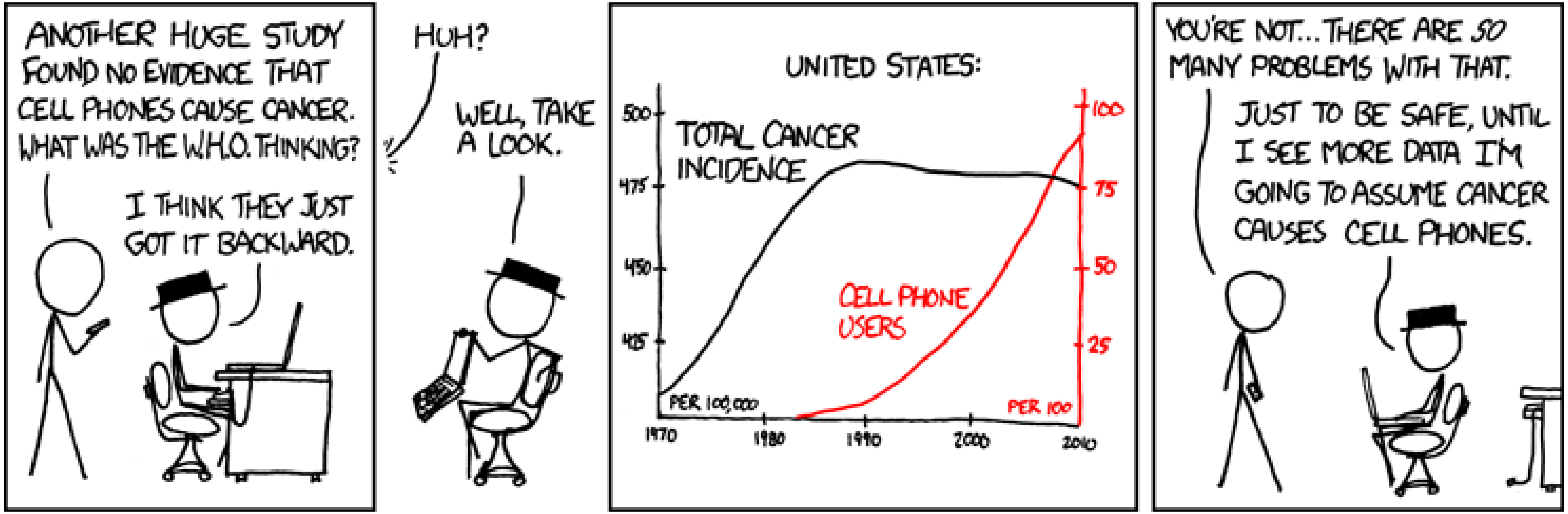


More Joint Distributions, Tail Bounds

CSE 312 26Su

Lecture 18

https://xkcd.com/925/

Announcements

joint
CLT

> Quiz 5 on Friday

Reference sheet on Ed

> Homework 5 due tonight

> Homework 6 released this evening

> After that, we'll be going over some applications which will also allow us to have some more review and practice with the main course content

Where Are We?

Today:

- Joint Distributions
  - Joint Support
  - Joint PMF / PDF
  - Joint CDF
  - Joint Expectation
- Conditional Distributions
  - Conditional Expectation
- Law of Total Expectation
- ~~Covariance~~

covariance

Today:

- Recap
  - One more joint distribution example
  - One more LTE example
- Tail Bounds
  - Markov's Inequality
  - Chebyshev's Inequality
  - (next: Chernoff Bound)
  - (next: not a tail bound, but Union Bound)

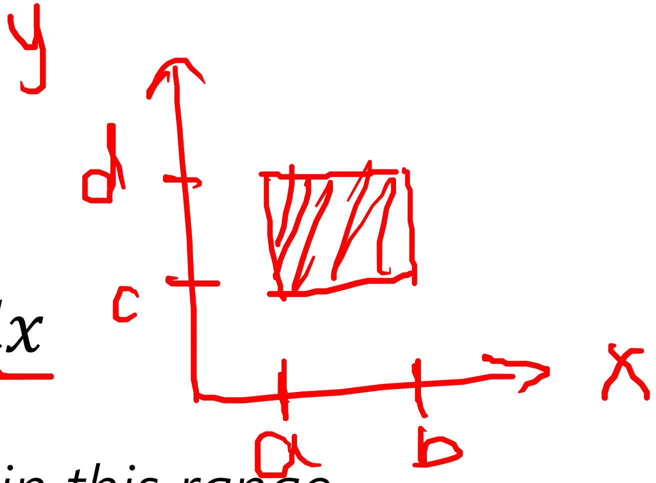
Joint Probabilities

To find probability of X and Y being in ranges, we use the joint distribution:

If $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are discrete….*

$$\mathbb{P}(a \le X \le b \cap c \le Y \le d) = \Sigma_{x \in a \le X \le b} \Sigma_{y \in c \le Y \le d}\, p_{X,Y}(x, y)$$

sum over the joint PMF for all pairs of x and y that fall in this range



If $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are continuous…*

$$\mathbb{P}(a \le X \le b \cap c \le Y \le d) = \int_a^b \int_c^d f_{X,Y}(x, y) dy\, dx$$

integrate over the joint PDF for all pairs of x and y that fall in this range

Example: Continuous Servers

$\Omega_X = [0, \infty]$

The time until server 1 crashes is $X \sim \text{Exp}(u)$, and the time until server 2 crashes is $Y \sim \text{Exp}(v)$. Both servers are independent of each other.

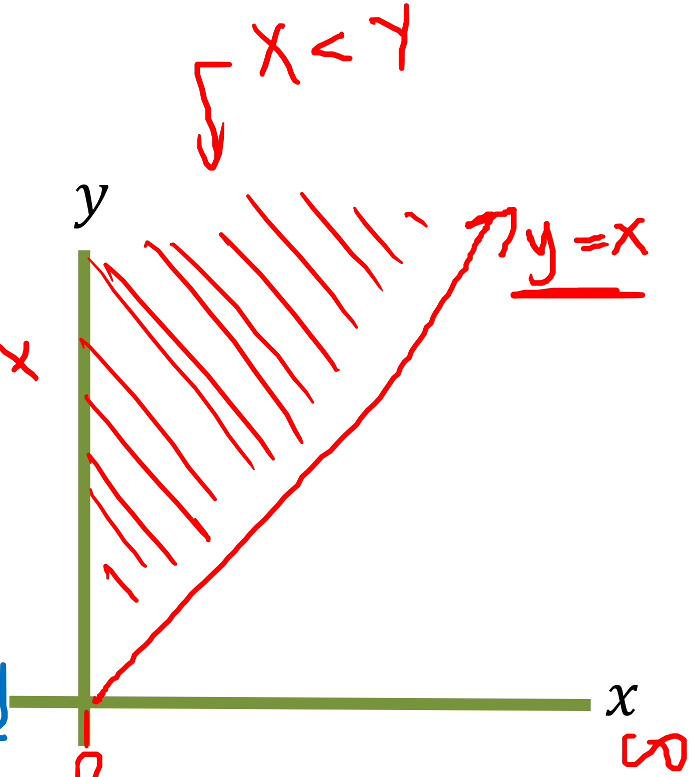
What is the probability server 1 crashes before server 2?

$$\mathbb{P}(X < Y) =$$

$$\int_0^\infty \int_x^\infty f_{X,Y}(x,y)\, dy\, dx$$

apply indp. $\rightarrow$ $\int_0^\infty \int_x^\infty f_X(x) f_Y(y)\, dy\, dx$

$\int_x^\infty \int_0^\infty f_{XY}(x,y)\, dx\, dy$



pollev.com/jolie312

Example: Continuous Servers

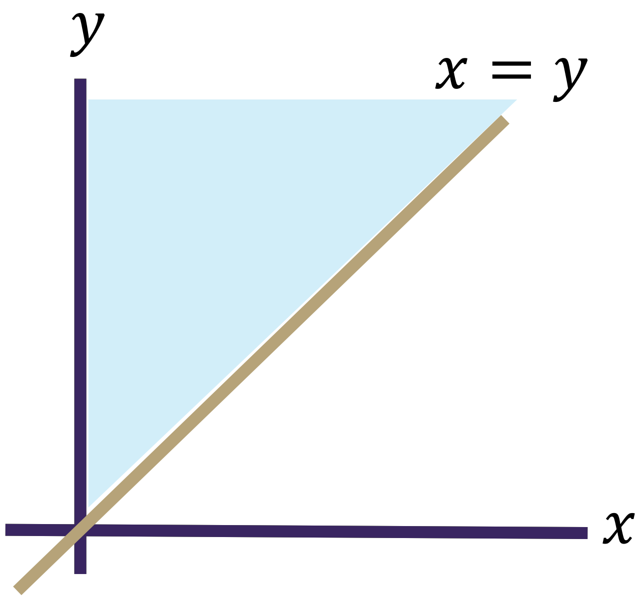
The time until server 1 crashes is $X \sim \text{Exp}(u)$, and the time until server 2 crashes is $Y \sim \text{Exp}(v)$. Both servers are independent of each other.

What is the probability server 1 crashes before server 2?

$$\mathbb{P}(X < Y)$$

$$= \int_0^\infty \int_x^\infty f_{X,Y}(x, y)\, dy\, dx$$

$$= \int_0^\infty \int_x^\infty f_X(x) f_Y(y)\, dy\, dx \text{ by independence}$$



Discrete vs. Continuous Joint Distributions

	Discrete	Continuous
Joint PMF/PDF	$p_{X,Y}(x,y) = P(X=x, Y=y)$	$f_{X,Y}(x,y) \neq P(X=x, Y=y)$
Joint CDF	$F_{X,Y}(x,y) = \sum_{t \leq x} \sum_{s \leq y} p_{X,Y}(t,s)$	$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(t,s)\,ds\,dt$
Normalization	$\sum_x \sum_y p_{X,Y}(x,y) = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$
Marginal PMF/PDF	$p_X(x) = \sum_y p_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$
Expectation	$E[g(X,Y)] = \sum_x \sum_y g(x,y) p_{X,Y}(x,y)$	$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy$
Conditional PMF/PDF	$p_{X\mid Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\mid Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
Conditional Expectation	$E[X \mid Y=y] = \sum_x x p_{X\mid Y}(x \mid y)$	$E[X \mid Y=y] = \int_{-\infty}^{\infty} x f_{X\mid Y}(x \mid y)\,dx$
Independence	$\forall x,y,\ p_{X,Y}(x,y) = p_X(x) p_Y(y)$	$\forall x,y,\ f_{X,Y}(x,y) = f_X(x) f_Y(y)$

Law of Total Expectation

Let $A_1, A_2, \ldots, A_k$ be a partition of the sample space, then

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X|A_i]\mathbb{P}(A_i)$$

Let X, Y be discrete random variables, then

$$\mathbb{E}[X] = \sum_{y\in\Omega_Y} \mathbb{E}[X|Y = y]\mathbb{P}(Y = y)$$

Similar in form to law of total probability, and the proof goes that way as well.

LTE Example: Elevator Rides 🛗

X people enter an elevator on the ground floor, where $X \sim Poi(10)$.
There are N floors above the ground floor, and each person is equally likely to get off at any of the N floors, independently of others.
What is the expected number of stops the elevator will make before discharging all the passengers?

LTE Example: Elevator Rides

X people enter an elevator on the ground floor, where $X \sim Poi(10)$.

There are N floors above the ground floor, and each person is equally likely to get off at any of the N floors, independently of others.

What is the expected number of stops the elevator will make before discharging all the passengers?

key insight: number of stops depends on number of passengers

LTE Example: Elevator Rides

X people enter an elevator on the ground floor, where $X \sim Poi(10)$.
There are N floors above the ground floor, and each person is equally likely to get off at any of the N floors, independently of others.
What is the expected number of stops the elevator will make before discharging all the passengers?

key insight: number of stops depends on number of passengers

Let S be the number of stops
We want to find $\mathbb{E}[S]$. Since S depends on the value of X, use LTE, partition on X

LTE Example: Elevator Rides

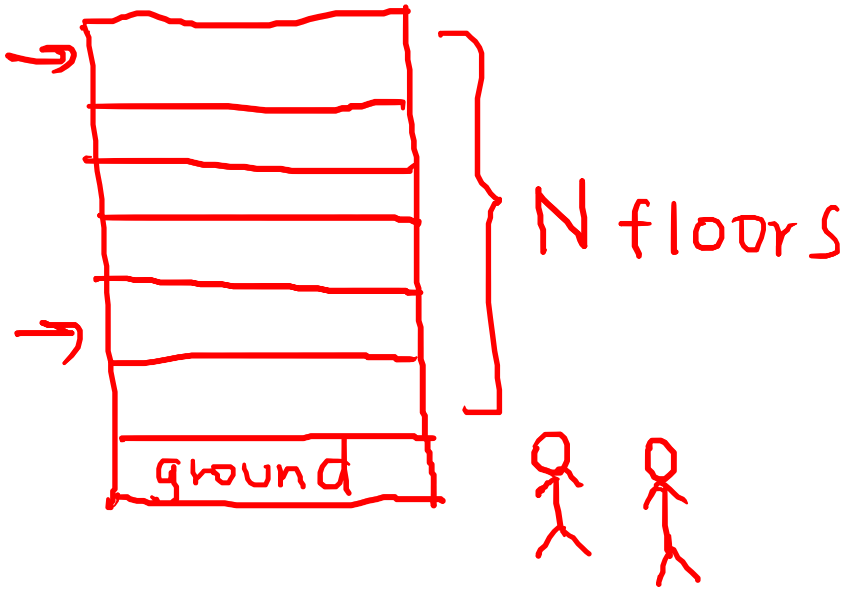
X - # of people
S - # of stops

$\Omega_X = \{0, \dots, \infty\}$

X people enter an elevator on the ground floor, where $X \sim Poi(10)$. There are N floors above the ground floor, and each person is equally likely to get off at any of the N floors, independently of others. What is the expected number of stops $\mathbb{E}[S]$ the elevator will make before discharging all the passengers?

LTE:

$$S = \sum_{i=1}^{N} S_i \qquad S_i - \begin{cases} 1 & \text{if stop on floor } i \\ 0 & \text{o.w.} \end{cases}$$



LOE:

$$E[S] = \sum_{i=1}^{N} E[S_i]$$

LTE:

$$E[S_i] = \sum_{k=0}^{\infty} E[S_i \mid X = k]\, P(X = k)$$

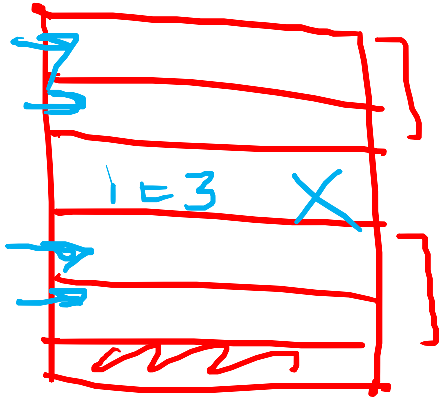
LTE Example: Elevator Rides

$S_i = \begin{cases} 1 & \text{if stop on } i\text{th fl.} \\ 0 & \text{ow} \end{cases}$

now:

$$\underline{E[S_i \mid X = k]} = \underline{P(S_i = 1 \mid X = k)}$$

k ppl on ground
prob stop on ith fl.



$$= 1 - \underline{P(S_i = 0 \mid X = k)}$$

no one gets off on ith floor

$$= 1 - \left(\frac{N-1}{N}\right)^k$$

$$E[S] = \sum_{i=1}^{N} \sum_{k=0}^{\infty} \left(1 - \left(\frac{N-1}{N}\right)^k\right) \cdot e^{-10} \cdot \frac{10^k}{k!}$$

Covariance

Covariance

$\mathrm{Cov}(X, Y)$ measures how much one random variable varies with a second.

Can be:

Positive (↑ in one var, ↑ in other)

Zero

Negative (↑ in one var, ↓ in other)

↓ ↑

Covariance

$\text{Cov}(X, Y)$ measures how much one random variable varies with a second.

- How "intertwined" X and Y are – how much knowing about one of them will affect the other
- If X is "big" how likely is it that Y will be "big"? How much do X and Y vary together?

Covariance

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

Properties of Covariance

Covariance

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

1. $Cov(X, Y) = Cov(Y, X)$ mult. comm.
2. $Var(X) = Cov(X, X)$ $Cov(X,X) = E[X^2] - E[X]^2$
3. $Cov(aX + b, Y) = a \cdot Cov(X, Y)$ LOE
4. $Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} Cov(X_i, Y_j)$ LOE

Properties of Covariance

Covariance

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

X, Y independent

If $X \perp Y$, what do you expect $Cov(X, Y)$ to be? 0

Properties of Covariance

Covariance

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

If $X \perp Y$, what do you expect $Cov(X, Y)$ to be? 0.

For independent X, Y, $\underline{\mathbb{E}[XY]} = \underline{\mathbb{E}[X]\mathbb{E}[Y]}$

Properties of Covariance

Does the opposite hold (i.e., does zero covariance imply independence?)

$X = \{-1, 0, 1\}$ with equal probability. $Y = 1$ if $X = 0$ and 0 otherwise.

$Y \backslash X$	-1	0	1	$p_Y(y)$
y = 0 0	$1/3$	0	$1/3$	$2/3$
y = 1 1	0	$1/3$	0	$1/3$
$p_X(x)$	$1/3$	$1/3$	$1/3$	

← joint pmf

marginal

Properties of Covariance

$Cov(X,Y) = E[XY] - E[X]E[Y]$

Does the opposite hold (i.e., does zero covariance imply independence?)

$X = \{-1, 0, 1\}$ with equal probability. $Y = 1$ if $X = 0$ and 0 otherwise.

$Y \backslash X$	-1	0	1	$p_Y(y)$
0	1/3	0	1/3	2/3
1	0	1/3	0	1/3
$p_X(x)$	1/3	1/3	1/3	

$$\mathbb{E}[X] = -1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0$$

$$\mathbb{E}[Y] = 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} = \frac{1}{3}$$

$$\mathbb{E}[XY] = (-1 \cdot 0) \cdot \frac{1}{3} + (0 \cdot 1) \cdot \frac{1}{3} + (1 \cdot 0) \cdot \frac{1}{3} = \boxed{0}$$

Properties of Covariance

Does the opposite hold (i.e., does zero covariance imply independence?)

$X = \{-1, 0, 1\}$ with equal probability. $Y = 1$ if $X = 0$ and 0 otherwise.

$Y \backslash X$	-1	0	1	$p_Y(y)$
0	$1/3$	0	$1/3$	$2/3$
1	0	$1/3$	0	$\boxed{1/3}$
$p_X(x)$	$1/3$	$1/3$	$1/3$	

$$\mathbb{E}[X] = -1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0$$
$$\mathbb{E}[Y] = 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} = \frac{1}{3}$$
$$\mathbb{E}[XY] = (-1 \cdot 0) \cdot \frac{1}{3} + (0 \cdot 1) \cdot \frac{1}{3} + (1 \cdot 0) \cdot \frac{1}{3} = 0$$

$$Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 \cdot \frac{1}{3} = \underline{0}$$
$$\mathbb{P}(\underline{Y = 1 | X = 0}) = \underline{1} \neq \frac{1}{3} = \mathbb{P}(Y = 1)$$

X, Y are dependent.

Properties of Covariance

Does the opposite hold (i.e., does zero covariance imply independence?)

$X = \{-1, 0, 1\}$ with equal probability. $Y = 1$ if $X = 0$ and 0 otherwise.

$Y \backslash X$	-1	0	1	$p_Y(y)$
0	$1/3$	0	$1/3$	$2/3$
1	0	$1/3$	0	$1/3$
$p_X(x)$	$1/3$	$1/3$	$1/3$	

$$\mathbb{E}[X] = -1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0$$
$$\mathbb{E}[Y] = 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} = \frac{1}{3}$$
$$\mathbb{E}[XY] = (-1 \cdot 0) \cdot \frac{1}{3} + (0 \cdot 1) \cdot \frac{1}{3} + (1 \cdot 0) \cdot \frac{1}{3} = 0$$

$$Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 \cdot \frac{1}{3} = 0$$

$\mathbb{P}(Y = 1 | X = 0) = 1 \neq \frac{1}{3} = \mathbb{P}(Y = 1)$ (i.e., X and Y are not independent)

So, zero covariance **does not imply** independence

Covariance *used for variance of a sum*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$

^ *we do not need* X and Y to be independent to use this formula!

Proof:

$$E[(X+Y)^2] - E[X+Y]^2$$

$$\begin{aligned}
&\mathrm{Var}(X+Y) \\
&= \mathrm{Cov}(X+Y, X+Y) \\
&= \mathrm{Cov}(X,X) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X) + \mathrm{Cov}(Y,Y) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)
\end{aligned}$$

Covariance *used for variance of a sum*

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

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$$

You and your friend are playing a game, you flip a coin: if heads you pay your friend a dollar, if tails they pay you a dollar. Let X be your profit and Y be your friend's profit.

What is $\mathrm{Var}(X + Y)$?

pollev.com/jolie312

Before you calculate, make a prediction.
What should it be?

Covariance *used for variance of a sum*

You and your friend are playing a game, you flip a coin: if heads you pay your friend a dollar, if tails they pay you a dollar. Let X be your profit and Y be your friend's profit.

What is $\text{Var}(X+Y)$?

$\text{Var}(X) = \text{Var}(Y) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 1 - 0^2 = 1$

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

$\mathbb{E}[XY] = \frac{1}{2} \cdot (-1 \cdot 1) + \frac{1}{2}(1 \cdot -1) = -1$

$\text{Cov}(X, Y) = -1 - 0 \cdot 0 = -1.$

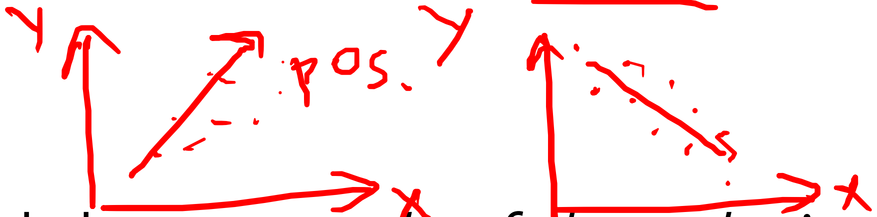
$\text{Var}(X+Y) = 1 + 1 + 2 \cdot -1 = 0$

Covariance

The magnitude of covariance is affected by the units of the random variables involved because $\mathrm{Cov}(2X, Y) = 2\mathrm{Cov}(X, Y)$, so we can't really compare and it's not very helpful

is covariance big because of the units or because of a very strong relationship?

The sign of the covariance (positive or negative) is helpful but it only tells us the direction.



We want to understand the *strength of the relationship!*

Pearson Correlation *(normalized covariance!)*

To understand the strength, we normalize the covariance!

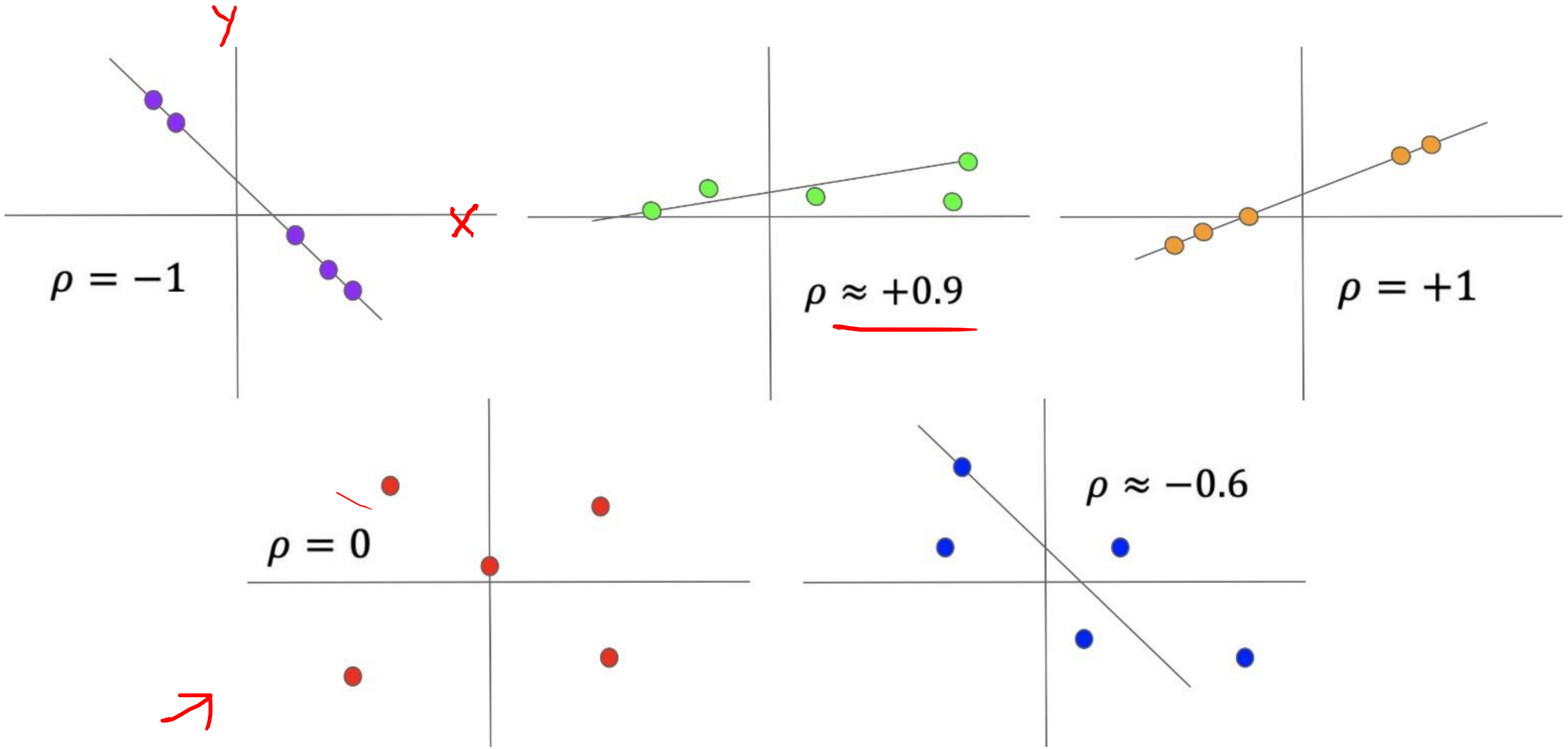
Pearson correlation: $\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$

> divide the covariance by the product of the standard deviation of X and the standard deviation of Y

> a value in the range $-1 \leq \rho(X,Y) \leq 1$ [−1, 1]

-1 means STRONG negative correlation, $+1$ means STRONG positive correlation

Pearson Correlation *(normalized covariance!)*

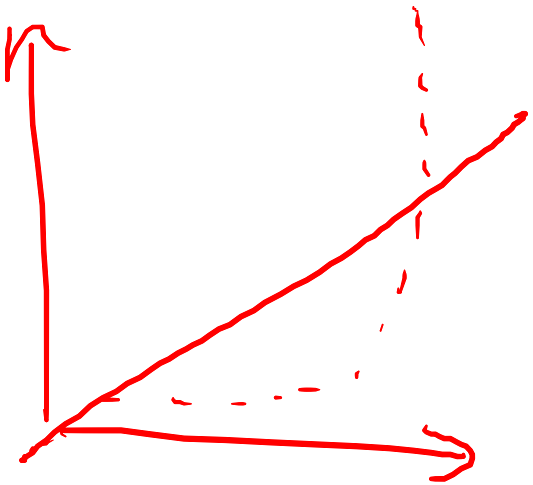


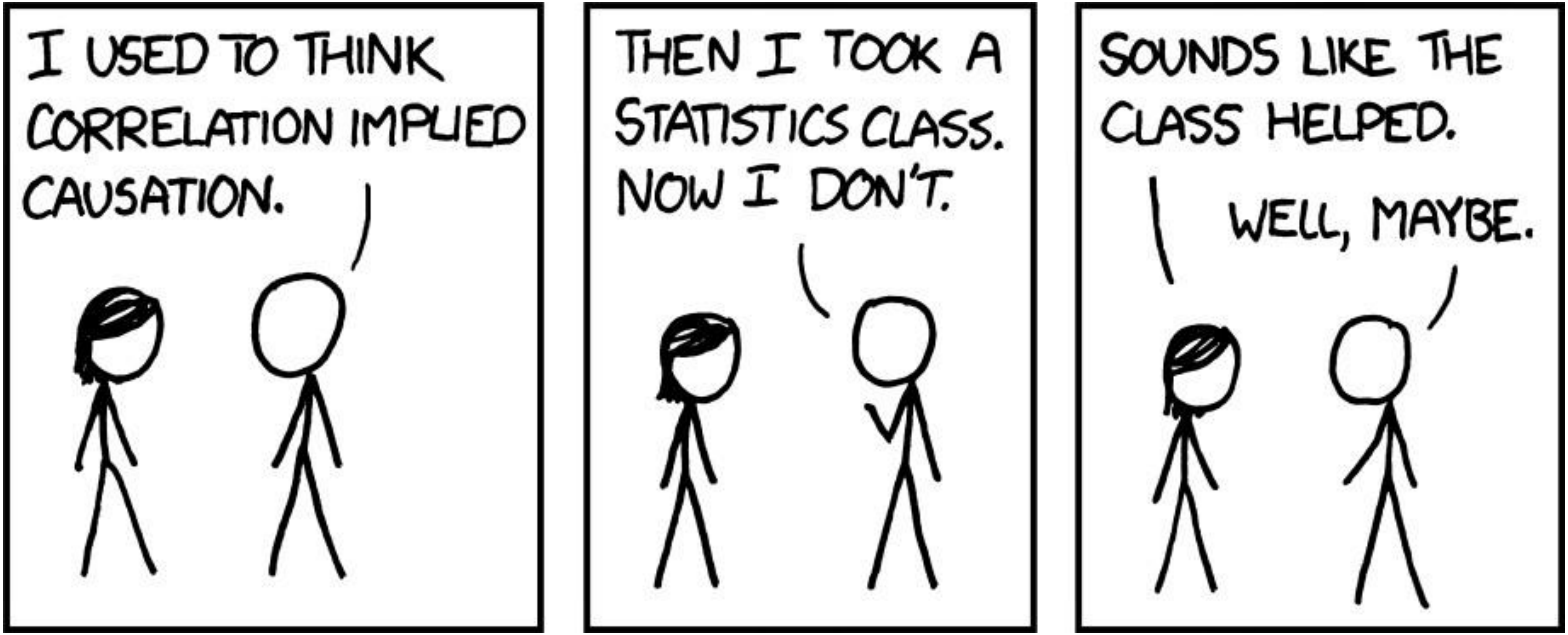
Additional Notes

Covariance directly measures only "linear" relationships; if Y depends on X^2, the covariance might not be as high as you expect.

$Z = X^2$

If dealing with real data, look at a plot to see if you should be looking for a linear relationship in the first place.





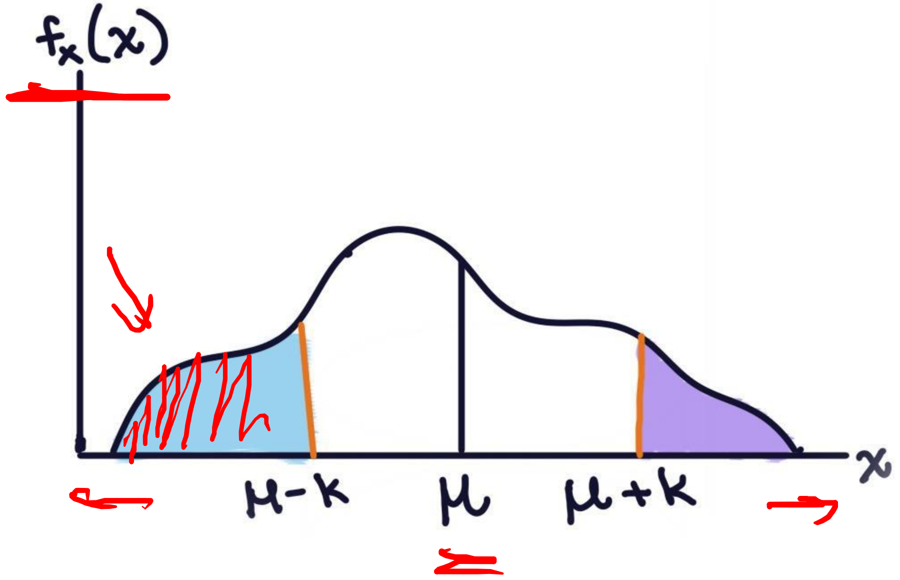
https://xkcd.com/552/

Tail Bounds

What's a Tail Bound?

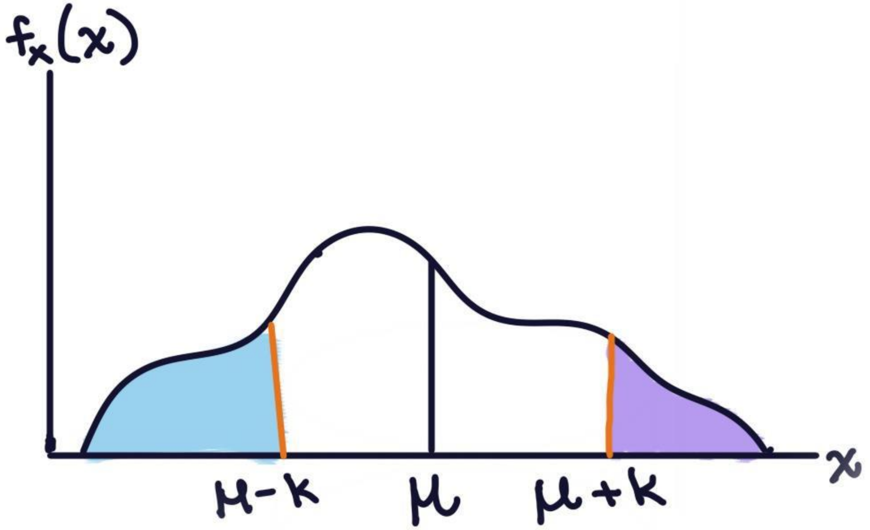
The "tails" of a probability distribution are the extreme regions to the left or right of the expectation

e.g., the shaded regions $X \leq \mu + k$ *and* $X \geq \mu - k$ *are "tails" of the distribution*



What's a Tail Bound?

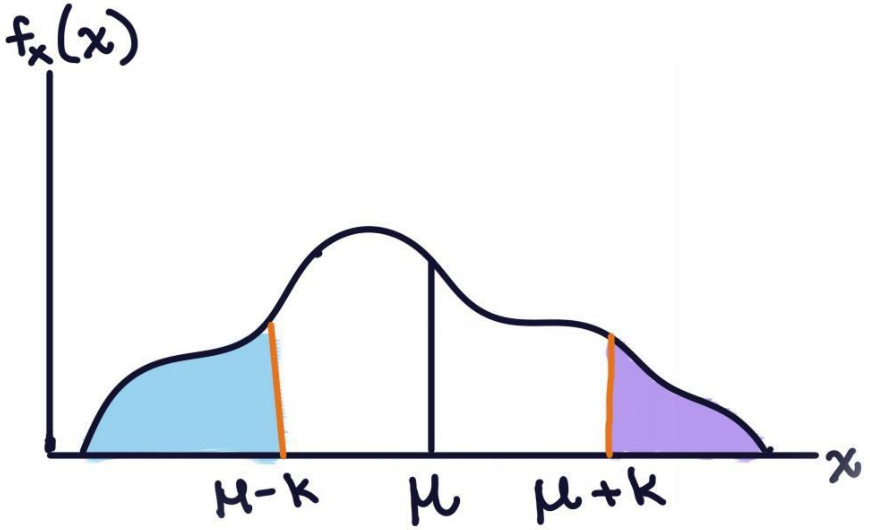
The "tails" of a probability distribution are the extreme regions to the left or right of the expectation
e.g., the shaded regions $X \leq \mu + k$ *and* $X \geq \mu - k$ *are "tails" of the distribution*



- Often, we want to make some guarantees about the probability of being in a tail is (e.g., $\mathbb{P}(X \geq k) \leq ??$)
guarantees about the running time (the chance of being > 5sec is no more than __)

What's a Tail Bound?

The "tails" of a probability distribution are the extreme regions to the left or right of the expectation
e.g., the shaded regions $X \leq \mu + k$ *and* $X \geq \mu - k$ *are "tails" of the distribution*



Often, we want to make some guarantees about the probability of being in a tail is (e.g., $\mathbb{P}(X \geq k) \leq ??$)
guarantees about the running time (the chance of being > 5sec is no more than __)

A **tail bound** (or concentration inequality) is a statement that bounds the probability in the "tails" of the distribution (e.g., there's little probability far from the center) or (equivalently) the probability is concentrated near the expectation.

What's a Tail Bound?

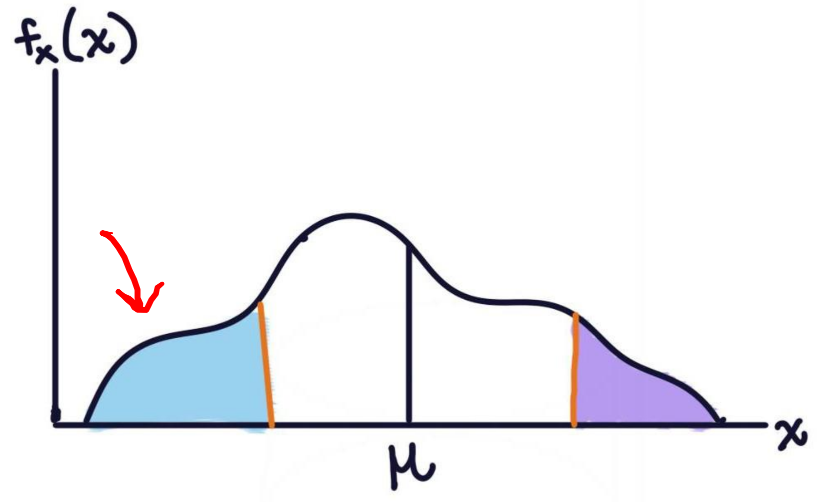
A tail bound (or concentration inequality) bounds the probability in the "tails" of the distribution. e.g., statements like $\mathbb{P}(X \geq 4) \leq 0.8$, $\mathbb{P}(X \geq 4) \leq 0.8$

What’s a Tail Bound?

A tail bound (or concentration inequality) bounds the probability in the “tails” of the distribution. e.g., statements like $\mathbb{P}(X \geq 4) \leq 0.8$, $\cancel{\mathbb{P}(X \geq 4) \leq 0.8}$

We’ve seen this before! We can:

- Compute these probabilities exactly in some cases
- Approximate X as normal using CLT if X is the sum of a bunch of i.i.d random variables

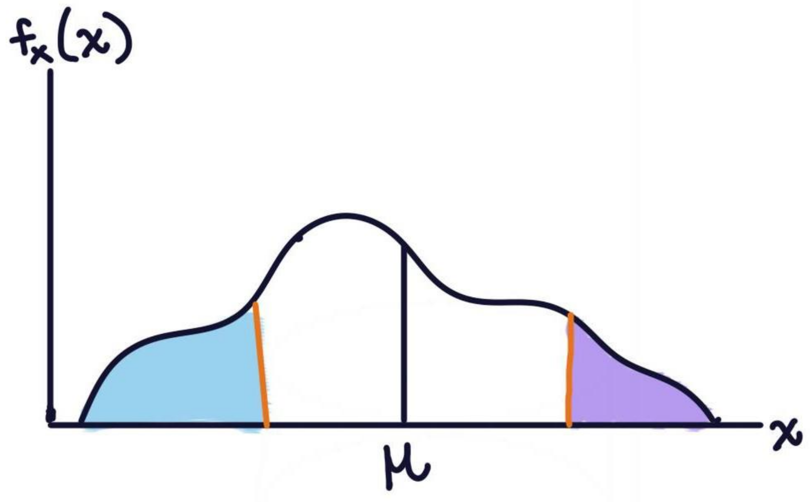


What’s a Tail Bound?

A tail bound (or concentration inequality) bounds the probability in the “tails” of the distribution. e.g., statements like $\mathbb{P}(X \geq 4) \leq 0.8$, $\mathbb{P}(X \geq 4) \leq 0.8$

We’ve seen this before! We can:

- Compute these probabilities exactly in some cases
- Approximate X as normal using CLT if X is the sum of a bunch of i.i.d random variables



But what if we barely know anything about X and it doesn't fit into the frameworks we've learned about? Can we still make some tail bound guarantees?

Tail Bounds

We're going to learn about 3 tail bounds that we can use when all we know about X is it's expected value and/or variance:

$E[X]$ Var(X)

- Markov's Inequality
- Chebysev's Inequality
- Chernoff Bound

Fri

Upper vs. Lower Bound

If we find something like $\mathbb{P}(A) \boxed{\leq b}$, we found an upper bound

This *highest/"uppermost"* value the probability of A could be is b

If we find something like $\mathbb{P}(A) \boxed{\geq b}$, we found a lower bound

This *lowest/smallest* value the probability of A could be is b

Markov's Inequality

Two statements are equivalent.
Left form is often easier to use.
Right form is more intuitive.

Markov's Inequality

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

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

Markov's Inequality

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

$$\mathbb{P}(X \geq k\mathbb{E}[X]) \leq \frac{1}{k}$$

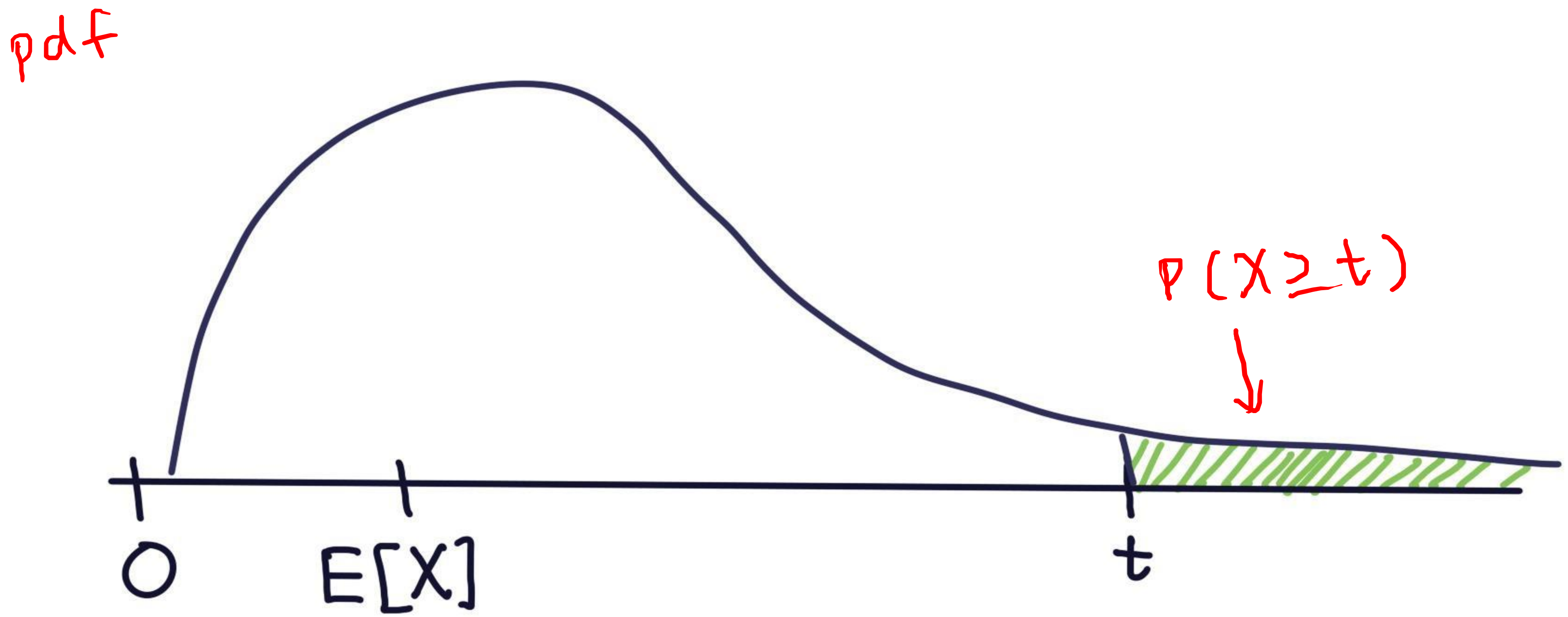
let $t = k \cdot E[X]$

$k > 0$

Requirements:

1. X must be non-negative
2. We know the expectation of X

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



Proof

partion $X<t$, $X\geq t$

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

$$\mathbb{E}[X] = \mathbb{E}[X|X<t]\mathbb{P}(X<t) + \mathbb{E}[X|X\geq t]\mathbb{P}(X\geq t)$$

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

$\underbrace{\mathbb{E}[X|X\geq t]}_{\geq 0}\underbrace{\mathbb{P}(X\geq t)}_{\in[0,1]} \geq 0$ if X is non-negative

$$\geq t\,\mathbb{P}(X\geq t)$$

$$\boxed{\mathbb{E}[X]\geq t\cdot\mathbb{P}(X\geq t)}$$

Doing some algebra…we get exactly what's in Markov's inequality! ➔

Markov's Inequality

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

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

Example: Let's see how good this bound is…

Suppose you roll a fair (6-sided) die until you see a 6. Let X be the number of rolls. Bound the probability that $X \geq 12$.

$p = 1/6$

$X \sim \text{Geo}\left(\frac{1}{6}\right)$, so $\mathbb{E}[X] = 1/(\frac{1}{6}) = 6$

x is non-neg ✓

E[X] = 6 ✓

Applying Markov's Inequality…

$$\mathbb{P}(X \geq 12) \leq \frac{\mathbb{E}[X]}{12} = \frac{6}{12} = \boxed{\frac{1}{2}}$$

Markov's Inequality

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

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

Example: Let's see how good this bound is…

Suppose you roll a fair (6-sided) die until you see a 6. Let X be the number of rolls. Bound the probability that $X \geq 12$.

$X \sim \mathrm{Geo}\left(\frac{1}{6}\right)$, so $\mathbb{E}[X] = 1/(\frac{1}{6}) = 6$

Applying Markov's Inequality…

$$\mathbb{P}(X \geq 12) \leq \frac{\mathbb{E}[X]}{12} = \frac{6}{12} = \boxed{\frac{1}{2}}$$

Exact probability?

$$\underline{1 - \mathbb{P}(X < 12)} \approx 1 - 0.865 = \boxed{0.135}$$

Markov's Inequality

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

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

Example: Ads

X - # of ads $\quad\quad\quad\quad$ $E[X] = 25$

Suppose the average number of ads you see on a website is 25. Give an upper bound on the probability of seeing a website with 75 or more ads.

$$P(X \geq 75) \leq \frac{E[X]}{75} = \frac{25}{75} = \frac{1}{3}$$

$t = 75$

Markov's Inequality

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

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

Example: Ads

Suppose the average number of ads you see on a website is 25. Give an upper bound on the probability of seeing a website with 75 or more ads.

$$\mathbb{P}(X \geq 75) \leq \frac{\mathbb{E}[X]}{75} = \frac{25}{75} = \boxed{\frac{1}{3}}$$

Markov's Inequality

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

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

Example: More Ads

X — # of ads $E[X]=25$

Suppose the average number of ads you see on a website is 25. Give an upper bound on the probability of seeing a website **with 20 or more ads**.

$$P(X \geq 20) \leq \frac{E[X]}{20} = \frac{25}{20} = 1.25$$

~~Fill out the poll everywhere: pollev.com/jolie312~~

Markov's Inequality

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

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

Example: More Ads

Suppose the average number of ads you see on a website is 25. Give an upper bound on the probability of seeing a website with 20 or more ads.

$$\mathbb{P}(X \geq 20) \leq \frac{\mathbb{E}[X]}{20} = \frac{25}{20} = \underline{1.25}$$

$[0,1]$

Well, that's…true. Technically.

But without more information we couldn't hope to do much better. What if every page gives exactly 25 ads? Then the probability really is 1.

So…what do we do?

A better inequality!

We're trying to bound the tails of the distribution.

What parameter of a random variable describes the tails?

The variance!

Chebyshev's Inequality

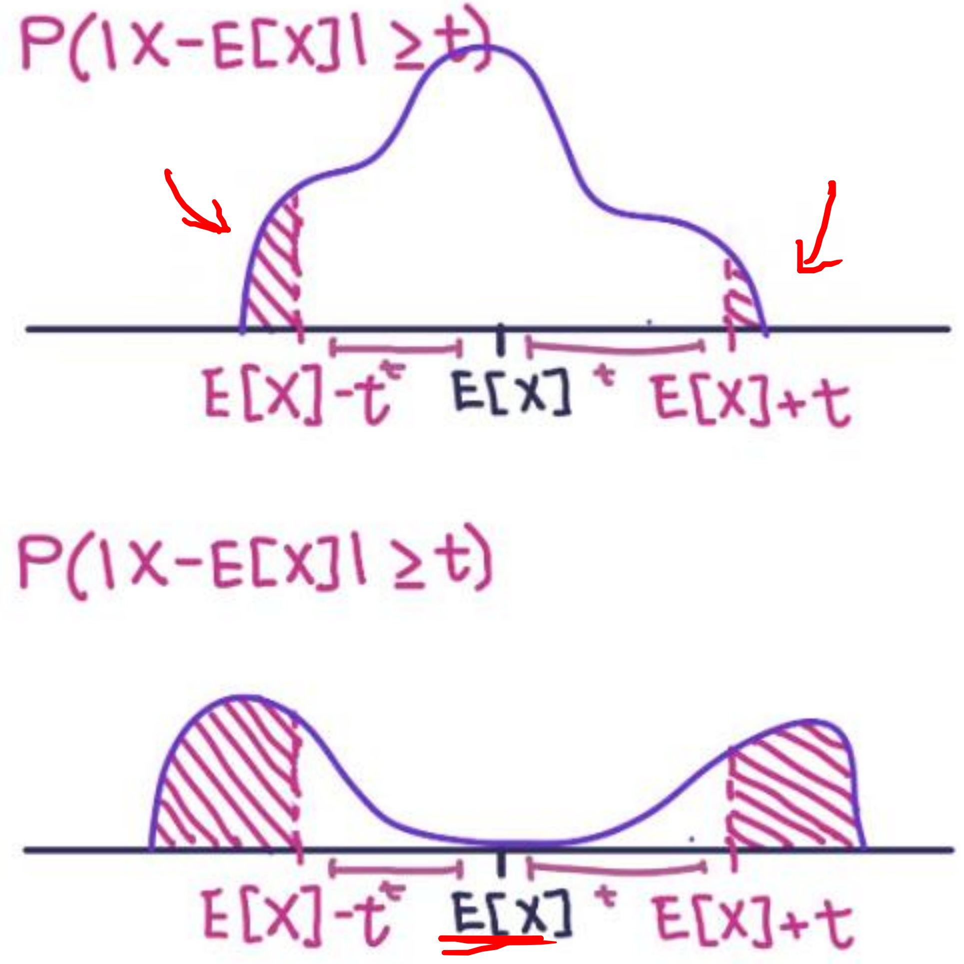
Chebyshev's Inequality

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

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

Requirements:

1. We know the expectation of X
2. We know the variance of X



Chebyshev's Inequality

Two statements are equivalent.
Left form is often easier to use.
Right form is more intuitive.

Chebyshev's Inequality

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

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

Chebyshev's Inequality

Let X be a random variable. For any $\mathrm{k} > 0$

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq k\sqrt{\mathrm{Var}(X)}\right) \leq \frac{1}{k^2}$$

Requirements:

1. We know the expectation of X
2. We know the variance of X

"probability we're at least k standard deviations away from the mean is $\leq \frac{1}{k^2}$"

Proof of Chebyshev

Chebyshev's Inequality

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

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

Let $Z = X - \mathbb{E}[X]$

$$\mathbb{P}(|Z| \geq t)$$

$$= \mathbb{P}(Z^2 \geq t^2)$$

Inequalities are equivalent (square each side)

$$\leq \frac{\mathbb{E}[Z^2]}{t^2}$$

Markov's Inequality

$$= \frac{\mathbb{E}[Z^2] - (\mathbb{E}[Z])^2}{t^2}$$

$\mathbb{E}[Z] = 0$

$$= \frac{\mathrm{Var}(Z)}{t^2}$$

$$\boxed{= \frac{\mathrm{Var}(X)}{t^2}}$$

Z is just X shifted. Variance is unchanged.

Example with geometric RV

Suppose you roll a fair (6-sided) die until you see a 6. Let X be the number of rolls. Bound the probability that $X \geq 12$

$$\mathbb{P}(X \geq 12) \leq \mathbb{P}(|X-6| \geq 6) \leq \frac{\frac{5/6}{1/36}}{6^2} = \frac{5}{6}$$

Not any better than Markov ☹

Chebyshev's Inequality

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

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

Example with geometric RV

Let X be a geometric rv with parameter p

Bound the probability that $X \geq \frac{2}{p}$

$$\mathbb{P}(X \geq 2/p) \leq \mathbb{P}(|X - 1/p| \geq 1/p) \leq \frac{\frac{1-p}{p^2}}{1/p^2} = 1 - p$$

Markov gives:

$$\mathbb{P}\left(X \geq \frac{2}{p}\right) = \frac{\mathbb{E}[X]}{2/p} = \frac{1}{p} \cdot \frac{p}{2} = \frac{1}{2}.$$

For large p, Chebyshev is better.

Chebyshev's Inequality

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

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

Markov's inequality gave $\mathbb{P}(X \geq 75) \leq 1/3$

Example: Ads (but better!)

Suppose the average number of ads you see on a website is 25. And the variance of the number of ads is 16. Give an upper bound on the probability of seeing a website with 75 or more ads.

Chebyshev's Inequality

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

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

Example: Ads (but better!)

Suppose the average number of ads you see on a website is 25. And the variance of the number of ads is 16. Give an upper bound on the probability of seeing a website with 75 or more ads.

$$\mathbb{P}(X \geq 75) = \mathbb{P}(X - 25 \geq 75 - 25) \leq \mathbb{P}(|X - 25| \geq 50) \leq \frac{16}{50^2}$$

Chebyshev's Inequality

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

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

Tail Bounds – Takeaways

Useful when an experiment is complicated and you just need the probability to be small (you don't need the exact value).

Choosing a minimum n for a poll – don't need exact probability of failure, just to make sure it's small.

Designing probabilistic algorithms – just need a guarantee that they'll be extremely accurate

Learning more about the situation (e.g. learning variance instead of just mean) usually lets you get more accurate bounds.

Next: more assumptions to get better bounds.