Rightarrow$ $\mathbb{E}[X]=\mu=0$ (notice probabilities are symmetric around 0)

$\Rightarrow$ $\text{Var}(X)=\mathbb{E}[X^2]=\frac{p}{2}\cdot c^2+\frac{p}{2}\cdot c^2=p\cdot c^2$

Chebyshev Inequality gives $\mathbb{P}[|X-\mu|\geq c]\leq\frac{Var(X)}{c^2}=\frac{p\cdot c^2}{c^2}=p$

Exact probability $\mathbb{P}[|X-\mu|\geq c]=p$ – Chebyshev is tight!

Summary

Markov's Inequality: $\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$

- X must be a non-negative random variable (i.e., $\mathbb{P}[X < 0] = 0$)
- All you need to know about X is $\mathbb{E}[X]$
- Best bound possible if all you know is $\mathbb{E}[X]$, but often not good enough

Chebyshev's Inequality: $\mathbb{P}[|X - \mu| \geq c] \leq \frac{\mathrm{Var}(X)}{c^2}$

- Uses $\mathrm{Var}(X)$ as indication of "spread" of distribution
- Best possible bound if all you know is $\mathbb{E}[X]$ and $\mathrm{Var}(X)$
- Gives good results for estimating probabilities and polling
- Confidence intervals – estimating $\mathbb{E}[X]$ with average of n samples A_n:

  $n = \frac{Var(X)}{\delta\varepsilon^2}$ guarantees $\mathbb{P}[|A_n - \mu| \geq \varepsilon] \leq \delta$

  Estimating probability p (polling): Use $n = \frac{1}{4\delta\varepsilon^2}$

(Weak) Law of Large Numbers $\lim_{n\to\infty} \mathbb{P}[|A_n - \mu| \geq \varepsilon] = 0$