Estimating p for X~Bernoulli(p)

Lecture 23, CS70 Summer 2025

Motivating Problem of the Day

Zaphod Beeblebrox is running for President of the Universe.

We want to estimate how many people will vote for Zaphod.

One approach: Ask random people.

- As long as we ask enough of them, we'll get an accurate result.
- But how many people do we need to ask?

Motivating Problem of the Day: Samples and Coin Tosses

Total number of voters: v Actual number of Zaphod voters: z

Assuming: A person's vote is fixed (they don't change) and they don't lie...

Pick a voter at random – this is a "sample"

- Probability that this is a Zaphod-voter is z/v
- This is a Bernoulli trial with "success" probability $p = z/v$

Same problem: Estimate the heads probability p of a biased coin.

We switch to this terminology for the following slides.

Experimental Procedure for Determining p

Suppose you have a coin with probability of heads p. You don't know p, but you want to figure it out experimentally.

Method: Toss the coin n times. Let h be the number of heads and let $\hat{p} = \frac{h}{n}$.

What we want: $|p - \hat{p}| < \epsilon$

- Unfortunately, we can never guarantee this!

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

- Example: $|p - \hat{p}| < \epsilon$ with 95% confidence ($\delta = 0.05$)
- Or in terms of probabilities, we'll settle for $P(|p - \hat{p}| < \epsilon) \geq 1 - \delta$

Today's Answer (Spoiler)

To achieve our desired goal: $P(|p - \hat{p}| \leq \epsilon) \geq 1 - \delta$

We can show that $n \geq \frac{1}{4\epsilon^2\delta}$ coin flips is sufficient.

- We'll use something called Chebyshev's inequality to derive this later.

Example: If $\delta = 1\%$ and $\epsilon = 0.05$, then target n is $\frac{1}{4(0.05)^2 0.01} = 10{,}000$

- That is, if we flip the coin 10,000 times, then there is only a 1% chance that our estimate $\hat{p}$ is off by more than 0.05.
  - Example: If we get $\hat{p} = 0.43$, the bound says the chance is less than 1% that the true p lies outside the range $[0.38, 0.48]$.

Today's Answer (Spoiler)

To achieve our desired goal: $P(|p - \hat{p}| \leq \epsilon) \geq 1 - \delta$

We can show that $n \geq \frac{1}{4\epsilon^2\delta}$ coin flips is sufficient.

- We'll use something called Chebyshev's inequality to derive this later.

Example: If $\delta = 1\%$ and $\epsilon = 0.05$, then target n is $\frac{1}{4(0.05)^2 0.01} = 10{,}000$

- That is, if we flip the coin 10,000 times, then there is only a 1% chance that our estimate $\hat{p}$ is off by more than 0.05.
- I observe that it is also true that $\frac{1}{4(0.01)^2 0.25} = 10{,}000$. As a result, what else can I say about $\hat{p}$?

Today's Answer (Spoiler)

To achieve our desired goal: $P(|p - \hat{p}| \leq \epsilon) \geq 1 - \delta$

We can show that $n \geq \frac{1}{4\epsilon^2\delta}$ coin flips is sufficient.

- We'll use something called Chebyshev's inequality to derive this later.

Example: If $\delta = 1\%$ and $\epsilon = 0.05$, then target n is $\frac{1}{4(0.05)^2 0.01} = 10{,}000$

- I observe that it is also true that $\frac{1}{4(0.01)^2 0.25} = 10{,}000$.
  - There is a less than 25% chance we're off by more than 0.01.
  - Example: If $\hat{p} = 0.4$, there is a $\geq 75\%$ chance the true p is in the range $[0.39, 0.41]$.

In Terms of Election Polling

To achieve our desired goal: $P(|p - \hat{p}| \leq \epsilon) \geq 1 - \delta$

We can show (later!) that $n \geq \frac{1}{4\epsilon^2\delta}$ coin flips is sufficient.

(Surprising) Observation: It doesn't matter how big the population of the country is, the number of people we need to poll to get a desired degree of accuracy is constant.

- Ask 10,000 people, and our bound says we have a 99% chance of being within 0.05 of p. True if we're surveying Canada (40M) or China (1.4B).

Later: why the actual practice of polling is more difficult than this.

Deriving this Expression and Concentration Inequalities

I've given this formula without any proof or rationale!

$$n \geq \frac{1}{4\epsilon^2\delta}$$

We'll derive this result using a powerful idea called a "concentration inequality".

- A concentration inequality (of "tail bound") is a mathematical bound on how likely a random variable can stray from some quantity.
- Example: For a non-negative random variable X with expectation 45, what is the probability that $X > 90$?
  - Surprisingly, we can bound this probability without knowing anything else about the RV (e.g., variance, distribution, etc).

Markov's Inequality

Lecture 23, CS70 Summer 2025

Thought Experiment 1

Imagine that $n = 100$ students take a midterm with average score $\mu = 45$.

- Let individual scores be $s_1, \dots, s_n$, so $\mu = \frac{1}{n}\sum_{i=1}^{n} s_i$
- Can all 100 students score greater than $\mu = 45$ on the midterm?

Thought Experiment 1

Imagine that $n = 100$ students take a midterm with average score $\mu = 45$.

- Let individual scores be $s_1, \ldots, s_n$, so $\mu = \frac{1}{n}\sum_{i=1}^{n} s_i$
- Can all 100 students score greater than $\mu = 45$ on the midterm?

No! This is basic "proof by contradiction" reasoning:

Assume for the sake of contradiction that $s_i > \mu$ for all i. Then

$$\mu = \frac{1}{n}\sum_{i=1}^{n} s_i > \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

So $\mu > \mu$, a contradiction. ■

The "Lake Wobegon Theorem"

Thought Experiment 2

Imagine that $n = 100$ students take a midterm with average score $\mu = 45$.

- Let individual scores be $s_1, \ldots, s_n$, so $\mu = \frac{1}{n} \sum_{i=1}^{n} s_i$
- Can $\frac{n}{2} = 50$ students score greater than $2\mu = 90$ on the midterm?

Thought Experiment 2

Imagine that $n = 100$ students take a midterm with average score $\mu = 45$.

- Let individual scores be $s_1, \dots, s_n$, so $\mu = \frac{1}{n}\sum_{i=1}^{n} s_i$
- Can $\frac{n}{2} = 50$ students score greater than $2\mu = 90$ on the midterm?

No! Again, assume for the sake of contradiction that $\frac{n}{2}$ students score $> 2\mu$.

Let H (with $|H| = \frac{n}{2}$) be the set of "high scorers", so $i \in H \implies s_i > 2\mu$

What about students who are not high-scorers? We only know $i \in \overline{H} \implies s_i \geq 0$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} s_i = \frac{1}{n}\left(\sum_{i\in H} s_i + \sum_{i\in \overline{H}} s_i\right) > \frac{1}{n}\left(\sum_{i\in H} 2\mu + \sum_{i\in \overline{H}} 0\right) = \frac{1}{n}\left(\frac{n}{2}\cdot 2\mu\right) = \mu$$

Again we get the contradiction $\mu > \mu$. ■

Thought Experiment 3

Consider a **non-negative** random variable with $E[X] = 45$.

- Claim: $P(X \geq 90) \leq \frac{1}{2}$.

Why?

Thought Experiment 3

Consider a **non-negative** random variable with $E[X] = 45$.

- Claim: $P(X \geq 90) \leq \frac{1}{2}$.

Let A be the event that $X \geq 90$, and assume for contradiction that $P(A) > \frac{1}{2}$

What can we say about conditional expectations?

- $E[X|A]$? $E[X|A] \geq 90$
- $E[X|\bar{A}]$? $E[X|\bar{A}] \geq 0$

So:

$$E[X] = E[X|A] \cdot P(A) + E[X|\bar{A}] \cdot P(\bar{A}) \geq 90 \cdot P(A) + 0 \cdot P(\bar{A}) > 90 \cdot \frac{1}{2} = 45$$

So we have $E[X] > 45$, a contradiction. ■

Markov's Inequality

Markov's Inequality: For a nonnegative random variable X (i.e., $X(\omega) \geq 0$ for all $\omega \in \Omega$) with finite mean, and any constant $c > 0$, we have:

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

Example: Suppose $E[X] = 45$, with $c = 90$: $P(X \geq 90) \leq 45/90 = 1/2$

- We saw this for these specific values already
- Reasoning: If $P(X \geq 90) > 1/2$, then these values would push $E[X]$ above 45.

Markov's Inequality

Markov's Inequality: For a nonnegative random variable X (i.e., $X(\omega) \geq 0$ for all $\omega \in \Omega$) with finite mean, and any constant $c > 0$, we have:

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

Proof:

$$E[X] = \sum_{\omega \in \Omega} X(\omega) \times P(\omega)$$

Assigning all "small" outcomes to zero. →

$$\geq \sum_{\omega: X(\omega) \geq c} X(\omega) \times P(\omega)$$

Assigning all "large" outcomes to c. →

$$\geq \sum_{\omega: X(\omega) \geq c} c \times P(\omega)$$

$$= c \sum_{\omega: X(\omega) \geq c} P(\omega) = c\, P(X \geq c)$$

Example Using Markov's Inequality 1: Money Giveaway

Markov's Inequality: For a nonnegative random variable X (i.e., $X(\omega) \geq 0$ for all $\omega \in \Omega$) with finite mean, and any constant $c > 0$, we have:

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

MacKenzie Scott is giving away random amounts of money to people, where X is a random variable (unknown distribution) for the amount given to someone. What is the probability that someone gets more than 5 times the average?

Here, $c = 5\,E[X]$, so we have:

$$P(X \geq 5\,E[X]) \leq \frac{E[X]}{5\,E[X]} = \frac{1}{5}$$

Example Using Markov's Inequality 2: Coin Toss

Markov's Inequality: For a nonnegative random variable X (i.e., $X(\omega) \geq 0$ for all $\omega \in \Omega$) with finite mean, and any constant $c > 0$, we have:

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

You toss a fair coin n times. Let X be the number of heads. What does Markov's Inequality tell us about the probability that $X \geq \frac{3}{4}n$?

Hint: What is $E[X]$, i.e., expected number of heads for n coin flips?

Example Using Markov's Inequality 2: Coin Toss

Markov's Inequality: For a nonnegative random variable X (i.e., $X(\omega) \geq 0$ for all $\omega \in \Omega$) with finite mean, and any constant $c > 0$, we have:

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

You toss a fair coin n times. Let X be the number of heads. What does Markov's Inequality tell us about the probability that $X \geq \frac{3}{4}n$?

$$E[X] = n/2$$

$$P\left(X \geq \frac{3}{4}n\right) \leq \frac{n/2}{3n/4} = \frac{1/2}{3/4} = \frac{1}{2} \cdot \frac{4}{3} = \frac{2}{3}$$

Example Using Markov's Inequality 2: Coin Toss

Markov's Inequality: For a nonnegative random variable X (i.e., $X(\omega) \geq 0$ for all $\omega \in \Omega$) with finite mean, and any constant $c > 0$, we have:

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

You toss a fair coin n times. Let X be the number of heads. What does Markov's Inequality tell us about the probability that $X \geq \frac{3}{4}n$?

$$P\left(X \geq \frac{3}{4}n\right) \leq \frac{2}{3}$$

For this example, is this bound "good" or perhaps "tight"?

- No, it says if you flip a coin 1,000,000 times, probability of getting 750,000 heads is less than 2/3. Actual probability is far lower.

Concentration Inequalities and Bound Tightness

You can't assume concentration inequalities are tight bounds!

- Just like the union bound, they are simply bounds on a probability.

In the case of a binomial random variable, Markov's Inequality is pretty useless.

- It is true that the probability of getting 3/4s heads is less than 2/3, but this fact is not useful.

Note: Markov's Inequality can actually be a tight bound.

- Consider RV where $X = \mu k$ with probability $1/k$, and otherwise $X = 0$. In this case, setting $c = \mu k$ we have $P(X \geq c) = 1/k = \mu/c$, and the Markov bound is tight at that point.

Markov's Inequality

Markov's Inequality: For a nonnegative random variable X (i.e., $X(\omega) \geq 0$ for all $\omega \in \Omega$) with finite mean, and any constant $c > 0$, we have:

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

What happens if we drop non-negativity?

- Naturally, Markov's Inequality as stated above is no longer true. Trivial example: one student gets -1,000,000 points on a midterm, then everyone else will do way better than twice the mean.

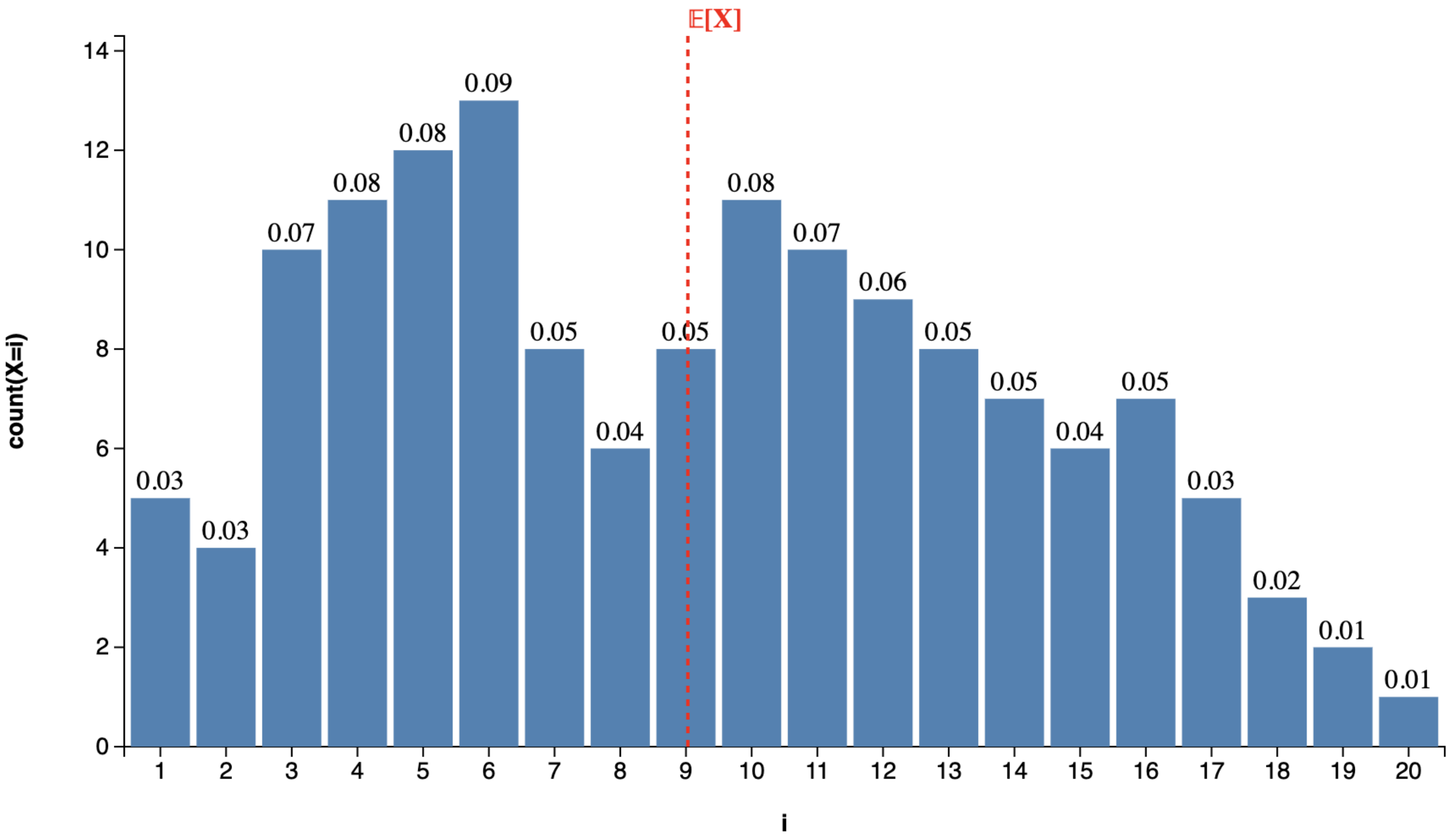
Chebyshev's Inequality

Lecture 23, CS70 Summer 2025

Expectations and Distributions (Visually)

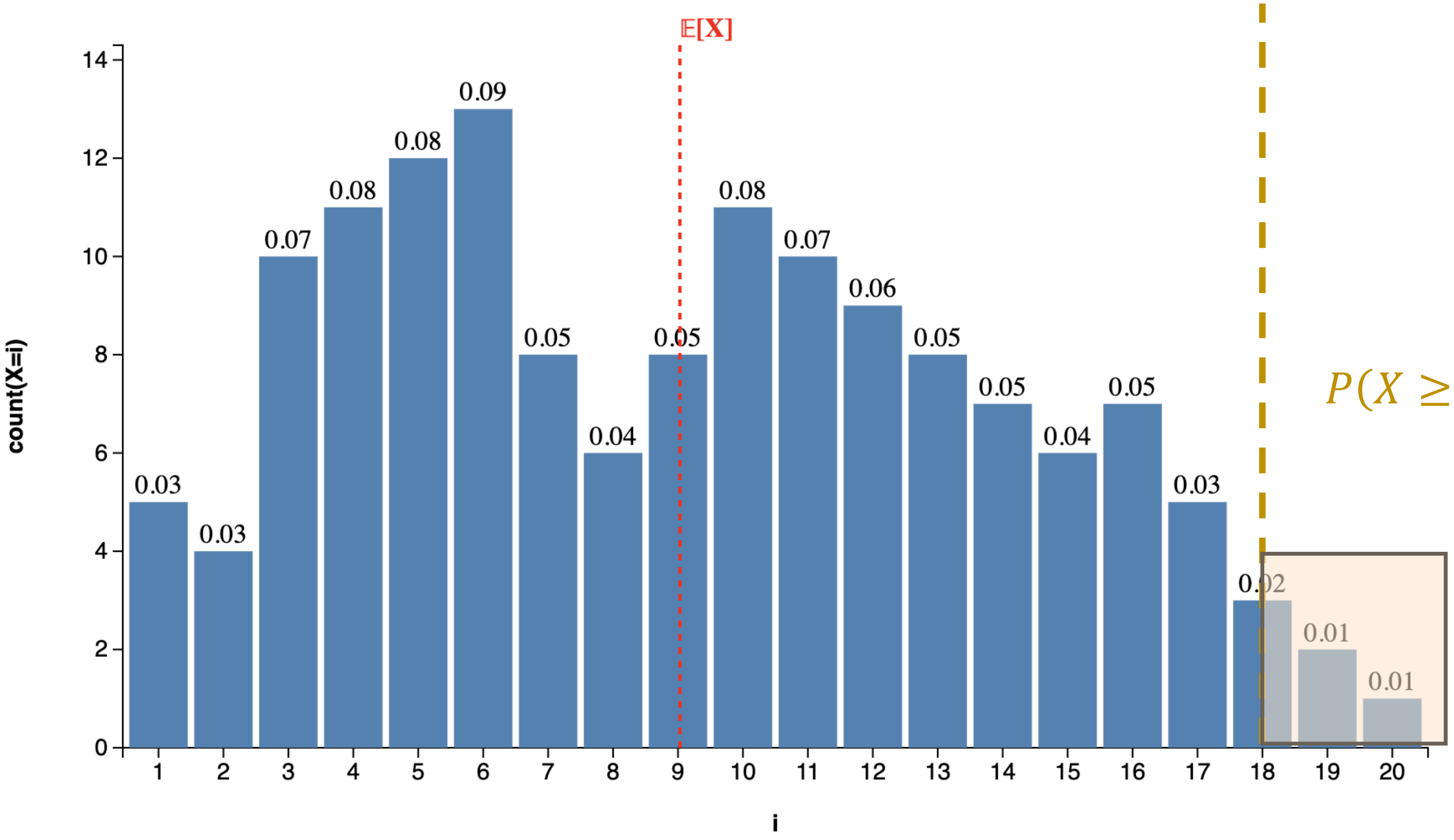
Consider the distribution below.

- We can think of the expectation as the "center of mass" of the distribution.
- Each value has mass equal to the height of the respective bar.



Markov's Inequality Visually

Markov's inequality gives us a bound on how much mass there can be above some constant c. Example if $c = 2\,E[X]$ shown below.

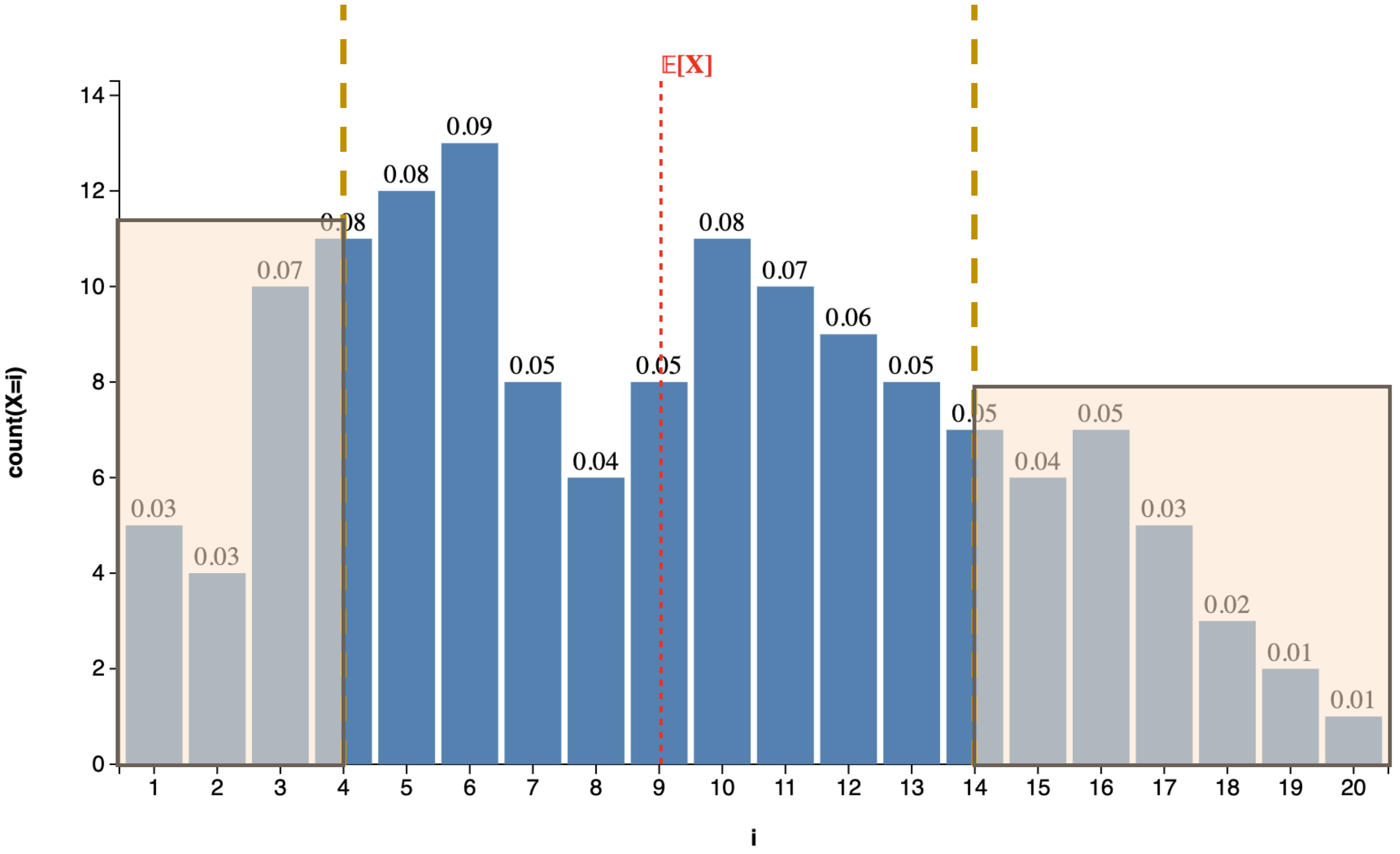


$P(X \geq 2\,E[X]) \leq 1/2$

Chebyshev's Inequality Visually

Observation: It might be more useful to have an inequality that bounds the amount of mass that is far from the mean (in either direction), and allows negative values for the rv.

$$P(|X - \mu| \geq c) \leq ??$$



Chebyshev's Inequality

We'll use two summary statistics: $E[X]$ and $\mathrm{var}(X)$ to provide a more useful bound on how far we stray from the mean.

Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X - \mu| \geq c) \leq \frac{\mathrm{var}(X)}{c^2}$$

Before we set about proving this inequality, let's build some intuition for what it means.

Chebyshev's Inequality: Intution 1, the Alternate Form

Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X - \mu| \geq c) \leq \frac{\text{var}(X)}{c^2}$$

Suppose we pick the constant $c = k\sigma$, what is the upper bound?

$$P(|X - \mu| \geq k\sigma) \leq ??$$

Chebyshev's Inequality

Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X - \mu| \geq c) \leq \frac{\text{var}(X)}{c^2}$$

Suppose we pick the constant $c = k\sigma$, what is the upper bound?

$$P(|X - \mu| \geq k\sigma) \leq \frac{\text{var}(X)}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Chebyshev's Inequality

Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X - \mu| \geq c\,) \leq \frac{\text{var}(X)}{c^2}$$

Or equivalently:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Suppose we have a random variable, what bound can we place on the probability that we are more than three standard deviations from the mean? More than one standard deviation from the mean?

Chebyshev's Inequality

Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X - \mu| \geq c\ ) \leq \frac{\text{var}(X)}{c^2}$$

Or equivalently:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Suppose we have a random variable, what bound can we place on the probability that we are more three and one standard deviations?

- The probabilities are ≤ 1/9 and 1, respectively.
- Note, for one standard deviation, the result is trivial, i.e., Chebyshev's tells us literally nothing. In general, Chebyshev's is a loose bound.

Chebyshev's is a Loose Bound: Normal Distribution Example

There's a distribution we'll talk about next week called the normal distribution.

- Probability that you are less than 3 standard deviations from the mean is 99.7%.
- Or equivalently: Probability that are you are more than 3 standard deviations from the mean is 0.3%.

Chebyshev's is considerably looser. It says the probability of being more than 3 standard deviations from the mean is 11.1%

Key difference: Chebyshev is true of all possible distributions. The 99.7% rule is only for the Normal.

Chebyshev Proof

Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X - \mu| \geq c) \leq \frac{\text{var}(X)}{c^2}$$

Dead End Proof: Observe that $|X - \mu|$ is a random variable.

- Since $|X - \mu|$ is non-negative, we could try to use Markov's inequality to try to bound $P(|X - \mu| \geq c)$, but expectation of an absolute value is a pain.

Chebyshev Proof

Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X - \mu| \geq c) \leq \frac{\text{var}(X)}{c^2}$$

Proof:

- Observe that $|X - \mu|$ is a random variable.
- Observe that any outcome ω, $|X(\omega) - \mu| \geq c$ if and only if $(X(\omega) - \mu)^2 \geq c^2$. This means that:

$$P(|X - \mu| \geq c) = P\left((X - \mu)^2 \geq c^2\right)$$

- Let $Y = (X - \mu)^2$.
- Then by Markov's inequality: $P(Y \geq c^2) \leq \frac{E[Y]}{c^2}$

Chebyshev Proof

Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X - \mu| \geq c\,) \leq \frac{\text{var}(X)}{c^2}$$

Proof:

- Observe that $|X - \mu|$ is a random variable.
- Observe that any outcome ω, $|X(\omega) - \mu| \geq c$ if and only if $(X(\omega) - \mu\,)^2 \geq c^2$. This means that:

$$P(|X - \mu| \geq c\,) = P\left((X - \mu)^2 \geq c^2\,\right)$$

- Let $Y = (X - \mu)^2$.
- Then by Markov's inequality: $P(Y \geq c^2) \leq \dfrac{E[Y]}{c^2}$

What is E[Y]?

Chebyshev Proof

Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X-\mu| \geq c) \leq \frac{\text{var}(X)}{c^2}$$

Proof :

- Observe that $|X-\mu|$ is a random variable.
- Observe that any outcome ω, $|X(\omega)-\mu| \geq c$ if and only if $(X(\omega)-\mu)^2 \geq c^2$. This means that:

$$P(|X-\mu| \geq c) = P\left((X-\mu)^2 \geq c^2\right)$$

- Let $Y = (X-\mu)^2$. $E[Y] = \text{var}(X)$
- Then by Markov's inequality: $P(Y \geq c^2) \leq \frac{E[Y]}{c^2} = \frac{\text{var}(X)}{c^2}$

Chebyshev's Inequality and Markov's Inequality

Even though Markov's Inequality is weak for understanding X, we were able to enhance its power by considering a new random variable $Y = (X - \mu)^2$ and considering bounds on this variable.

- Yes! You can enhance the bound further by making up new random variables such as $(X - \mu)^{10}$, but then you need to be able to compute $E[(X - \mu)^{10}]$.

The Chebyshev Bound for Coin Flips

Lecture 23, CS70 Summer 2025

Application: Probability of Coin Toss Outcomes

Earlier, we tried to use Markov's Inequality to understand the behavior of $X \sim \text{Binomial}(n, 0.5)$.

Specifically, we asked: What is the probability that at least 3/4s of our flips come up heads.

Using Markov's Inequality, we found that $P\left(X \geq \frac{3}{4}n\right) \leq \frac{2}{3}$.

This bound was not useful! As we flip more coins, the actual probability drops.

- If we use Chebyshev's Inequality instead, we'll see a dependence on n.

Application: Probability of Coin Toss Outcomes

For $X \sim \text{Binomial}(n, 0.5)$, let's bound $P\left(X \geq \frac{3}{4}n\right)$, using Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X - \mu| \geq c\ ) \leq \frac{\text{var}(X)}{c^2}$$

First, we write expectation and variance: $E[X] = \frac{n}{2}$ $\quad \text{var}(X) = n \cdot \text{p} \cdot (1 - p) = \frac{n}{4}$

Then, get $P\left(X \geq \frac{3}{4}n\right)$ into the appropriate form to apply Chebyshev:

$$P\left(X \geq \frac{3}{4}n\right) = P\left(X - E[X] \geq \frac{3}{4}\text{n} - E[X]\right) = P\left(X - E[X] \geq \frac{n}{4}\right)$$

$$\leq P\left(|X - E[X]| \geq \frac{n}{4}\right)$$

Application: Probability of Coin Toss Outcomes

For $X \sim \text{Binomial}(n, 0.5)$, let's bound $P\left(X \geq \frac{3}{4}n\right)$, using Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X - \mu| \geq c\,) \leq \frac{\text{var}(X)}{c^2}$$

First, we write expectation and variance: $E[X] = \frac{n}{2}$ $\quad \text{var}(X) = n \cdot \text{p} \cdot (1 - p) = \frac{n}{4}$

Then, $P\left(X \geq \frac{3}{4}n\right) \leq P\left(|X - E[X]| \geq \frac{n}{4}\right)$

$$\leq \frac{\text{var}(X)}{\left(\frac{n}{4}\right)^2} = \frac{n/4}{\left(\frac{n}{4}\right)^2} = \frac{4}{n}$$

Application: Probability of Coin Toss Outcomes

For $X \sim \text{Binomial}(n, 0.5)$, let's bound $P\left(X \geq \frac{3}{4}n\right)$, using Chebyshev's Inequality: If X is a random variable with $E[X] = \mu$, then for any constant $c > 0$:

$$P(|X - \mu| \geq c) \leq \frac{\text{var}(X)}{c^2}$$

Our Chebyshev bound is: $P\left(X \geq \frac{3}{4}n\right) \leq \frac{4}{n}$

In other words, as we flip more coins, the chance of having 75% (or more) heads drops.

- For n = 1,000, this bound says probability is less than or equal to 0.4%
- Note: This is still a very loose bound! Actual probability is much lower.

Back to Estimating p for a Coin

We have a coin with unknown heads probability p. We estimate it by tossing a coin n times.

- X is the number of heads. $X = X_1 + X_2 + \cdots + X_n$.
- Our empirical mean is $\hat{p} = \frac{X}{n}$

Goal: How big does n need to be so that $P(|\hat{p} - p| \geq \epsilon) \leq \delta$

Observation, we can bound $P(|\hat{p} - p| \geq \epsilon)$ using Chebyshev's inequality.

- To show this, we'll need $E[\hat{p}]$.

$$E[\hat{p}] = E\left[\frac{X}{n}\right] = \frac{1}{n} E[X] = \frac{pn}{n} = p$$

Back to Estimating p for a Coin

We have a coin with unknown heads probability p. We estimate it by tossing a coin n times.

- X is the number of heads. $X = X_1 + X_2 + \cdots + X_n$.
- Our empirical mean is $\hat{p} = \frac{X}{n}$

Observation, we can bound $P(|\hat{p} - p| \geq \epsilon)$ using Chebyshev's inequality.

- Since $E[\hat{p}] = p$, we have that:

$$P(|\hat{p} - E[\hat{p}]| \geq \epsilon) = P(|\hat{p} - p| \geq \epsilon) \leq \frac{\mathrm{var}(\hat{p})}{\epsilon^2}$$

- So now we need to know $\mathrm{var}(\hat{p})$.

$$\mathrm{var}[\hat{p}] = \mathrm{var}\left(\frac{X}{n}\right) = \frac{\mathrm{var}(X)}{n^2} = \frac{\mathrm{var}(X_1 + \cdots + X_n)}{n^2} = \frac{n \cdot \mathrm{var}(X_1)}{n^2} = \frac{n \cdot p(1-p)}{n^2}$$

$$= \frac{p(1-p)}{n}$$

Back to Estimating p for a Coin

We have a coin with unknown heads probability p. We estimate it by tossing a coin n times.

- X is the number of heads. $X = X_1 + X_2 + \cdots + X_n$.
- Our empirical mean is $\hat{p} = \frac{X}{n}$

Observation, we can bound $P(|\hat{p} - p| \geq \epsilon)$ using Chebyshev's inequality.

- Since $E[\hat{p}] = p$, and $\mathrm{var}(\hat{p}) = p(1-p)/n$ we have that:

$$P(|\hat{p} - E[\hat{p}]| \geq \epsilon) = P(|\hat{p} - p| \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2}$$

At this point, it might seem like we're a bit stuck! The variance of $\hat{p}$ seems to depend on p, which is what we're trying to estimate! Any idea how to proceed?

Back to Estimating p for a Coin

We have a coin with unknown heads probability p. We estimate it by tossing a coin n times.

- X is the number of heads. $X = X_1 + X_2 + \cdots + X_n$.
- Our empirical mean is $\hat{p} = \frac{X}{n}$

Observation, we can bound $P(|\hat{p} - p| \geq \epsilon)$ using Chebyshev's inequality.

- Since $\mathbb{E}[\hat{p}] = p$, and $\mathrm{var}(\hat{p}) = p(1-p)/n$ we have that:

$$P(|\hat{p} - E[\hat{p}]| \geq \epsilon) = P(|\hat{p} - p| \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2}$$

Solution: Be conservative and replace $p(1-p)$ with its maximum possible value. What is it?

Back to Estimating p for a Coin

We have a coin with unknown heads probability p. We estimate it by tossing a coin n times.

- X is the number of heads. $X = X_1 + X_2 + \cdots + X_n$.
- Our empirical mean is $\hat{p} = \frac{X}{n}$

Observation, we can bound $P(|\hat{p} - p| \geq \epsilon)$ using Chebyshev's inequality.

- Since $\mathbb{E}[\hat{p}] = p$, and $\mathrm{var}(\hat{p}) = p(1-p)/n$ we have that:

$$P(|\hat{p} - E[\hat{p}]| \geq \epsilon) = P(|\hat{p} - p| \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$

Solution: Be conservative and replace $p(1-p)$ with its maximum possible value of 1/4, which we get when $p = 1/2$.

Back to Estimating p for a Coin

We have a coin with unknown heads probability p. We estimate it by tossing a coin n times.

- X is the number of heads. $X = X_1 + X_2 + \cdots + X_n$.
- Our empirical mean is $\hat{p} = \frac{X}{n}$

Using Chebychev's inequality we have shown: $P(|\hat{p} - p| \geq \epsilon) \leq \frac{1}{4n\epsilon^2}$

Earlier, our goal was to have our probability $P(|\hat{p} - p| \geq \epsilon) \leq \delta$

To achieve this goal, we need:

$$\frac{1}{4n\epsilon^2} \leq \delta \qquad \Leftrightarrow \qquad \frac{1}{4\delta\epsilon^2} \leq n$$

Weak Law of Large Numbers

Lecture 23, CS70 Summer 2025

Weak Law of Large Numbers

As a quick detour before wrapping up today, let's talk about the Weak Law of Large Numbers.

If $X_1, X_2, X_3, \ldots$ are independent and identically distributed (i.i.d.) random variables with $E[X_i] = \mu$, and $\mathrm{var}(X_i) = \sigma^2 < \infty$, then for every $\epsilon > 0$, we have that:

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

Informal interpretation: As we run the same experiment over and over, the difference between the empirical average and the expectation converges to zero.

Weak Law of Large Numbers

If $X_1, X_2, X_3, \ldots$ are independent and identically distributed (i.i.d.) random variables with $E[X_i] = \mu$, and $\text{var}(X_i) = \sigma^2 < \infty$, then for every $\epsilon > 0$, we have that:

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

Proof: Let $\sigma^2 = \text{var}(X_i)$. Let $S_n = X_1 + \cdots + X_n$. Let $Y_n = S_n/n$

Questions:

- What is $E[Y_n]$?
- What is $\text{var}(Y_n)$?

Weak Law of Large Numbers

If $X_1, X_2, X_3, \ldots$ are independent and identically distributed (i.i.d.) random variables with $E[X_i] = \mu$, and $\operatorname{var}(X_i) = \sigma^2 < \infty$, then for every $\epsilon > 0$, we have that:

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

Proof: Let $\sigma^2 = \operatorname{var}(X_i)$. Let $S_n = X_1 + \cdots + X_n$. Let $Y_n = S_n/n$

Questions:

- What is $E[Y_n]$? $E[Y_n] = E[(X_1 + \cdots + X_n)/n] = (E[X_1] + \cdots + E[X_n])/n = \mu n/n = \mu$
- What is $\operatorname{var}(Y_n)$?

$$\operatorname{var}(Y_n) = \operatorname{var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\operatorname{var}(X_1 + \cdots + X_n)}{n^2} = \frac{n\operatorname{var}(X_i)}{n^2} = \frac{\sigma^2}{n}$$

Weak Law of Large Numbers

If $X_1, X_2, X_3, \ldots$ are independent and identically distributed (i.i.d.) random variables with $E[X_i] = \mu$, and $\text{var}(X_i) = \sigma^2 < \infty$, then for every $\epsilon > 0$, we have that:

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

Proof: Let $\sigma^2 = \text{var}(X_i)$. Let $S_n = X_1 + \cdots + X_n$. Let $Y_n = S_n/n$

- $E[Y_n] = \mu$
- $\text{var}(Y_n) = \sigma^2/n$

Chebyshev's inequality tells us:

$$P(|Y_n - \mu| \geq \epsilon) \leq \frac{\text{var}(Y_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

So: $$\lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

Back to Polling

Lecture 23, CS70 Summer 2025

Back to the Polling Problem

Total number of voters: v Actual number of Zaphod voters: z

Assuming: A person's vote is fixed (they don't change) and they don't lie...

Pick a voter at random – this is a "sample"

- Probability that this is a Zaphod-voter is z/v
- This is a Bernoulli trial with "success" probability $p = z/v$

Randomly sample $n \geq \frac{1}{4\epsilon^2\delta}$ voters to get $P(|p - \hat{p}| \leq \epsilon) \geq 1 - \delta$

Back to the Polling Problem

Total number of voters: v Actual number of Zaphod voters: z

Randomly sample $n \geq \frac{1}{4\epsilon^2\delta}$ voters to get $P(|p - \hat{p}| \leq \epsilon) \geq 1 - \delta$

How can this go wrong? ... or: Why aren't polls always right (or at least better)?

- Are you really randomly sampling voters?

  What list are you using? Do all voters have phones? Do all answer their phones?
- Do people answer honestly?
- Do people never change their mind?
- Did you just get unlucky? (probability $\delta > 0$ of the bound not holding!)

Reminder: Real life is messy. We don't live in ideal mathematical models...