

COS 330: Great Ideas in Theoretical Computer Science

Fall 2026

Lecture 6

*Probability and Computation**Module: Randomization*

1 Motivation

Suppose we want to estimate the fraction of voters who support a party R , with all other voters supporting a party D . We randomly sample s potential voters and report the fraction of sampled voters who support R . How large should s be to estimate the true fraction to within one percentage point?

Consider polling in the United States and in India, whose populations are around 340 million and 1.5 billion, respectively (4.4 times larger). For the same accuracy and confidence, how much larger must the sample be in India than in the United States? We will return to this question after developing the tools to analyze sampling error.

Polling and coin tosses. If a fraction p of the population supports R , each independent uniform sample returns an R supporter with probability p . Sampling with replacement therefore behaves like tossing a coin whose probability of heads is p : an R supporter corresponds to heads, and a D supporter to tails. The fraction of heads plays the role of the estimated fraction of R voters.

The simplest case is a fair coin. In n independent tosses, roughly half should be heads, but the observed fraction need not be exactly one half. Our running example will ask for the probability of seeing at least 51% heads. In polling terms, this is the probability that the estimate is at least 51% when the true support is 50%. Studying how this probability decreases with the sample size will lead to a quantitative answer to the polling question.

1.1 Probability and computation

Consider the computational problem of estimating the fraction of ones in a fixed array $a_1, \dots, a_N \in \{0, 1\}$. An exact answer requires inspecting every entry in the worst case: an unread entry could change the answer. An approximate answer may require much less work. A sampling algorithm chooses s indices independently and uniformly from $\{1, \dots, N\}$, with replacement, and returns the fraction of inspected entries equal to one. The central question is how many entries it must inspect to achieve a given accuracy with a small probability of failure.

Polling fits this problem by representing each voter as an array entry: one for an R supporter and zero for a D supporter. Interviewing a voter corresponds to inspecting an entry, and the sample fraction is the algorithm's output. The population size becomes the input length; the sample size measures the number of entries accessed.

The input is fixed; the algorithm introduces the randomness through its choice of indices. A guarantee should hold for every fixed input array, with probability taken over these choices. We will develop the probability tools needed to analyze the accuracy of this algorithm and the computational cost of obtaining a reliable answer.

2 Probability Review

We work only with discrete probability spaces: their outcomes form a finite or countable set. All experiments in this lecture have finitely many outcomes, so all sums below are finite.

2.1 Sample spaces and events

Definition 2.1 (Finite probability space). A *sample space* Ω is the set of possible outcomes of an experiment. Each outcome $\omega \in \Omega$ has a probability mass $p(\omega) \geq 0$, with

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

For n coin tosses, $\Omega = \{H, T\}^n$: all length- n sequences of heads and tails. There are 2^n outcomes. For independent fair tosses, every sequence has probability 2^{-n} .

Definition 2.2 (Event). An *event* is a subset $A \subseteq \Omega$. Its probability is

$$\Pr[A] = \sum_{\omega \in A} p(\omega).$$

For example, “at least 51% heads” is the set of sequences containing at least $\lceil 0.51n \rceil$ heads.

Definition 2.3 (Conditional probability). For events A, B with $\Pr[B] > 0$, the probability of A conditional on B is

$$\Pr[A \mid B] = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

Conditioning restricts attention to outcomes in B , then rescales their probabilities so that they sum to one.

2.2 Random variables and independence

Definition 2.4 (Random variable). A real-valued *random variable* is a function $X : \Omega \rightarrow \mathbb{R}$.

Despite the name, the function is fixed. The randomness comes from the outcome to which we apply it. For the coin experiment, define

$$X_i(\omega) = \begin{cases} 1, & \text{if toss } i \text{ is heads,} \\ 0, & \text{otherwise.} \end{cases}$$

This is an *indicator random variable*: it records whether an event occurs. The total number of heads is $X = \sum_{i=1}^n X_i$.

Definition 2.5 (Bernoulli random variable). A random variable Y is *Bernoulli with parameter* $p \in [0, 1]$ if $\Pr[Y = 1] = p$ and $\Pr[Y = 0] = 1 - p$.

An event indicator is Bernoulli with parameter equal to the event's probability. For a fair-coin indicator, the parameter is $p = 1/2$.

Definition 2.6 (Independence). Two random variables X, Y are *independent* if, for all a, b ,

$$\Pr[X = a, Y = b] = \Pr[X = a] \Pr[Y = b].$$

Equivalently, whenever $\Pr[Y = b] > 0$, $\Pr[X = a \mid Y = b] = \Pr[X = a]$. A collection $X_1, \dots, X_n$ is *mutually independent* if

$$\Pr[X_1 = a_1, \dots, X_n = a_n] = \prod_{i=1}^n \Pr[X_i = a_i]$$

for every choice of values $a_1, \dots, a_n$.

The indicators of independent coin tosses are mutually independent. Throughout, an independent collection means a mutually independent one.

2.3 Expectation and linearity

Definition 2.7 (Expectation). The *expectation*, or mean, of a random variable X is

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} p(\omega) X(\omega).$$

For a fair-coin indicator, $\mathbb{E}[X_i] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 = \frac{1}{2}$. More generally, an indicator that is 1 with probability p has expectation p .

Theorem 2.1 (Linearity of expectation). For any random variables $X_1, \dots, X_n$,

$$\mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i].$$

Independence is not required.

Proof. Interchange the two finite sums:

$$\sum_{\omega \in \Omega} p(\omega) \sum_{i=1}^n X_i(\omega) = \sum_{i=1}^n \sum_{\omega \in \Omega} p(\omega) X_i(\omega) = \sum_{i=1}^n \mathbb{E}[X_i].$$

■

In particular, the expected number of heads in n fair tosses is

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = \frac{n}{2}.$$

Products behave differently: there is no general formula for $\mathbb{E}[XY]$ in terms of $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ alone. Independence is one thing makes such a formula (sometimes) possible.

Lemma 2.1 (Expectation of a product of independent variables). If $X_1, \dots, X_n$ are mutually independent random variables, then

$$\mathbb{E}[X_1 X_2 \cdots X_n] = \prod_{i=1}^n \mathbb{E}[X_i].$$

Proof. For two independent variables X, Y ,

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{a,b} \Pr[X = a, Y = b] ab = \sum_{a,b} \Pr[X = a] \Pr[Y = b] ab \\ &= \left(\sum_a \Pr[X = a] a \right) \left(\sum_b \Pr[Y = b] b \right) = \mathbb{E}[X] \mathbb{E}[Y], \end{aligned}$$

using independence to factor the joint probability into a product over a and b separately. The general case follows by induction on n : group $X_1, \dots, X_{n-1}$ and X_n as the two factors, using that their product is independent of X_n . ■

Unlike linearity of expectation, this identity genuinely needs independence: without it, $\mathbb{E}[XY]$ need not equal $\mathbb{E}[X] \mathbb{E}[Y]$. We will use this multiplicativity later, when computing exponential moments in the proof of the Chernoff bound.

2.4 The union bound

Lemma 2.2 (Union bound). For any events $A_1, \dots, A_m$,

$$\Pr \left[\bigcup_{i=1}^m A_i \right] \leq \sum_{i=1}^m \Pr[A_i].$$

Proof. For every outcome, the indicator of the union is at most the sum of the individual indicators. Take expectations and use linearity. ■

For two events this says $\Pr[A_1 \cup A_2] \leq \Pr[A_1] + \Pr[A_2]$. Again, there is no independence assumption. This is the same tool used to bound the chance that *any* server is overloaded in Lecture 1.

3 Concentration Inequalities

Knowing an expectation does not by itself show that a typical outcome is close to it. For example, a variable that equals 0 or n with equal probability has mean $n/2$, but never takes a value close to that mean. To justify using a sample average as an estimate, we need to bound the probability of a large deviation.

Definition 3.1 (Concentration inequality). A *concentration inequality* bounds the probability that a random variable deviates from its expectation. A typical form is

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq \text{a bound depending on } t \text{ and properties of } X,$$

where $t > 0$. One-sided versions bound the upper or lower tail separately.

A small bound means that X lies near its mean with high probability. For coin tosses, this quantifies how unlikely an unusually large fraction of heads is. For polling and our sampling algorithm, it turns an error tolerance into a bound on the chance of an inaccurate answer. Different assumptions lead to different guarantees: a mean alone gives a basic tail bound, variance gives more information, and independent bounded summands allow much stronger concentration. We begin with Markov's inequality, which is also the main tool used to prove the stronger bounds.

3.1 Markov's Inequality

Can we use the expectation to bound the probability of seeing many heads?

Theorem 3.1 (Markov's inequality). If $X \geq 0$ and $a > 0$, then

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.$$

In particular, if $\mathbb{E}[X] > 0$ and $t > 0$, then

$$\Pr[X \geq t\mathbb{E}[X]] \leq \frac{1}{t}.$$

Proof. Pointwise, $X \geq a \mathbf{1}_{\{X \geq a\}}$. Taking expectations gives $\mathbb{E}[X] \geq a \Pr[X \geq a]$, and division by a proves the claim. Nonnegativity ensures that outcomes below the threshold cannot contribute negative values to the expectation. ■

3.1.1 Applying Markov to coin tosses

Write our threshold as $(1 + \delta)n/2$. To obtain $0.51n$, take $\delta = 0.02$. Markov gives

$$\Pr\left[X \geq (1 + \delta)\frac{n}{2}\right] \leq \frac{1}{1 + \delta}, \quad \Pr[X \geq 0.51n] \leq \frac{1}{1.02} \approx 0.9804.$$

This only says that we cannot see at least 51% heads all the time. The bound does not improve as n grows.

Remark 3.1 (Markov's inequality is tight). For any $k \geq 1$, let $Y = 1$ with probability $1/k$, and $Y = 0$ otherwise. Then $\mathbb{E}[Y] = 1/k$ and

$$\Pr[Y \geq k\mathbb{E}[Y]] = \Pr[Y \geq 1] = \frac{1}{k}.$$

So a stronger general bound cannot follow from nonnegativity and the mean alone. To improve our coin-toss bound, we need to use more structure.

3.2 Variance

The guiding principle is that sums of independent, bounded random variables concentrate around their means. More broadly, we expect concentration when a function of independent inputs is not too sensitive to any single input. Changing one coin toss changes the number of heads by at most one.

Definition 3.2 (Variance). The *variance* of X is its average squared deviation from its mean:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Its *standard deviation* is $\sqrt{\text{Var}(X)}$.

For a Bernoulli variable X_i with success probability p , $\mathbb{E}[X_i] = p$, so

$$\begin{aligned}\text{Var}(X_i) &= p(1-p)^2 + (1-p)(0-p)^2 \\ &= p - 2p^2 + p^3 + p^2 - p^3 = p(1-p).\end{aligned}$$

Lemma 3.1 (Variance of a sum of independent random variables). If $X_1, \dots, X_n$ are independent and $X = \sum_i X_i$, then

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(X_i).$$

Proof. Let $\mu_i = \mathbb{E}[X_i]$. By linearity, $X - \mathbb{E}[X] = \sum_i (X_i - \mu_i)$. Expanding the square gives

$$\begin{aligned}\text{Var}(X) &= \mathbb{E} \left[\left(\sum_i (X_i - \mu_i) \right)^2 \right] \\ &= \sum_i \mathbb{E}[(X_i - \mu_i)^2] + \sum_{i \neq j} \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)].\end{aligned}$$

For $i \neq j$, independence makes the cross term equal to

$$\mathbb{E}[X_i - \mu_i] \mathbb{E}[X_j - \mu_j] = 0.$$

Thus only the individual variances remain. ■

Unlike linearity of expectation, this conclusion does need an assumption on the dependence between variables. Pairwise independence already suffices for this proof (we'll revisit this concept in a later lecture).

For n independent Bernoulli variables with the same parameter p ,

$$\text{Var}(X) = np(1-p).$$

For fair coins, this becomes $\text{Var}(X) = n/4$, so the standard deviation is $\sqrt{n}/2$.

3.3 Chebyshev's Inequality

Theorem 3.2 (Chebyshev's inequality). For any random variable X with finite variance and any $a > 0$,

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

If $\text{Var}(X) > 0$, then for $t > 0$ this is equivalently

$$\Pr[|X - \mathbb{E}[X]| \geq t\sqrt{\text{Var}(X)}] \leq \frac{1}{t^2}.$$

Proof. Apply Markov's inequality to the nonnegative variable $(X - \mathbb{E}[X])^2$:

$$\begin{aligned}\Pr[|X - \mathbb{E}[X]| \geq a] &= \Pr[(X - \mathbb{E}[X])^2 \geq a^2] \\ &\leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{a^2} = \frac{\text{Var}(X)}{a^2}.\end{aligned}$$
■

This is a simple example of the *moment method*: transform a random variable before applying Markov. Here we square the absolute deviation. More generally, a nonnegative increasing function of the absolute deviation can give a different bound. Later we will use an exponential transformation.

3.3.1 Returning to coin tosses

For $\delta > 0$, the event $X \geq (1 + \delta)n/2$ implies $|X - n/2| \geq \delta n/2$. Since the variance is $n/4$,

$$\Pr\left[X \geq (1 + \delta)\frac{n}{2}\right] \leq \frac{n/4}{(\delta n/2)^2} = \frac{1}{\delta^2 n}.$$

With $\delta = 0.02$, this gives $\Pr[X \geq 0.51n] \leq 2500/n$. For large n , this is much better than Markov's bound: it decreases inversely with n . As always, a probability bound larger than one can be replaced by one.

3.4 Chernoff–Hoeffding Bounds

For sums of independent random variables, the moment method gives a much stronger conclusion: an exponentially small probability of a large deviation.

Theorem 3.3 (Multiplicative Chernoff bounds). Let $X = \sum_{i=1}^n X_i$, where the X_i are independent Bernoulli random variables with $\Pr[X_i = 1] = p_i$. Write $\mu = \mathbb{E}[X] = \sum_i p_i$ and assume $\mu > 0$. Then

$$\Pr[X \geq (1 + \delta)\mu] \leq \exp\left(-\frac{\delta^2 \mu}{2 + \delta}\right) \quad (\delta > 0),$$

and

$$\Pr[X \leq (1 - \delta)\mu] \leq \exp\left(-\frac{\mu \delta^2}{2}\right) \quad (0 < \delta < 1).$$

Consequently, for $0 < \delta \leq 1$,

$$\Pr[|X - \mu| \geq \delta \mu] \leq 2 \exp\left(-\frac{\mu \delta^2}{3}\right).$$

The two-sided bound follows by the union bound and $2 + \delta \leq 3$. The lower-tail bound extends to $\delta = 1$ by taking a limit. The upper-tail proof and the connection between the additive and multiplicative forms appear in Section 5.1 and will be covered in precept.

3.4.1 Coin tosses: an exponential improvement

For fair coins, $\mu = n/2$. For $0 < \delta \leq 1$, the upper-tail bound gives

$$\Pr\left[X \geq (1 + \delta)\frac{n}{2}\right] \leq e^{-\delta^2 n/6}.$$

Taking $\delta = 2/100$ yields $\Pr[X \geq 0.51n] \leq e^{-n/15000}$. For a fixed positive $\delta \leq 1$, compare our three bounds:

Tool	Upper bound	Dependence on n
Markov	$1/(1 + \delta)$	constant
Chebyshev	$1/(\delta^2 n)$	inverse polynomial
Chernoff	$e^{-\delta^2 n/6}$	exponentially decreasing

The exponential bound is eventually far smaller than either earlier bound.

3.4.2 An additive form

Theorem 3.4 (Hoeffding's inequality). Let $X_1, \dots, X_n$ be independent random variables taking values in $[0, 1]$. Set $X = \sum_i X_i$ and $\mu = \mathbb{E}[X]$. For every $t > 0$,

$$\Pr[|X - \mu| \geq t] \leq 2 \exp\left(-\frac{2t^2}{n}\right).$$

We state this additive form without proof. The variables need not be indicators: any independent variables bounded between zero and one satisfy the theorem. Unlike the multiplicative form, the threshold is an absolute deviation t , rather than a fraction of μ .

4 Applications

4.1 Polling

Suppose a fraction p of the population supports R and the remaining fraction $1 - p$ supports D . Take n independent, uniformly random samples from the population (sampling with replacement), and let

$$X_i = \begin{cases} 1, & \text{if sample } i \text{ supports } R, \\ 0, & \text{otherwise.} \end{cases}$$

Then $\mathbb{E}[X_i] = p$ and $\text{Var}(X_i) = p(1 - p)$. If $X = \sum_i X_i$ is the number of sampled R supporters, then $\mathbb{E}[X] = np$. Our estimate of the population fraction is $\hat{p} = X/n$. Here n is the sample size, denoted s in the opening question.

An estimation error of at least ε occurs exactly when $|\hat{p} - p| \geq \varepsilon$, or equivalently when $|X - np| \geq \varepsilon n$. This expresses the failure event as a deviation of a sum of independent indicators.

For an additive error tolerance $\varepsilon > 0$, Hoeffding gives

$$\begin{aligned} \Pr[|\hat{p} - p| \geq \varepsilon] &= \Pr[|X - np| \geq \varepsilon n] \\ &\leq 2 \exp\left(-\frac{2\varepsilon^2 n^2}{n}\right) = 2e^{-2\varepsilon^2 n}. \end{aligned}$$

Thus, to make the failure probability at most $\alpha \in (0, 1)$, it suffices to take

$$n \geq \frac{\ln(2/\alpha)}{2\varepsilon^2}.$$

4.1.1 Two percentage points with 95% confidence

Take $\varepsilon = 0.02$ and $\alpha = 0.05$. We require

$$2e^{-2(0.02)^2 n} \leq 0.05, \quad \text{or equivalently} \quad n \geq \frac{\ln 40}{0.0008} \approx 4611.10.$$

So 4612 samples suffice, and a convenient round number is 5000. Indeed, for $n = 5000$ the failure bound is $2e^{-4} \approx 0.0366 < 0.05$.

The opening question asked for *one* percentage point of error. With $\varepsilon = 0.01$ and the same confidence, the analogous requirement is

$$n \geq \frac{\ln 40}{0.0002} \approx 18444.40,$$

so 18445 samples suffice. Halving the tolerated error multiplies this sample-size bound by four.

Remark 4.1 (Population size does not appear). The sufficient sample size depends on the error tolerance and failure probability, but not on the population size. This answers the opening question: the same sample-size bound works for both India and the United States, despite their very different population sizes. This conclusion concerns the idealized model of independent uniform samples; the sampling assumptions are what make each X_i an independent Bernoulli variable with parameter p .

4.2 Estimating an Array Average by Sampling

Fix an input array $a_1, \dots, a_N \in \{0, 1\}$, and write

$$p = \frac{1}{N} \sum_{j=1}^N a_j$$

for the fraction of ones. The array and p are fixed; only the algorithm's sampled indices are random. Let $J_1, \dots, J_s$ be independent uniform indices in $\{1, \dots, N\}$. Define

$$X_i = a_{J_i}, \quad X = \sum_{i=1}^s X_i, \quad \hat{p} = \frac{X}{s}.$$

Here J_i describes a random choice, X_i describes the observed value, and $\hat{p}$ is the algorithm's output. Each X_i is an indicator, with

$$\Pr[X_i = 1] = \frac{|\{j : a_j = 1\}|}{N} = p, \quad \mathbb{E}[X_i] = p.$$

The X_i are independent because each is a fixed function of its own independently chosen index. Repeated indices are allowed; sampling with replacement preserves independence between draws. By linearity,

$$\mathbb{E}[X] = sp, \quad \mathbb{E}[\hat{p}] = \frac{1}{s} \mathbb{E}[X] = p.$$

Thus the estimate is correct in expectation. To bound its error on an individual execution, apply Hoeffding to X :

$$\begin{aligned} \Pr[|\hat{p} - p| \geq \varepsilon] &= \Pr[|X - sp| \geq s\varepsilon] \\ &\leq 2 \exp\left(-\frac{2(s\varepsilon)^2}{s}\right) = 2e^{-2s\varepsilon^2}. \end{aligned}$$

For $0 < \varepsilon \leq 1$ and $0 < \alpha < 1$, choosing

$$s = \left\lceil \frac{\ln(2/\alpha)}{2\varepsilon^2} \right\rceil$$

makes this probability at most α . This guarantee holds for every fixed input array, regardless of the fraction or arrangement of its ones.

Assume generating a uniform random index, accessing an array entry, and arithmetic each take constant time. The algorithm takes s samples, maintains their sum, and divides by s . Its running time is therefore

$$O(s) = O\left(\varepsilon^{-2} \ln \frac{2}{\alpha}\right),$$

independently of N . The array is assumed to be already available for random access; loading the entire input is not part of this bound. Halving the error tolerance multiplies the sufficient sample

size by roughly four, whereas the dependence on the inverse failure probability is only logarithmic. For a large array, allowing a small error and failure probability can save us from inspecting most of its entries.

The analysis illustrates a modeling pattern: define variables for the random observations, express the output through them, and translate an inaccurate answer into a deviation of their sum. The algorithm does not know p and cannot check its own error, but the analysis can still bound the probability of that error. Lecture 7 develops further uses of randomness in algorithms and methods for making their answers reliable.

5 Extra Proofs

5.1 Proof of the Chernoff Bounds

This proof was skipped in lecture and will be covered in precept.

We follow the proof in Precept 4, using t for an additive deviation and λ for the exponential parameter. Let $X = \sum_{i=1}^n X_i$, where X_i are independent Bernoulli variables with parameters p_i , and let $\mu = \mathbb{E}[X] = \sum_i p_i > 0$. We first prove

$$\Pr[X \geq \mu + t] \leq \exp\left(-\frac{t^2}{2\mu + t}\right) \quad (t > 0).$$

Here t is an absolute deviation from the mean, whereas the earlier statement uses a relative deviation δ . Setting $t = \delta\mu$ gives $\mu + t = (1 + \delta)\mu$ and

$$\frac{t^2}{2\mu + t} = \frac{\delta^2\mu^2}{2\mu + \delta\mu} = \frac{\delta^2\mu}{2 + \delta},$$

so this is exactly the upper-tail bound in Theorem 3.3. Conversely, any $t > 0$ corresponds to $\delta = t/\mu$.

5.1.1 The upper tail: the proof from Precept 4

Proof. The idea is the same as before, but we apply Markov's inequality to the non-negative random variable $e^{\lambda X}$, for some $\lambda > 0$ to be chosen later. We have:

$$\Pr[X \geq \mu + t] = \Pr[e^{\lambda X} \geq e^{\lambda(\mu+t)}] \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda(\mu+t)}}.$$

Now we need to compute $\mathbb{E}[e^{\lambda X}]$. Since $X = \sum_{i=1}^n X_i$ and the X_i are independent, we have:

$$\mathbb{E}[e^{\lambda X}] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^n X_i}\right] = \mathbb{E}\left[\prod_{i=1}^n e^{\lambda X_i}\right] = \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}],$$

where the last equality follows from the independence of the X_i . Now, since X_i is a Bernoulli random variable with parameter p_i , we have:

$$\mathbb{E}[e^{\lambda X_i}] = p_i e^{\lambda} + (1 - p_i) = 1 + p_i(e^{\lambda} - 1) \leq e^{p_i(e^{\lambda} - 1)},$$

where the last inequality follows from the fact that $1 + x \leq e^x$ for all x . Thus, we have:

$$\mathbb{E}[e^{\lambda X}] \leq \prod_{i=1}^n e^{p_i(e^{\lambda} - 1)} = e^{(e^{\lambda} - 1) \sum_{i=1}^n p_i} = e^{(e^{\lambda} - 1)\mu}.$$

Plugging this back into our bound gives:

$$\Pr[X \geq \mu + t] \leq \frac{e^{(e^\lambda - 1)\mu}}{e^{\lambda(\mu + t)}} = e^{(e^\lambda - 1)\mu - \lambda(\mu + t)}.$$

Now we need to choose λ to minimize the right-hand side. Some simple calculus shows that the optimal choice is $\lambda = \ln(1 + t/\mu)$. Plugging this in gives:

$$\Pr[X \geq \mu + t] \leq e^{t - (\mu + t) \ln(1 + t/\mu)}.$$

Finally, we can use the inequality $\ln(1 + x) \geq \frac{x}{1+x/2}$ for $x > 0$ to get:

$$\Pr[X \geq \mu + t] \leq e^{t - (\mu + t) \frac{t/\mu}{1 + t/(2\mu)}} = e^{-\frac{t^2}{2\mu + t}}.$$

■

For completeness, the optimizing parameter follows from

$$\frac{d}{d\lambda} [\mu(e^\lambda - 1) - \lambda(\mu + t)] = \mu e^\lambda - (\mu + t).$$

It vanishes at $\lambda = \ln(1 + t/\mu)$, and the second derivative $\mu e^\lambda > 0$ proves this is the minimum. The logarithmic inequality used above follows by setting $h(x) = \ln(1 + x) - 2x/(2 + x)$: $h(0) = 0$ and

$$h'(x) = \frac{x^2}{(1 + x)(2 + x)^2} \geq 0 \quad (x \geq 0).$$

5.1.2 Why the exponential?

The exponential turns a sum into a product, whose expectation factors by independence. The quantity $\mathbb{E}[e^{\lambda X}]$ is the *moment-generating function*: expanding the exponential gives

$$\mathbb{E}[e^{\lambda X}] = \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E}[X^k]}{k!}.$$

The k th moment is $\mathbb{E}[X^k]$. Note that Chebyshev's inequality uses the second moment, and Markov's inequality uses the first moment. The higher moments we use the tighter the bound. This explains the connection with applying Markov to powers of X : the exponential incorporates all these moments in a single expression.

5.1.3 A more general form

The proof uses independence to factor the exponential moment, and then an upper bound on each factor. Thus the method applies whenever we can control the exponential moments of the independent summands; it is not limited to Bernoulli variables.

In particular, suppose $0 \leq X_i \leq 1$, and write $p_i = \mathbb{E}[X_i]$. By convexity, for $0 \leq x \leq 1$, $e^{\lambda x} \leq (1 - x) + xe^\lambda$. Taking expectations gives

$$\mathbb{E}[e^{\lambda X_i}] \leq 1 + p_i(e^\lambda - 1) \leq e^{p_i(e^\lambda - 1)}.$$

This is the same bound used in the Bernoulli proof, so the upper-tail argument continues unchanged. A similar argument with a negative exponential parameter gives the lower-tail bound; we state it without carrying out that calculation.

Theorem 5.1 (Additive Chernoff bounds). Let $X_1, \dots, X_n$ be independent random variables with $0 \leq X_i \leq 1$. Set $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X] > 0$. For every $t > 0$,

$$\Pr[X \geq \mu + t] \leq \exp\left(-\frac{t^2}{2\mu + t}\right), \quad \Pr[X \leq \mu - t] \leq \exp\left(-\frac{t^2}{2\mu}\right).$$

Consequently, for $0 < t \leq \mu$,

$$\Pr[|X - \mu| \geq t] \leq 2 \exp\left(-\frac{t^2}{3\mu}\right).$$

The two-sided statement follows by the union bound: add the upper- and lower-tail bounds and use $2\mu + t \leq 3\mu$. This last simplification is why we restrict to $t \leq \mu$.

5.1.4 Recovering the multiplicative statements

Bernoulli variables are a special case of variables in $[0, 1]$. As above, substitute $t = \delta\mu$ to express additive deviations as fractions of the mean. For the lower tail,

$$\mu - t = (1 - \delta)\mu, \quad \frac{t^2}{2\mu} = \frac{\mu\delta^2}{2}.$$

For the two-sided bound,

$$|X - \mu| \geq t \iff |X - \mu| \geq \delta\mu, \quad \frac{t^2}{3\mu} = \frac{\mu\delta^2}{3}.$$

The restriction $0 < t \leq \mu$ becomes $0 < \delta \leq 1$. Together with the upper-tail substitution at the start of this section, these give all three statements in Theorem 3.3.