Randomized Algorithms

From expectation to tail bounds: Markov and Chebyshev

Alex Gittens
Department of Computer Science, Rensselaer Polytechnic Institute

Fall 2026

The question for today

Last lecture, expectation gave us a way to quantify *average* cost. Today we ask a different question:

Tail question

If a randomized algorithm has small expected cost, how much confidence should we have that a particular run is not much worse?

The question for today

Last lecture, expectation gave us a way to quantify *average* cost. Today we ask a different question:

Tail question

If a randomized algorithm has small expected cost, how much confidence should we have that a particular run is not much worse?

We will build two increasingly informative answers:

mean $\Longrightarrow$ Markov and mean + variance $\Longrightarrow$ Chebyshev.

The question for today

Last lecture, expectation gave us a way to quantify *average* cost. Today we ask a different question:

Tail question

If a randomized algorithm has small expected cost, how much confidence should we have that a particular run is not much worse?

We will build two increasingly informative answers:

mean $\Longrightarrow$ Markov and mean + variance $\Longrightarrow$ Chebyshev.

Our goal is just to memorize these inequalities. We want to understand

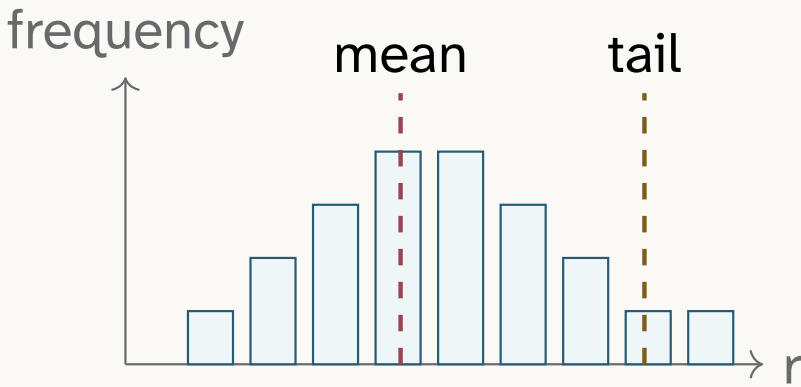
- what information each inequality uses,
- what sort of guarantee that information can support, and
- when the guarantee is too weak and we need more structure.

Expectation does not describe reliability

Two runtime distributions can have the same mean and behave very differently.

A small value of $\mathbb{E}[T]$ does *not* rule out rare, extremely expensive runs. What we need is control of a **tail probability**, such as

$$\Pr[T \geq t].$$



First principle

An expectation bound is an average-cost statement. A tail bound is a reliability statement.

PART 1

From expectation to a tail bound

1

Markov's inequality

Markov

If $X \geq 0$ and $a > 0$, then

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

Markov's inequality

Markov

If $X \geq 0$ and $a > 0$, then

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

The proof is almost obvious once we look at the bad event $\{X \geq a\}$. On that event, X is at least a; outside it, the indicator contributes nothing:

$$X = X\mathbf{1}_{\{X \geq a\}} + X\mathbf{1}_{\{X < a\}}$$

Markov's inequality

Markov

If $X \geq 0$ and $a > 0$, then

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

The proof is almost obvious once we look at the bad event $\{X \geq a\}$. On that event, X is at least a; outside it, the indicator contributes nothing:

$$\begin{aligned} X &= X\mathbf{1}_{\{X \geq a\}} + X\mathbf{1}_{\{X < a\}} \\ &\geq a\,\mathbf{1}_{\{X \geq a\}} + X\mathbf{1}_{\{X < a\}} \end{aligned}$$

Markov's inequality

Markov

If $X \geq 0$ and $a > 0$, then

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

The proof is almost obvious once we look at the bad event $\{X \geq a\}$. On that event, X is at least a; outside it, the indicator contributes nothing:

$$\begin{aligned} X &= X\mathbf{1}_{\{X \geq a\}} + X\mathbf{1}_{\{X < a\}} \\ &\geq a\,\mathbf{1}_{\{X \geq a\}} + X\mathbf{1}_{\{X < a\}} \\ &\geq a\,\mathbf{1}_{\{X \geq a\}}. \end{aligned}$$

Markov's inequality

Markov

If $X \geq 0$ and $a > 0$, then

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

The proof is almost obvious once we look at the bad event $\{X \geq a\}$. On that event, X is at least a; outside it, the indicator contributes nothing:

$$\begin{aligned} X &= X\mathbf{1}_{\{X \geq a\}} + X\mathbf{1}_{\{X < a\}} \\ &\geq a\,\mathbf{1}_{\{X \geq a\}} + X\mathbf{1}_{\{X < a\}} \\ &\geq a\,\mathbf{1}_{\{X \geq a\}}. \end{aligned}$$

Taking expectations,

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

Intuitively, *a large tail probability forces the mean to be large*.

What Markov can—and cannot—tell us

Suppose $\mathbb{E}[X] = \mu$ and $X \geq 0$. Setting $a = c\mu$ gives

$$\Pr[X \geq c\mu] \leq \frac{1}{c}.$$

What Markov can—and cannot—tell us

Suppose $\mathbb{E}[X] = \mu$ and $X \geq 0$. Setting $a = c\mu$ gives

$$\Pr[X \geq c\mu] \leq \frac{1}{c}.$$

So without knowing anything beyond the mean:

- exceeding 10μ has probability at most 0.1;

What Markov can—and cannot—tell us

Suppose $\mathbb{E}[X] = \mu$ and $X \geq 0$. Setting $a = c\mu$ gives

$$\Pr[X \geq c\mu] \leq \frac{1}{c}.$$

So without knowing anything beyond the mean:

- exceeding 10μ has probability at most 0.1;
- exceeding 100μ has probability at most 0.01.

What Markov can—and cannot—tell us

Suppose $\mathbb{E}[X] = \mu$ and $X \geq 0$. Setting $a = c\mu$ gives

$$\Pr[X \geq c\mu] \leq \frac{1}{c}.$$

So without knowing anything beyond the mean:

- exceeding 10μ has probability at most 0.1;
- exceeding 100μ has probability at most 0.01.

This $1/c$ dependence is the best possible guarantee from the mean alone.

What Markov can—and cannot—tell us

Suppose $\mathbb{E}[X] = \mu$ and $X \geq 0$. Setting $a = c\mu$ gives

$$\Pr[X \geq c\mu] \leq \frac{1}{c}.$$

So without knowing anything beyond the mean:

- exceeding 10μ has probability at most 0.1;
- exceeding 100μ has probability at most 0.01.

This $1/c$ dependence is the best possible guarantee from the mean alone.

Tight example

Let $X = c\mu$ with probability $1/c$, and $X = 0$ otherwise. Then $\mathbb{E}[X] = \mu$ and

$$\Pr[X \geq c\mu] = \frac{1}{c}.$$

A running example: randomized quickselect

Find the k-th smallest key without sorting the entire input

Choose a pivot uniformly at random, partition around it, and recurse only on the side that can contain rank k.

The algorithm can still take $\Theta(n^2)$ comparisons on an unlucky sequence of pivots. The real question is therefore:

Runtime question

Can we show that a run is unlikely to take much more than linear time?

Input: array A of distinct keys, target rank k
if $|A| = 1$ **then**
  return $A[1]$
choose $p \in A$ uniformly at random
partition A into L, $\{p\}$, and R
if $k \leq |L|$ **then**
  return Quickselect(L, k)
if $k = |L| + 1$ **then**
  return p
return Quickselect($R, k - |L| - 1$)

Quickselect: expected number of comparisons

Let T_n denote the number of comparisons. We use a simple worst-rank upper bound, so the argument applies to every target rank k.

The first partition uses $n-1$ comparisons. If the pivot has rank j, the recursive call has size at most

$$\max\{j-1, n-j\}.$$

Quickselect: expected number of comparisons

Let T_n denote the number of comparisons. We use a simple worst-rank upper bound, so the argument applies to every target rank k.

The first partition uses $n-1$ comparisons. If the pivot has rank j, the recursive call has size at most

$$\max\{j-1, n-j\}.$$

Because the pivot rank is uniform,

$$\mathbb{E}[T_n] \le n-1+\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[T_{\max\{j-1,n-j\}}\right].$$

Quickselect: expected number of comparisons

Let T_n denote the number of comparisons. We use a simple worst-rank upper bound, so the argument applies to every target rank k.

The first partition uses $n-1$ comparisons. If the pivot has rank j, the recursive call has size at most

$$\max\{j-1, n-j\}.$$

Because the pivot rank is uniform,

$$\mathbb{E}[T_n] \le n-1+\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[T_{\max\{j-1,n-j\}}\right].$$

Pair the symmetric ranks j and $n-j+1$. Then

$$\mathbb{E}[T_n] \le n-1+\frac{2}{n}\sum_{s=\lceil n/2\rceil}^{n-1}\mathbb{E}[T_s].$$

Randomization gives a simple distribution over recursive subproblem sizes.

Quickselect: solving the recurrence

We prove by induction that $\mathbb{E}[T_n] \leq 4n$. Assume $\mathbb{E}[T_s] \leq 4s$ for every $s < n$. Then

$$\mathbb{E}[T_n] \leq n - 1 + \frac{8}{n} \sum_{s=\lceil n/2 \rceil}^{n-1} s.$$

Quickselect: solving the recurrence

We prove by induction that $\mathbb{E}[T_n] \le 4n$. Assume $\mathbb{E}[T_s] \le 4s$ for every $s < n$. Then

$$\mathbb{E}[T_n] \le n - 1 + \frac{8}{n} \sum_{s=\lceil n/2 \rceil}^{n-1} s.$$

The upper half contains at most $n/2$ terms, whose average is at most $3n/4$, so

$$\sum_{s=\lceil n/2 \rceil}^{n-1} s \le \frac{3n^2}{8}.$$

Quickselect: solving the recurrence

We prove by induction that $\mathbb{E}[T_n] \le 4n$. Assume $\mathbb{E}[T_s] \le 4s$ for every $s < n$. Then

$$\mathbb{E}[T_n] \le n - 1 + \frac{8}{n} \sum_{s=\lceil n/2 \rceil}^{n-1} s.$$

The upper half contains at most $n/2$ terms, whose average is at most $3n/4$, so

$$\sum_{s=\lceil n/2 \rceil}^{n-1} s \le \frac{3n^2}{8}.$$

Hence

$$\mathbb{E}[T_n] \le n - 1 + 3n < 4n.$$

Quickselect: solving the recurrence

We prove by induction that $\mathbb{E}[T_n] \le 4n$. Assume $\mathbb{E}[T_s] \le 4s$ for every $s < n$. Then

$$\mathbb{E}[T_n] \le n - 1 + \frac{8}{n} \sum_{s=\lceil n/2 \rceil}^{n-1} s.$$

The upper half contains at most $n/2$ terms, whose average is at most $3n/4$, so

$$\sum_{s=\lceil n/2 \rceil}^{n-1} s \le \frac{3n^2}{8}.$$

Hence

$$\mathbb{E}[T_n] \le n - 1 + 3n < 4n.$$

Expected-time result

Randomized quickselect uses $O(n)$ comparisons in expectation.

What does the expectation bound buy us?

Now Markov turns the expected-time calculation into a statement about an individual run:

$$\Pr[T_n \geq 40n] \leq \frac{\mathbb{E}[T_n]}{40n}$$

What does the expectation bound buy us?

Now Markov turns the expected-time calculation into a statement about an individual run:

$$\begin{aligned}\Pr[T_n \geq 40n] &\leq \frac{\mathbb{E}[T_n]}{40n} \\ &\leq \frac{4n}{40n}\end{aligned}$$

What does the expectation bound buy us?

Now Markov turns the expected-time calculation into a statement about an individual run:

$$\begin{aligned}\Pr[T_n \geq 40n] &\leq \frac{\mathbb{E}[T_n]}{40n} \\ &\leq \frac{4n}{40n} \\ &= 0.1.\end{aligned}$$

What does the expectation bound buy us?

Now Markov turns the expected-time calculation into a statement about an individual run:

$$\begin{aligned}\Pr[T_n \geq 40n] &\leq \frac{\mathbb{E}[T_n]}{40n} \\ &\leq \frac{4n}{40n} \\ &= 0.1.\end{aligned}$$

This is a genuine high-probability statement:

> *with probability at least* 0.9*, quickselect uses fewer than* $40n$ *comparisons.*

What does the expectation bound buy us?

Now Markov turns the expected-time calculation into a statement about an individual run:

$$\begin{aligned}\Pr[T_n \geq 40n] &\leq \frac{\mathbb{E}[T_n]}{40n} \\ &\leq \frac{4n}{40n} \\ &= 0.1.\end{aligned}$$

This is a genuine high-probability statement:

> *with probability at least* 0.9*, quickselect uses fewer than* $40n$ *comparisons.*

But notice the price. To reduce the failure probability from 0.1 to 0.01, Markov makes us increase the threshold by another factor of 10.

What is missing?

The mean does not tell us how tightly T_n is clustered around that mean.

PART 2

Variance and Chebyshev

2

Variance measures fluctuation around the mean

For $\mu = \mathbb{E}[X]$, define

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

Variance measures fluctuation around the mean

For $\mu = \mathbb{E}[X]$, define

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

The square is doing two jobs:

- positive and negative deviations cannot cancel;
- large deviations are penalized more heavily.

Variance measures fluctuation around the mean

For $\mu = \mathbb{E}[X]$, define

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

The square is doing two jobs:

- positive and negative deviations cannot cancel;
- large deviations are penalized more heavily.

Expanding the square gives the computationally useful identity

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

Variance measures fluctuation around the mean

For $\mu = \mathbb{E}[X]$, define

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

The square is doing two jobs:

- positive and negative deviations cannot cancel;
- large deviations are penalized more heavily.

Expanding the square gives the computationally useful identity

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

Example: Bernoulli

If $B \sim \mathrm{Bernoulli}(p)$, then $B^2 = B$, so

$$\mathrm{Var}(B) = p - p^2 = p(1-p).$$

Covariance determines the variance of a sum

Randomized algorithms often accumulate random contributions—comparisons, collisions, loads, or measurements. Controlling the total requires the variance of a sum.

In general, $\mathrm{Var}(X+Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$.

A simple picture

Expanding gives

$$\mathrm{Var}(X+Y) = \mathbb{E}[((X-\mathbb{E}X)+(Y-\mathbb{E}Y))^2]$$

Covariance determines the variance of a sum

Randomized algorithms often accumulate random contributions—comparisons, collisions, loads, or measurements. Controlling the total requires the variance of a sum.

In general, $\mathrm{Var}(X+Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$.

A simple picture

Expanding gives

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[((X-\mathbb{E}X)+(Y-\mathbb{E}Y))^2] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y)\end{aligned}$$

Covariance determines the variance of a sum

Randomized algorithms often accumulate random contributions—comparisons, collisions, loads, or measurements. Controlling the total requires the variance of a sum.

In general, $\mathrm{Var}(X+Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$.

A simple picture

Expanding gives

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[((X-\mathbb{E}X)+(Y-\mathbb{E}Y))^2] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) \\ &\quad + 2\mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)].\end{aligned}$$

Covariance determines the variance of a sum

Randomized algorithms often accumulate random contributions—comparisons, collisions, loads, or measurements. Controlling the total requires the variance of a sum.

In general, $\mathrm{Var}(X+Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$. Expanding gives

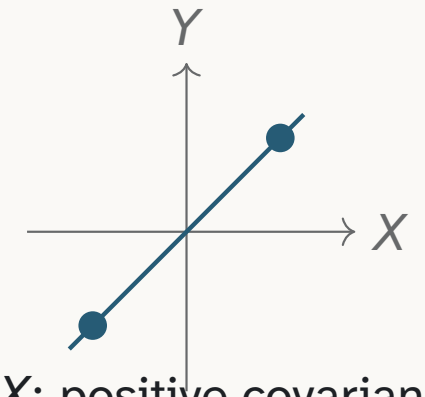
$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[((X-\mathbb{E}X)+(Y-\mathbb{E}Y))^2] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) \\ &\quad + 2\mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)].\end{aligned}$$

The last expectation measures whether the fluctuations move together. We call it

$$\mathrm{Cov}(X,Y) := \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)].$$

In particular, $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

A simple picture



$Y = X$: positive covariance
fluctuations reinforce

Covariance determines the variance of a sum

Randomized algorithms often accumulate random contributions—comparisons, collisions, loads, or measurements. Controlling the total requires the variance of a sum.

In general, $\mathrm{Var}(X+Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$.
Expanding gives

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[((X-\mathbb{E}X)+(Y-\mathbb{E}Y))^2] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) \\ &\quad + 2\mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)].\end{aligned}$$

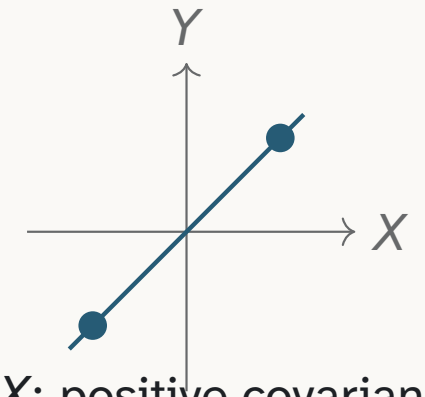
The last expectation measures whether the fluctuations move together. We call it

$$\mathrm{Cov}(X,Y) := \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)].$$

In particular, $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
Thus, for any $X_1, \ldots, X_n$,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_i \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

A simple picture



$Y = X$: positive covariance
fluctuations reinforce

Independent averaging reduces variance

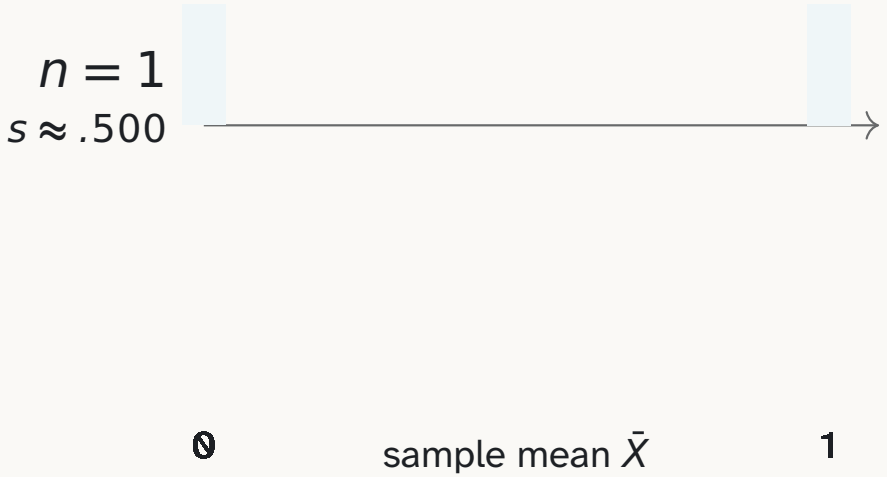
Repeated random trials help only if combining them improves reliability. The covariance formula shows why independence makes averaging work.

Let $X_1, \ldots, X_n$ be independent, each with mean μ and variance σ^2, and define the **sample mean**

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

10,000 simulated Bernoulli averages

$X_i \sim \text{Bernoulli}(1/2)$ independently



Independent averaging reduces variance

Repeated random trials help only if combining them improves reliability. The covariance formula shows why independence makes averaging work.

Let $X_1, \dots, X_n$ be independent, each with mean μ and variance σ^2, and define the **sample mean**

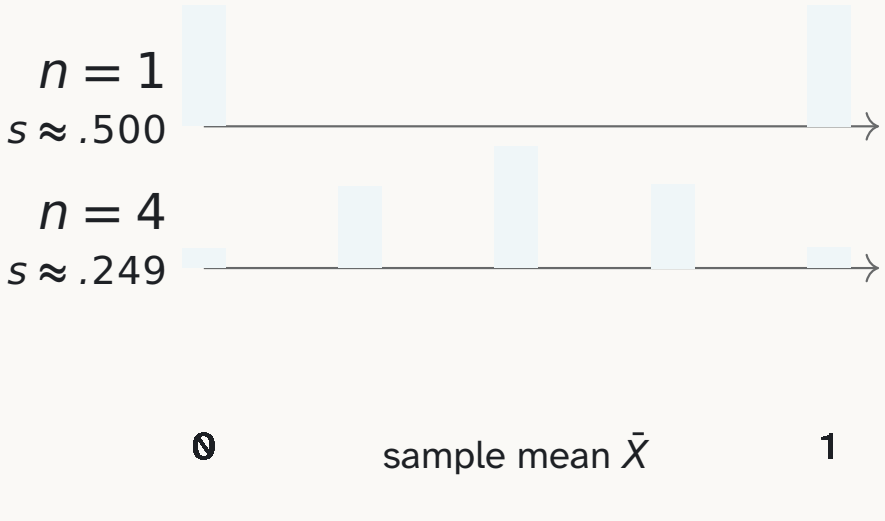
$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Because independence implies zero covariance,

$$\operatorname{Var}(\bar{X}) = \frac{\sigma^2}{n}, \qquad \operatorname{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

10,000 simulated Bernoulli averages

$X_i \sim \text{Bernoulli}(1/2)$ independently



Independent averaging reduces variance

Repeated random trials help only if combining them improves reliability. The covariance formula shows why independence makes averaging work.

Let $X_1, \dots, X_n$ be independent, each with mean μ and variance σ^2, and define the **sample mean**

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

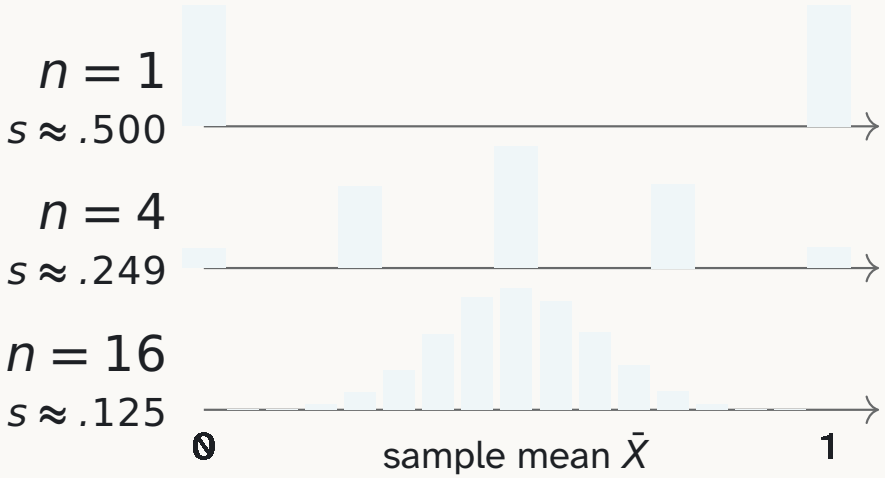
Because independence implies zero covariance,

$$\operatorname{Var}(\bar{X}) = \frac{\sigma^2}{n}, \qquad \operatorname{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

This is why we compute the variance now: Chebyshev will turn it into a high-probability guarantee.

10,000 simulated Bernoulli averages

$X_i \sim \text{Bernoulli}(1/2)$ independently



Chebyshev's inequality

Chebyshev

If X has mean μ and variance σ^2, then for every $t > 0$,

$$\Pr(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Chebyshev's inequality

Chebyshev

If X has mean μ and variance σ^2, then for every $t > 0$,

$$\Pr(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

The proof is Markov applied to an informative random variable:

$$Y = (X - \mu)^2 \geq 0.$$

Chebyshev's inequality

Chebyshev

If X has mean μ and variance σ^2, then for every $t > 0$,

$$\Pr(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

The proof is Markov applied to an informative random variable:

$$Y = (X - \mu)^2 \geq 0.$$

The events

$$|X - \mu| \geq t \qquad \text{and} \qquad Y \geq t^2$$

are identical. Therefore

$$\Pr(|X - \mu| \geq t) = \Pr(Y \geq t^2)$$

Chebyshev's inequality

Chebyshev

If X has mean μ and variance σ^2, then for every $t > 0$,

$$\Pr(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

The proof is Markov applied to an informative random variable:

$$Y = (X - \mu)^2 \geq 0.$$

The events

$$|X - \mu| \geq t \qquad \text{and} \qquad Y \geq t^2$$

are identical. Therefore

$$\Pr(|X - \mu| \geq t) = \Pr(Y \geq t^2) \leq \frac{\mathbb{E}[Y]}{t^2}$$

Chebyshev's inequality

Chebyshev

If X has mean μ and variance σ^2, then for every $t > 0$,

$$\Pr(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

The proof is Markov applied to an informative random variable:

$$Y = (X - \mu)^2 \geq 0.$$

The events

$$|X - \mu| \geq t \qquad \text{and} \qquad Y \geq t^2$$

are identical. Therefore

$$\Pr(|X - \mu| \geq t) = \Pr(Y \geq t^2) \leq \frac{\mathbb{E}[Y]}{t^2} = \frac{\sigma^2}{t^2}.$$

Read Chebyshev in units of standard deviation

Set $t = c\sigma$. Then

$$\Pr(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2}.$$

Read Chebyshev in units of standard deviation

Set $t = c\sigma$. Then

$$\Pr(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2}.$$

So, using only the mean and variance,

- the probability of being 2 or more standard deviations away is at most $1/4$;

Read Chebyshev in units of standard deviation

Set $t = c\sigma$. Then

$$\Pr(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2}.$$

So, using only the mean and variance,

- the probability of being 2 or more standard deviations away is at most $1/4$;
- 5 or more standard deviations away: at most $1/25$;

Read Chebyshev in units of standard deviation

Set $t = c\sigma$. Then

$$\Pr(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2}.$$

So, using only the mean and variance,

- the probability of being 2 or more standard deviations away is at most $1/4$;
- 5 or more standard deviations away: at most $1/25$;
- 10 or more standard deviations away: at most $1/100$.

Read Chebyshev in units of standard deviation

Set $t = c\sigma$. Then

$$\Pr(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2}.$$

So, using only the mean and variance,

- the probability of being 2 or more standard deviations away is at most $1/4$;
- 5 or more standard deviations away: at most $1/25$;
- 10 or more standard deviations away: at most $1/100$.

What changed from Markov?

Markov measures deviations against the mean. Chebyshev measures deviations against the natural fluctuation scale σ.

PART 3

Applications and limitations

3

Sample means: a first concentration bound

For independent $X_1, \ldots, X_n$ with common mean μ and variance σ^2, define the **sample mean** $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Sample means: a first concentration bound

For independent $X_1, \ldots, X_n$ with common mean μ and variance σ^2, define the **sample mean** $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Applying Chebyshev,

$$\Pr(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Sample means: a first concentration bound

For independent $X_1, \ldots, X_n$ with common mean μ and variance σ^2, define the **sample mean** $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\operatorname{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Applying Chebyshev,

$$\Pr(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Equivalently, to make the failure probability at most δ, it is enough that

$$n \geq \frac{\sigma^2}{\delta\varepsilon^2}.$$

Sample means: a first concentration bound

For independent $X_1, \ldots, X_n$ with common mean μ and variance σ^2, define the **sample mean** $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Applying Chebyshev,

$$\Pr(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Equivalently, to make the failure probability at most δ, it is enough that

$$n \geq \frac{\sigma^2}{\delta\varepsilon^2}.$$

Consequence: sample means are consistent

Chebyshev already proves consistency of the sample mean: for fixed ε, the probability of an ε-sized error goes to zero as n grows.

But the confidence dependence is expensive

Chebyshev gives

$$\Pr(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

But the confidence dependence is expensive

Chebyshev gives

$$\Pr(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

To reduce the failure probability by a factor of 100, we must increase the sample size by a factor of 100.

But the confidence dependence is expensive

Chebyshev gives

$$\Pr(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

To reduce the failure probability by a factor of 100, we must increase the sample size by a factor of 100.

For many randomized algorithms and statistical estimators, we would prefer a guarantee of the form

$$\Pr(\text{failure}) \leq e^{-cn},$$

where confidence improves *exponentially* with the amount of independent randomness.

But the confidence dependence is expensive

Chebyshev gives

$$\Pr(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

To reduce the failure probability by a factor of 100, we must increase the sample size by a factor of 100.

For many randomized algorithms and statistical estimators, we would prefer a guarantee of the form

$$\Pr(\text{failure}) \leq e^{-cn},$$

where confidence improves *exponentially* with the amount of independent randomness.

Moment bounds give only polynomial tail guarantees

Mean and variance alone generally support only polynomial tail guarantees. Exponential tails require additional information about the distribution.

Balls into bins: concentration for one server

Throw n tasks independently and uniformly onto n servers. For a fixed server i, let

$$L_i = \sum_{j=1}^{n} I_{ij}, \qquad I_{ij} = \mathbf{1}_{\{\text{task } j \text{ goes to server } i\}}.$$

Then $I_{ij} \sim \mathrm{Bernoulli}(1/n)$ independently across j, so

$$\mathbb{E}[L_i] = 1, \qquad \mathrm{Var}(L_i) = 1 - \frac{1}{n} < 1.$$

Balls into bins: concentration for one server

Throw n tasks independently and uniformly onto n servers. For a fixed server i, let

$$L_i = \sum_{j=1}^{n} I_{ij}, \qquad I_{ij} = \mathbf{1}_{\{\text{task } j \text{ goes to server } i\}}.$$

Then $I_{ij} \sim \mathrm{Bernoulli}(1/n)$ independently across j, so

$$\mathbb{E}[L_i] = 1, \qquad \mathrm{Var}(L_i) = 1 - \frac{1}{n} < 1.$$

Chebyshev therefore gives, for $t > 0$,

$$\Pr(L_i \geq 1 + t) \leq \Pr(|L_i - 1| \geq t) \leq \frac{1}{t^2}.$$

Balls into bins: concentration for one server

Throw n tasks independently and uniformly onto n servers. For a fixed server i, let

$$L_i = \sum_{j=1}^{n} I_{ij}, \qquad I_{ij} = \mathbf{1}_{\{\text{task } j \text{ goes to server } i\}}.$$

Then $I_{ij} \sim \mathrm{Bernoulli}(1/n)$ independently across j, so

$$\mathbb{E}[L_i] = 1, \qquad \mathrm{Var}(L_i) = 1 - \frac{1}{n} < 1.$$

Chebyshev therefore gives, for $t > 0$,

$$\Pr(L_i \geq 1 + t) \leq \Pr(|L_i - 1| \geq t) \leq \frac{1}{t^2}.$$

Local statement

A fixed server is unlikely to be very heavily loaded.

Balls into bins: from one server to the maximum

We want a statement about

$$L_{\max} = \max_{1 \le i \le n} L_i.$$

Balls into bins: from one server to the maximum

We want a statement about

$$L_{\max} = \max_{1 \le i \le n} L_i.$$

Using the union bound,

$$\Pr(L_{\max} \ge 1 + t) \le \sum_{i=1}^{n} \Pr(L_i \ge 1 + t) \le \frac{n}{t^2}.$$

Balls into bins: from one server to the maximum

We want a statement about

$$L_{\max} = \max_{1 \le i \le n} L_i.$$

Using the union bound,

$$\Pr(L_{\max} \ge 1 + t) \le \sum_{i=1}^{n} \Pr(L_i \ge 1 + t) \le \frac{n}{t^2}.$$

Set $t = c\sqrt{n}$:

$$\Pr(L_{\max} \ge 1 + c\sqrt{n}) \le \frac{1}{c^2}.$$

Balls into bins: from one server to the maximum

We want a statement about

$$L_{\max} = \max_{1 \le i \le n} L_i.$$

Using the union bound,

$$\Pr(L_{\max} \ge 1 + t) \le \sum_{i=1}^{n} \Pr(L_i \ge 1 + t) \le \frac{n}{t^2}.$$

Set $t = c\sqrt{n}$:

$$\Pr(L_{\max} \ge 1 + c\sqrt{n}) \le \frac{1}{c^2}.$$

This is valid, but extremely loose. In fact, with high probability the true maximum load is only

$$\Theta\left(\frac{\log n}{\log \log n}\right).$$

Balls into bins: from one server to the maximum

We want a statement about

$$L_{\max} = \max_{1 \le i \le n} L_i.$$

Using the union bound,

$$\Pr(L_{\max} \ge 1 + t) \le \sum_{i=1}^{n} \Pr(L_i \ge 1 + t) \le \frac{n}{t^2}.$$

Set $t = c\sqrt{n}$:

$$\Pr(L_{\max} \ge 1 + c\sqrt{n}) \le \frac{1}{c^2}.$$

This is valid, but extremely loose. In fact, with high probability the true maximum load is only

$$\Theta\left(\frac{\log n}{\log \log n}\right).$$

The source of this gap: Chebyshev uses only the variance of each load, and the union bound then pays for all n servers. Much of the structure of the problem has been ignored.

Hash collisions: set up the counting variable

Hash n distinct keys independently and uniformly into m buckets. Let

$$X_{ij} = \mathbf{1}_{\{h(i)=h(j)\}}, \qquad C = \sum_{i<j} X_{ij},$$

so C counts colliding pairs.

Hash collisions: set up the counting variable

Hash n distinct keys independently and uniformly into m buckets. Let

$$X_{ij} = \mathbf{1}_{\{h(i)=h(j)\}}, \qquad C = \sum_{i<j} X_{ij},$$

so C counts colliding pairs.
For each pair,

$$\mathbb{E}[X_{ij}] = \Pr(h(i) = h(j)) = \frac{1}{m}.$$

By linearity of expectation,

$$\mathbb{E}[C] = \binom{n}{2}\frac{1}{m}.$$

Hash collisions: set up the counting variable

Hash n distinct keys independently and uniformly into m buckets. Let

$$X_{ij} = \mathbf{1}_{\{h(i)=h(j)\}}, \qquad C = \sum_{i<j} X_{ij},$$

so C counts colliding pairs.
For each pair,

$$\mathbb{E}[X_{ij}] = \Pr(h(i) = h(j)) = \frac{1}{m}.$$

By linearity of expectation,

$$\mathbb{E}[C] = \binom{n}{2}\frac{1}{m}.$$

The expected number of colliding pairs becomes large precisely when

$$n \asymp \sqrt{m}.$$

Hash collisions: the covariance calculation

To use Chebyshev, we need $\mathrm{Var}(C)$. The indicators X_{ij} are not mutually independent, but every *pair* of distinct collision indicators is independent.

Hash collisions: the covariance calculation

To use Chebyshev, we need $\mathrm{Var}(C)$. The indicators X_{ij} are not mutually independent, but every *pair* of distinct collision indicators is independent.

For example, even when two indicators share a key,

$$\Pr(X_{12} = 1,\, X_{13} = 1) = \Pr(h(1) = h(2) = h(3)) = \frac{1}{m^2},$$

while

$$\Pr(X_{12} = 1)\Pr(X_{13} = 1) = \frac{1}{m} \cdot \frac{1}{m} = \frac{1}{m^2}.$$

Hash collisions: the covariance calculation

To use Chebyshev, we need $\mathrm{Var}(C)$. The indicators X_{ij} are not mutually independent, but every *pair* of distinct collision indicators is independent.

For example, even when two indicators share a key,

$$\Pr(X_{12} = 1,\, X_{13} = 1) = \Pr(h(1) = h(2) = h(3)) = \frac{1}{m^2},$$

while

$$\Pr(X_{12} = 1)\Pr(X_{13} = 1) = \frac{1}{m} \cdot \frac{1}{m} = \frac{1}{m^2}.$$

Thus every two distinct collision indicators have zero covariance:

$$\mathrm{Cov}(X_{ij}, X_{k\ell}) = 0 \qquad ((i,j) \neq (k,\ell)).$$

Hash collisions: the covariance calculation

To use Chebyshev, we need $\mathrm{Var}(C)$. The indicators X_{ij} are not mutually independent, but every *pair* of distinct collision indicators is independent.

For example, even when two indicators share a key,

$$\Pr(X_{12} = 1,\, X_{13} = 1) = \Pr(h(1) = h(2) = h(3)) = \frac{1}{m^2},$$

while

$$\Pr(X_{12} = 1)\Pr(X_{13} = 1) = \frac{1}{m} \cdot \frac{1}{m} = \frac{1}{m^2}.$$

Thus every two distinct collision indicators have zero covariance:

$$\mathrm{Cov}(X_{ij}, X_{k\ell}) = 0 \qquad ((i,j) \neq (k,\ell)).$$

Therefore

$$\mathrm{Var}(C) = \sum_{i<j} \mathrm{Var}(X_{ij}) \leq \sum_{i<j} \mathbb{E}[X_{ij}] = \mathbb{E}[C].$$

Hash collisions: the birthday threshold from Cheby

The event of no collision is $C = 0$. Since $\mathbb{E}[C] > 0$,

$$\{C = 0\} \subseteq \{|C - \mathbb{E}[C]| \geq \mathbb{E}[C]\}.$$

Hash collisions: the birthday threshold from Cheby

The event of no collision is $C = 0$. Since $\mathbb{E}[C] > 0$,

$$\{C = 0\} \subseteq \{|C - \mathbb{E}[C]| \geq \mathbb{E}[C]\}.$$

Chebyshev gives

$$\Pr(C = 0) \leq \frac{\mathrm{Var}(C)}{\mathbb{E}[C]^2} \leq \frac{1}{\mathbb{E}[C]} = \frac{m}{\binom{n}{2}}.$$

Hash collisions: the birthday threshold from Cheby

The event of no collision is $C = 0$. Since $\mathbb{E}[C] > 0$,

$$\{C = 0\} \subseteq \{|C - \mathbb{E}[C]| \geq \mathbb{E}[C]\}.$$

Chebyshev gives

$$\Pr(C = 0) \leq \frac{\mathrm{Var}(C)}{\mathbb{E}[C]^2} \leq \frac{1}{\mathbb{E}[C]} = \frac{m}{\binom{n}{2}}.$$

Hence, if $n/\sqrt{m} \to \infty$, then

$$\Pr(\text{no collision}) \to 0.$$

Hash collisions: the birthday threshold from Cheby

The event of no collision is $C = 0$. Since $\mathbb{E}[C] > 0$,

$$\{C = 0\} \subseteq \{|C - \mathbb{E}[C]| \geq \mathbb{E}[C]\}.$$

Chebyshev gives

$$\Pr(C = 0) \leq \frac{\mathrm{Var}(C)}{\mathbb{E}[C]^2} \leq \frac{1}{\mathbb{E}[C]} = \frac{m}{\binom{n}{2}}.$$

Hence, if $n/\sqrt{m} \to \infty$, then

$$\Pr(\text{no collision}) \to 0.$$

This argument is worth remembering because it combines four common tools: indicators, linearity of expectation, covariance calculations, and Chebyshev.

Choosing between Markov and Chebyshev

Markov	Chebyshev

Choosing between Markov and Chebyshev

	Markov	**Chebyshev**
Information	$X \geq 0$ and $\mathbb{E}[X]$	$\mathbb{E}[X]$ and $\mathrm{Var}(X)$

Choosing between Markov and Chebyshev

	Markov	**Chebyshev**
Information	$X \geq 0$ and $\mathbb{E}[X]$	$\mathbb{E}[X]$ and $\mathrm{Var}(X)$
Question	Is X very large?	Is X far from its mean?

Choosing between Markov and Chebyshev

	Markov	**Chebyshev**
Information	$X \geq 0$ and $\mathbb{E}[X]$	$\mathbb{E}[X]$ and $\mathrm{Var}(X)$
Question	Is X very large?	Is X far from its mean?
Bound	$\Pr(X \geq a) \leq \mathbb{E}[X]/a$	$\Pr(\lvert X - \mu \rvert \geq t) \leq \sigma^2/t^2$

Choosing between Markov and Chebyshev

	Markov	**Chebyshev**
Information	$X \geq 0$ and $\mathbb{E}[X]$	$\mathbb{E}[X]$ and $\mathrm{Var}(X)$
Question	Is X very large?	Is X far from its mean?
Bound	$\Pr(X \geq a) \leq \mathbb{E}[X]/a$	$\Pr(\lvert X - \mu \rvert \geq t) \leq \sigma^2/t^2$
Scale	multiples of the mean	multiples of the standard deviation

Choosing between Markov and Chebyshev

	Markov	**Chebyshev**
Information	$X \geq 0$ and $\mathbb{E}[X]$	$\mathbb{E}[X]$ and $\mathrm{Var}(X)$
Question	Is X very large?	Is X far from its mean?
Bound	$\Pr(X \geq a) \leq \mathbb{E}[X]/a$	$\Pr(\lvert X - \mu \rvert \geq t) \leq \sigma^2/t^2$
Scale	multiples of the mean	multiples of the standard deviation
Limitation	uses no spread information	still gives only polynomial tails

Choosing between Markov and Chebyshev

	Markov	**Chebyshev**
Information	$X \geq 0$ and $\mathbb{E}[X]$	$\mathbb{E}[X]$ and $\mathrm{Var}(X)$
Question	Is X very large?	Is X far from its mean?
Bound	$\Pr(X \geq a) \leq \mathbb{E}[X]/a$	$\Pr(\lvert X-\mu\rvert \geq t) \leq \sigma^2/t^2$
Scale	multiples of the mean	multiples of the standard deviation
Limitation	uses no spread information	still gives only polynomial tails

Methodological lesson

A concentration inequality is only as strong as the information fed into it. Before applying a bound, ask what features of the random variable the bound is actually using.

Where we go next

Today we used only the first two moments:

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

Where we go next

Today we used only the first two moments:

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

That was enough to obtain useful but coarse tail bounds.

For sums of independent bounded random variables, however, we know much more than their variance. We know how the summands are generated and how large each one can be.

Where we go next

Today we used only the first two moments:

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

That was enough to obtain useful but coarse tail bounds.

For sums of independent bounded random variables, however, we know much more than their variance. We know how the summands are generated and how large each one can be.

Next question

How can we exploit independence and boundedness to replace polynomial tail decay by exponential tail decay?

Where we go next

Today we used only the first two moments:

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

That was enough to obtain useful but coarse tail bounds.

For sums of independent bounded random variables, however, we know much more than their variance. We know how the summands are generated and how large each one can be.

Next question

How can we exploit independence and boundedness to replace polynomial tail decay by exponential tail decay?

That leads to moment-generating functions and Chernoff/Hoeffding bounds.

What to carry forward

- Expectation describes average behavior; a tail probability describes reliability.

What to carry forward

- Expectation describes average behavior; a tail probability describes reliability.
- Markov is the strongest universal tail statement available from a nonnegative mean alone.

What to carry forward

- Expectation describes average behavior; a tail probability describes reliability.
- Markov is the strongest universal tail statement available from a nonnegative mean alone.
- Variance measures fluctuation; covariance determines whether fluctuations reinforce or cancel in sums.

What to carry forward

- Expectation describes average behavior; a tail probability describes reliability.
- Markov is the strongest universal tail statement available from a nonnegative mean alone.
- Variance measures fluctuation; covariance determines whether fluctuations reinforce or cancel in sums.
- Independence kills covariance terms and makes averaging concentrate.

What to carry forward

- Expectation describes average behavior; a tail probability describes reliability.
- Markov is the strongest universal tail statement available from a nonnegative mean alone.
- Variance measures fluctuation; covariance determines whether fluctuations reinforce or cancel in sums.
- Independence kills covariance terms and makes averaging concentrate.
- Chebyshev converts variance into a deviation bound on the standard-deviation scale.

What to carry forward

- Expectation describes average behavior; a tail probability describes reliability.
- Markov is the strongest universal tail statement available from a nonnegative mean alone.
- Variance measures fluctuation; covariance determines whether fluctuations reinforce or cancel in sums.
- Independence kills covariance terms and makes averaging concentrate.
- Chebyshev converts variance into a deviation bound on the standard-deviation scale.
- When Chebyshev is dramatically loose, that is evidence that useful distributional structure is being discarded.

PART 4

Appendix

4

Independence forces covariance to vanish

Suppose X and Y are independent and have finite second moments. Recall that

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].$$

Independence forces covariance to vanish

Suppose X and Y are independent and have finite second moments. Recall that

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].$$

Because functions of independent random variables remain independent, $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$ are independent. Therefore their product expectation factors:

$$\operatorname{Cov}(X, Y) = \mathbb{E}[X - \mathbb{E}X]\,\mathbb{E}[Y - \mathbb{E}Y]$$

Independence forces covariance to vanish

Suppose X and Y are independent and have finite second moments. Recall that

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].$$

Because functions of independent random variables remain independent, $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$ are independent. Therefore their product expectation factors:

$$\begin{aligned}\operatorname{Cov}(X, Y) &= \mathbb{E}[X - \mathbb{E}X]\,\mathbb{E}[Y - \mathbb{E}Y] \\ &= 0 \cdot 0\end{aligned}$$

Independence forces covariance to vanish

Suppose X and Y are independent and have finite second moments. Recall that

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].$$

Because functions of independent random variables remain independent, $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$ are independent. Therefore their product expectation factors:

$$\begin{aligned}\operatorname{Cov}(X, Y) &= \mathbb{E}[X - \mathbb{E}X]\,\mathbb{E}[Y - \mathbb{E}Y] \\ &= 0 \cdot 0 \\ &= 0.\end{aligned}$$

Independence forces covariance to vanish

Suppose X and Y are independent and have finite second moments. Recall that

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].$$

Because functions of independent random variables remain independent, $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$ are independent. Therefore their product expectation factors:

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}[X - \mathbb{E}X]\,\mathbb{E}[Y - \mathbb{E}Y] \\ &= 0 \cdot 0 \\ &= 0.\end{aligned}$$

Variance of sums of independent r.v.s

If $X_1, \ldots, X_n$ are pairwise independent, then

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$