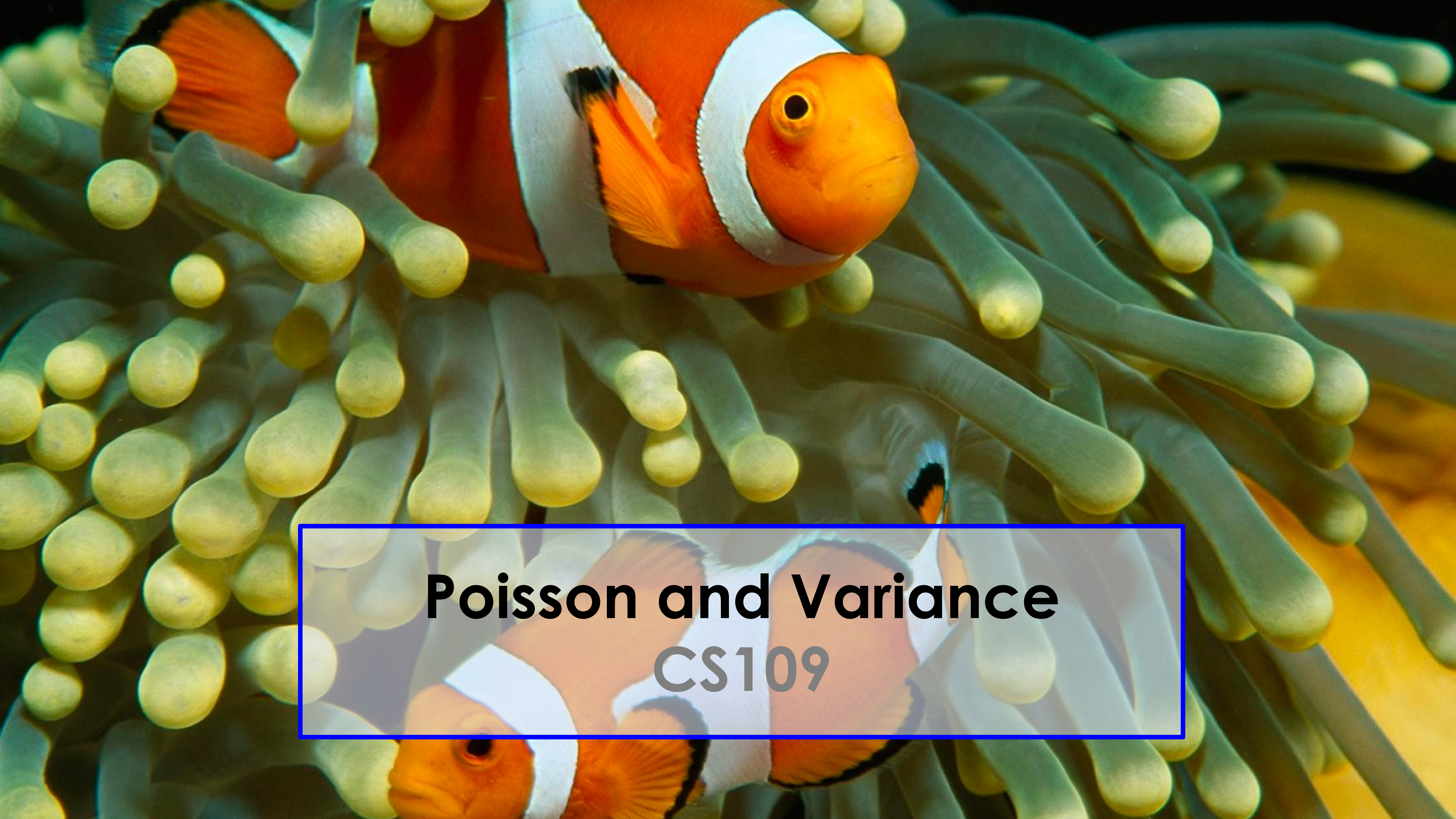


Poisson and Variance

CS109

Probability of Extreme Weather?

Review

Pokémon™

Gotta catch 'em all!™

(Classic Random Variables)

The Geometric Random Variable

Imagine flipping a coin *until you see your first heads*.

Each coin flip is an independent trial, with probability *p* of getting heads.

Want to model: how many coin flips until the first heads?

$$X \sim \text{Geo}(p)$$

$$P(X = n) = (1 - p)^{n-1} p$$

Like throwing pokeballs
until you catch a pokemon!

The Negative Binomial Random Variable

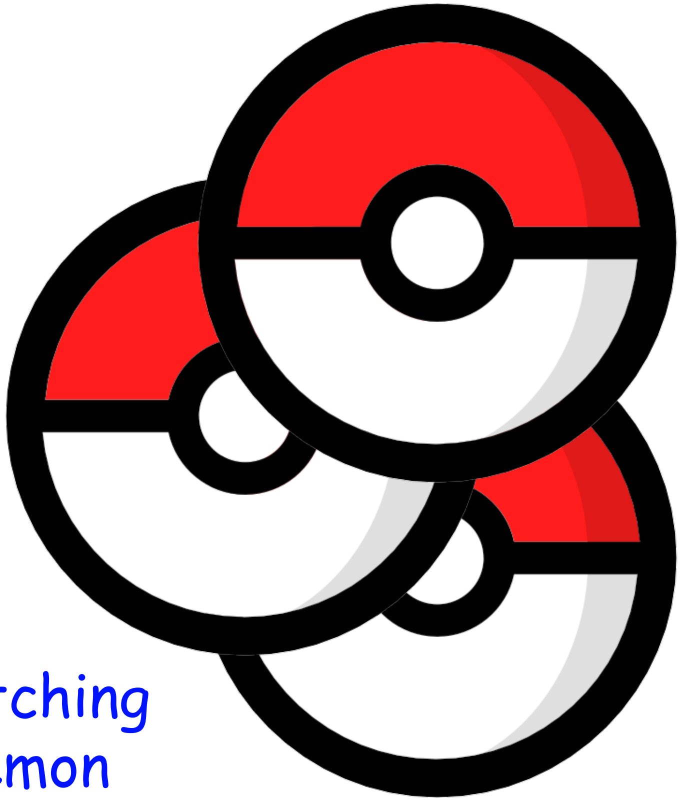
Imagine flipping a coin *until you see* ***r*** *heads*.

Each coin flip is an independent trial, with probability *p* of getting heads.

Want to model: how many coin flips until ***r*** heads?

$$X \sim \text{NegBin}(r, p)$$

$$P(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$



Like catching *r* pokemon

Can Jacob Bernoulli Have a Variable Named After Him?

Here yee. I want to have a random variable named after myself. Huzzah.

Yes - the Bernoulli random variable: $X \sim \text{Bern}(p)$

- The Bernoulli is an **indicator** random variable (value is either 0 or 1).
- $P(X = 1) = p$
- $P(X = 0) = 1 - p$

(this is the whole PMF)

- Examples: a single coin flip, one ad click, any binary event

Expected Value or Expectation

Expected value answers the question:

What is the average value we could expect some random variable to be?

A possible value of X

The probability of that value

$$E[X] = \sum_{x} x \cdot \mathrm{P}(X = x)$$

Loop over all values x that X can take on

Helpful Properties of Expectation

1. Linearity:

$$E[aX + b] = aE[X] + b$$

2. Expectation of a sum is the sum of expectations:

$$E[X + Y] = E[X] + E[Y]$$

These are all true, no matter what random variables X and Y are

3. Law of the Unconscious Statistician:

$$E[g(x)] = \sum_{x \in X} g(x) P(X = x)$$

Expectations of Classic Random Variables

$X \sim \mathrm{Geo}(p)$

$$E[X] = \frac{1}{p}$$

$Y \sim \mathrm{NegBin}(r, p)$

$$E[Y] = \frac{r}{p}$$

Expectations of Classic Random Variables

$X \sim \text{Geo}(p)$

$$E[X] = \frac{1}{p}$$

$X \sim \text{Bern}(p)$

$$E[X] = p$$

$Y \sim \text{NegBin}(r, p)$

$$E[Y] = \frac{r}{p}$$

$Y \sim \text{Bin}(n, p)$

$$E[Y] = n \cdot p$$

Pokemon: Actually Catching Them All

To catch a Pokemon, you throw a pokeball repeatedly until it’s caught.

Each pokeball has a 1/3 chance of catching the Pokemon.

What is the **expected number** of pokeballs needed to catch 1 Pokemon?

There are 151 Pokemon to catch in the game Pokemon Diamond.

What is the **expected number** of pokeballs needed to catch *every* Pokemon?

Pokemon: Actually Catching Them All

To catch a Pokemon, you throw a pokeball repeatedly until it's caught.

Each pokeball has a 1/3 chance of catching the Pokemon.

What is the **expected number** of pokeballs needed to catch 1 Pokemon?

Let X be the number of pokeballs we use.

$X \sim \text{Geo}(p = 1/3)$

$$E[X] = \frac{1}{p} = 3$$

There are 151 Pokemon to catch in the game Pokemon Diamond.

What is the **expected number** of pokeballs needed to catch *every* Pokemon?

Pokemon: Actually Catching Them All

To catch a Pokemon, you throw a pokeball repeatedly until it's caught.

Each pokeball has a 1/3 chance of catching the Pokemon.

What is the **expected number** of pokeballs needed to catch 1 Pokemon?

Let X be the number of pokeballs we use.

$X \sim \text{Geo}(p = 1/3)$

$$E[X] = \frac{1}{p} = 3$$

There are 151 Pokemon to catch in the game Pokemon Diamond.

What is the **expected number** of pokeballs needed to catch *every* Pokemon?

Let Y be the number of pokeballs we use in total.

$Y \sim \text{NegBin}(r = 151, p = 1/3)$

$$E[Y] = \frac{r}{p} = 151 \cdot 3 = 453$$

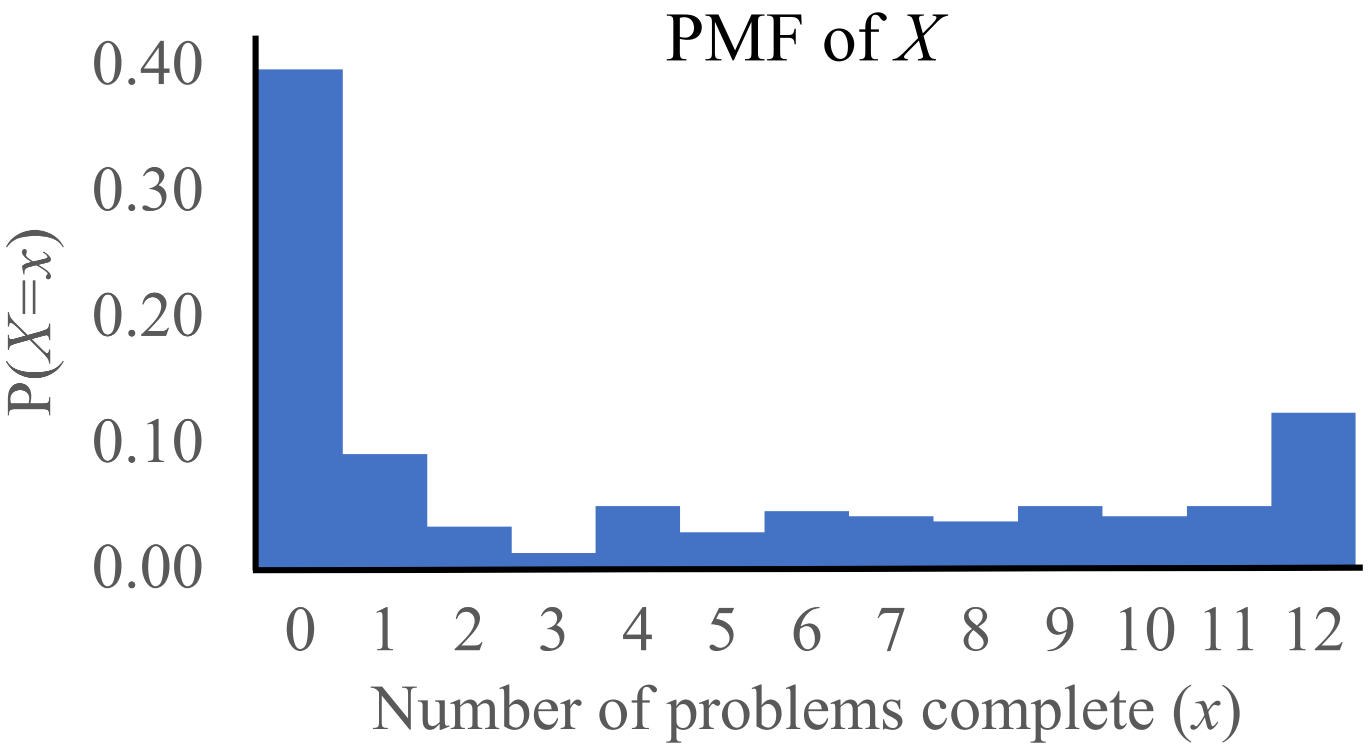
End Review

Expectation is only a single number summary…

Expectation Is Not All You Need

Let *X* be the number of problems on pset2 that a randomly selected student has completed, as of Monday morning.

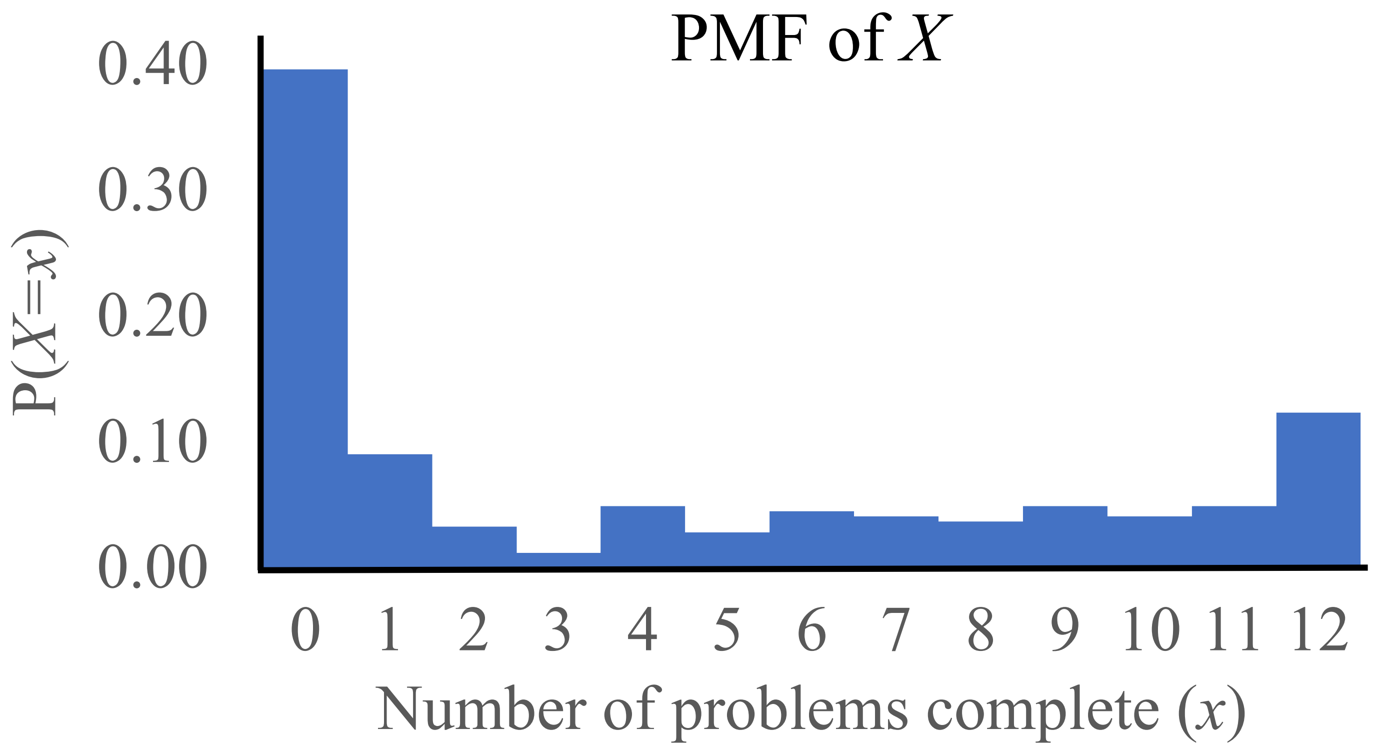
X takes on values with uncertainty, so *X* is a random variable.



Expectation Is Not All You Need

Let *X* be the number of problems on pset2 that a randomly selected student has completed, as of Monday morning.

X takes on values with uncertainty, so *X* is a random variable.

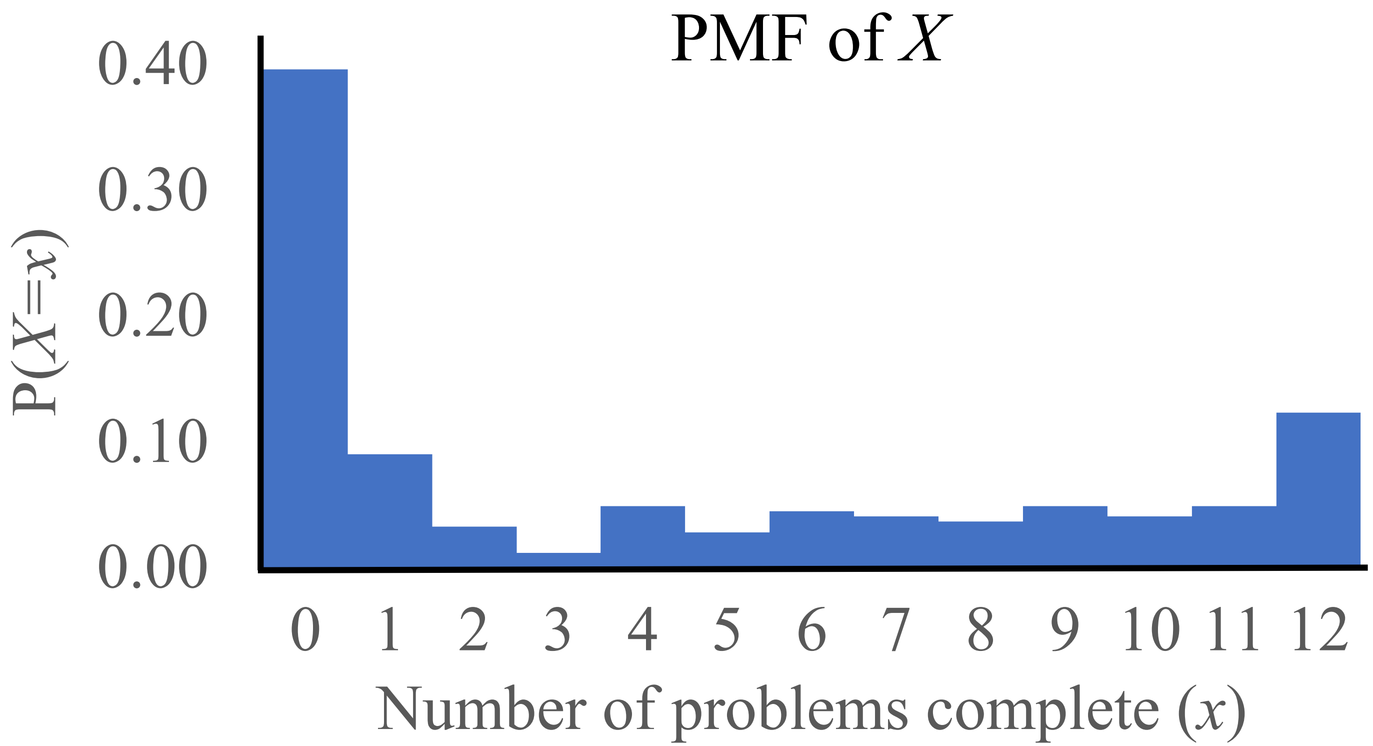


$$E[X] = 6$$

Expectation Is Not All You Need

Let X be the number of problems on pset2 that a randomly selected student has completed, as of Monday morning.

X takes on values with uncertainty, so X is a random variable.



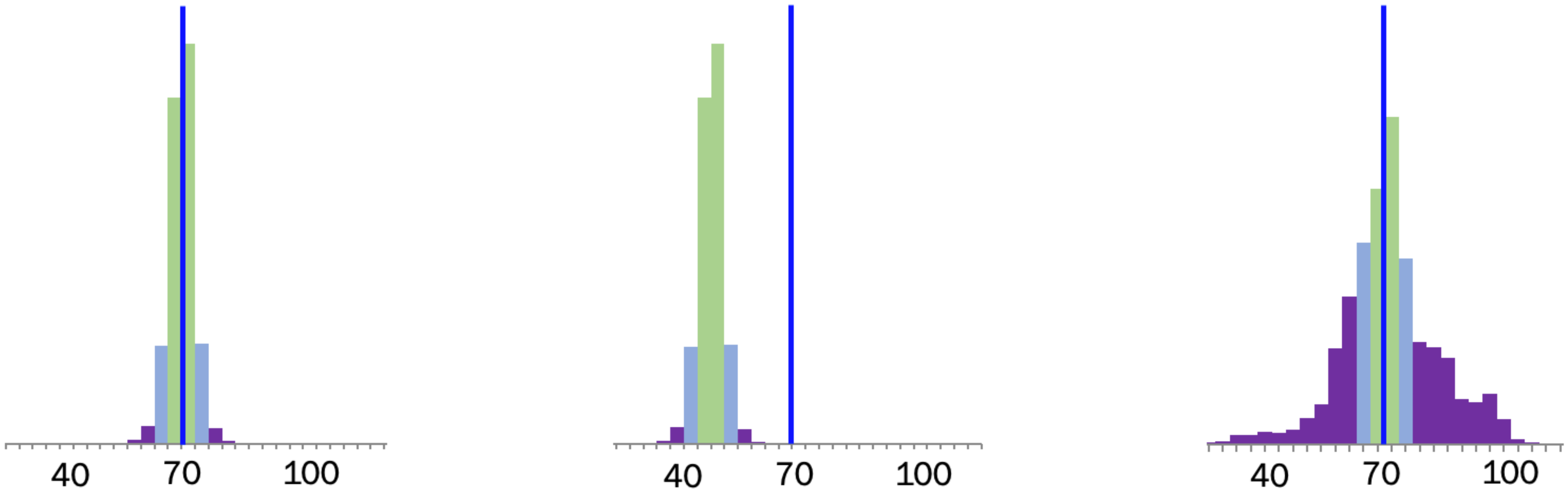
$$E[X] = 6$$

Does this expected value capture all the information in the data?
No!

Can we invent *another* summary number?

A Second Summary Statistic

Consider the following 3 distributions (PMFs):



How are they different from one another?

Variance

Variance is a formal definition of the **spread** of a random variable.

If X is a random variable with mean $\mu = \mathrm{E}[X]$, then the **variance** of X, denoted Var(X), is:

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

Variance

Variance is a formal definition of the **spread** of a random variable.

If X is a random variable with mean $\mu = \mathrm{E}[X]$, then the **variance** of X, denoted Var(X), is:

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

distance

On average…

The random variable X

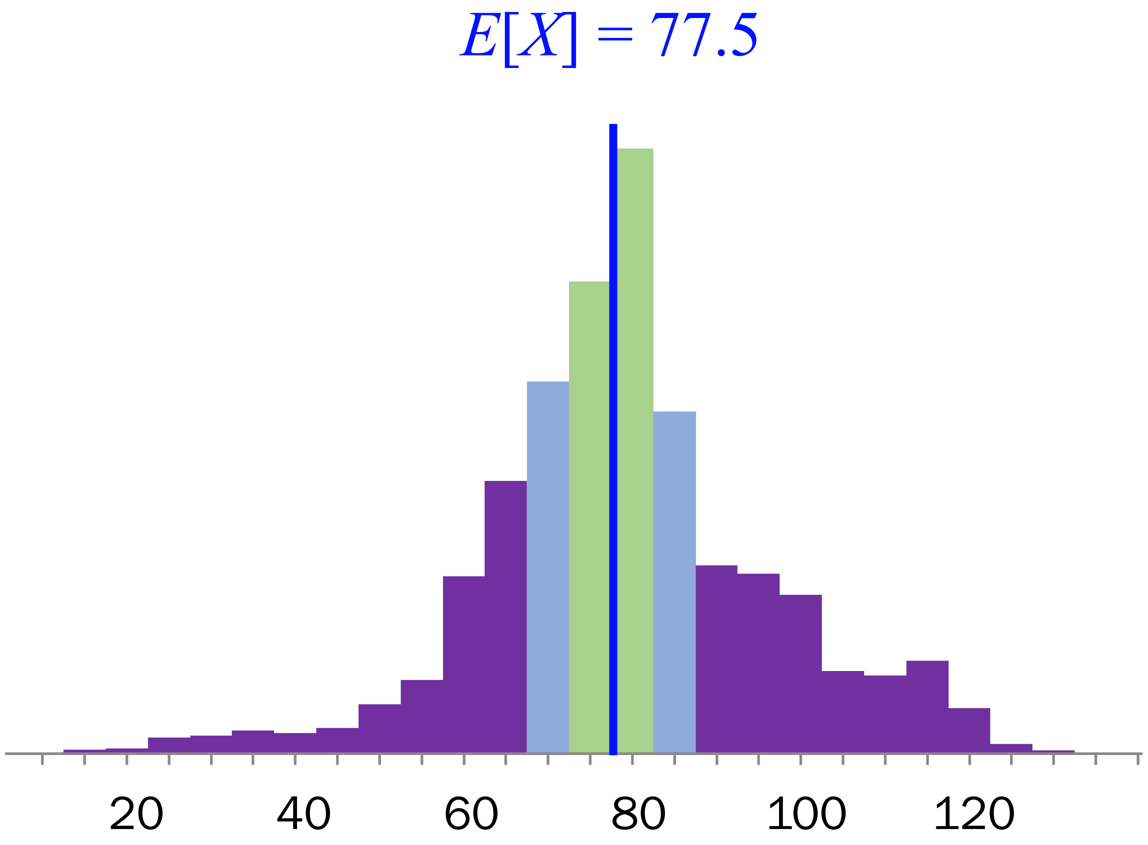
The mean of X

"How far away from the mean is X, on average?"

Variance Intuition

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

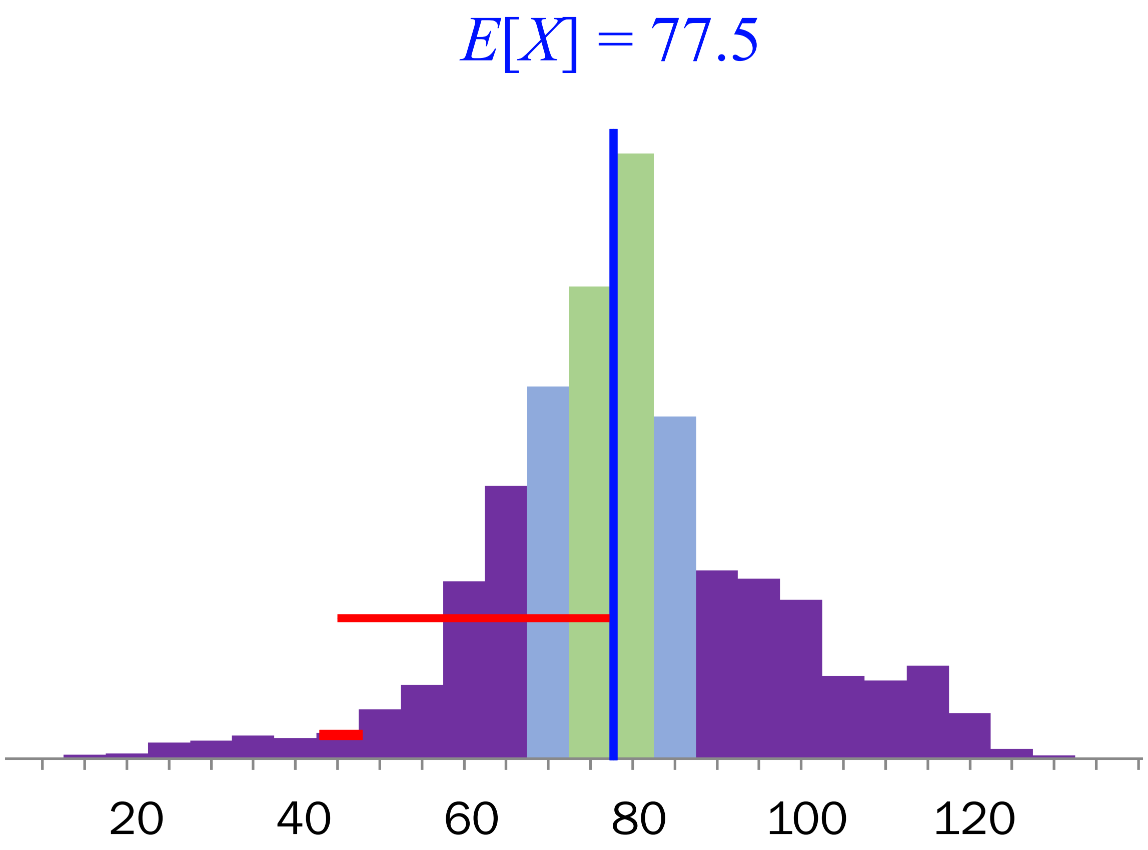
Let X be a random variable that represents a midterm exam grade.



Variance Intuition

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

Let X be a random variable that represents a midterm exam grade.



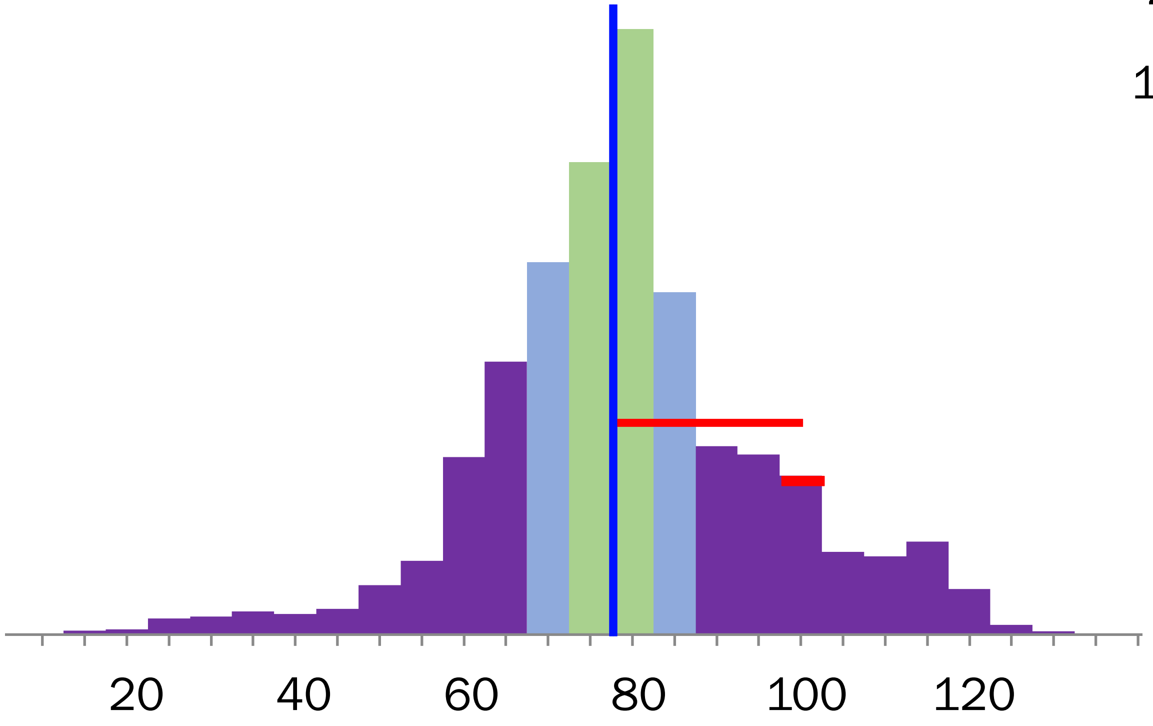
X	$(X - \mu)^2$
45 points	1056 points^2

Variance Intuition

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

Let *X* be a random variable that represents a midterm exam grade.

$E[X] = 77.5$

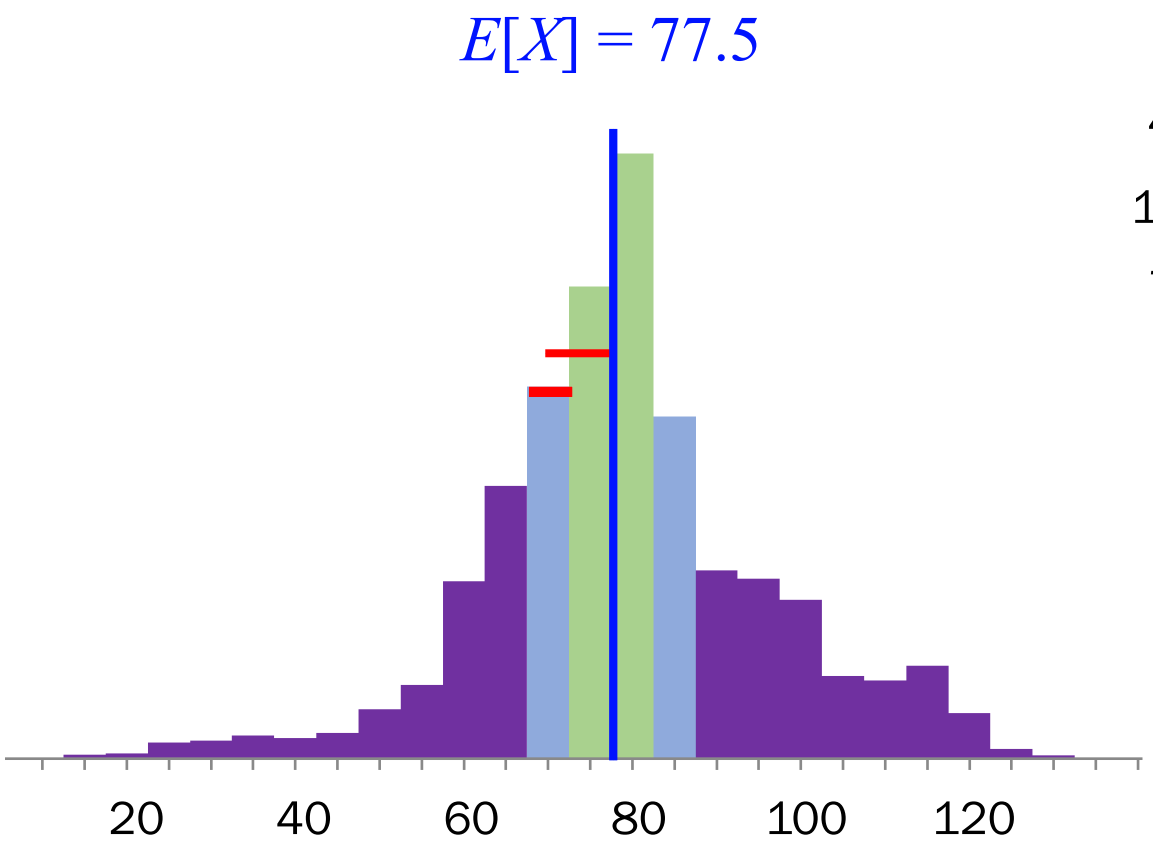


X	$(X-\mu)^2$
45 points	1056 points^2
100 points	506 points^2

Variance Intuition

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

Let X be a random variable that represents a midterm exam grade.



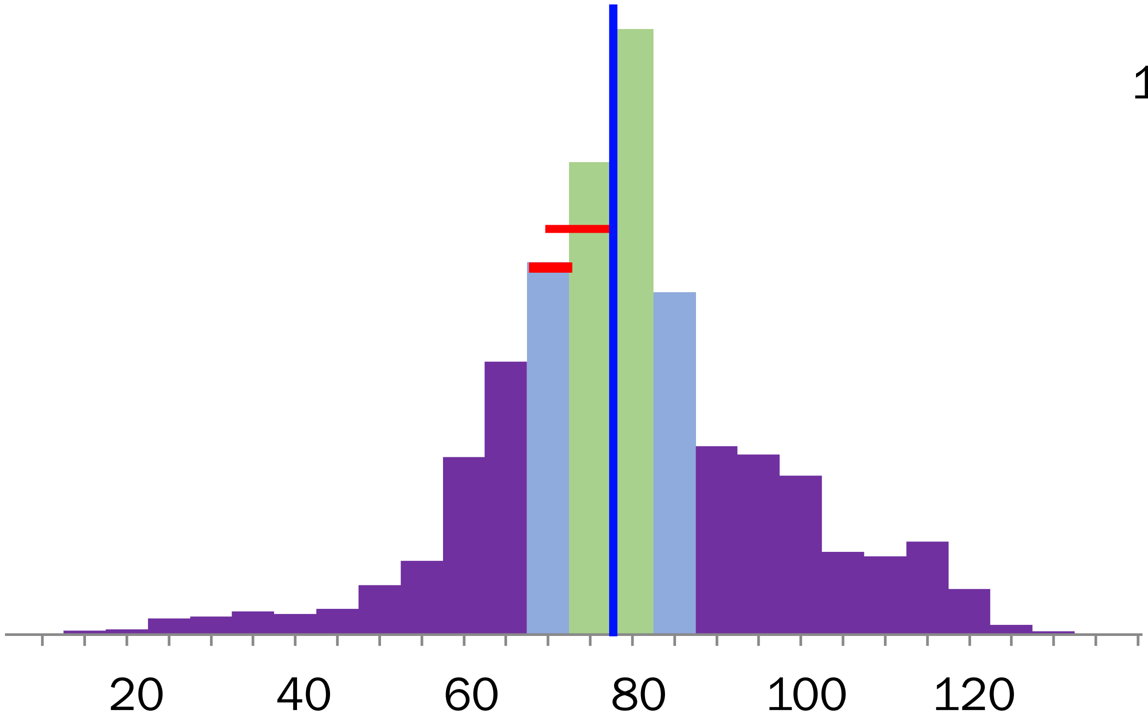
X	$(X - \mu)^2$
45 points	1056 points^2
100 points	506 points^2
70 points	56 points^2

Variance Intuition

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

Let X be a random variable that represents a midterm exam grade.

$E[X] = 77.5$



X	$(X-\mu)^2$
45 points	1056 points^2
100 points	506 points^2
70 points	56 points^2

…

$\text{Var}(X)$ = 52 points^2

Variance

Variance is a formal definition of the **spread** of a random variable.

If X is a random variable with mean $\mu = \mathrm{E}[X]$, then the **variance** of X, denoted Var(X), is:

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

In practice, it is usually easier to calculate this equivalent:

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

How to calculate $\mathrm{E}[X^2]$? Law of the Unconscious Statistician!

How To Get From E[(X – μ)²] to E[X²] - E[X]²

$$\begin{aligned}
\mathrm{Var}(X) &= E[(X-\mu)^2] \\
&= \sum_x (x-\mu)^2 p(x) \\
&= \sum_x (x^2 - 2\mu x + \mu^2) p(x) \\
&= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) \\
&= E[X^2] - 2\mu E[X] + \mu^2 \\
&= E[X^2] - 2\mu^2 + \mu^2 \\
&= E[X^2] - \mu^2 \\
&= E[X^2] - (E[X])^2
\end{aligned}$$

Law of Unconscious Statistician

Notation:

$$p(x) = P(X = x)$$

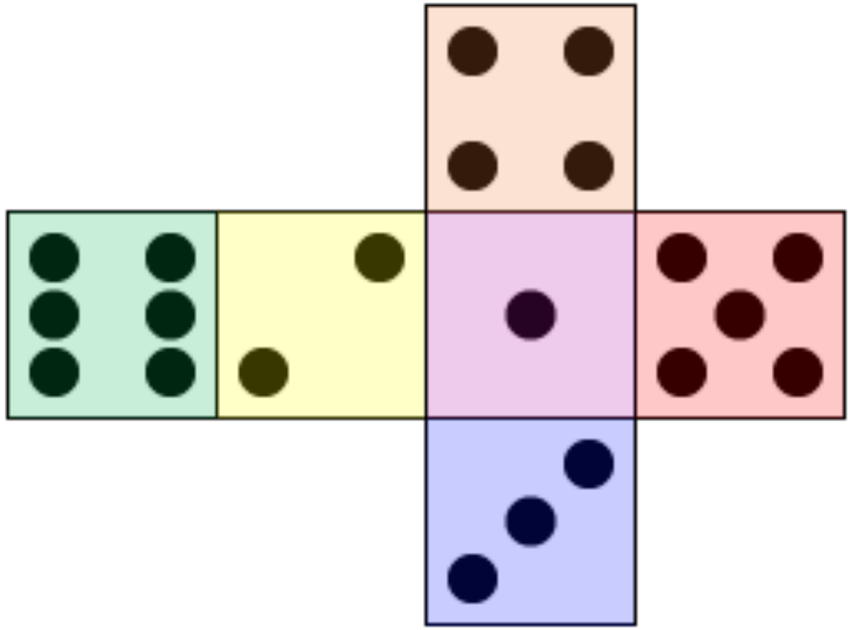
$$\mu = E[X]$$

Example: Variance of a Dice Roll

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

Let X be the result of rolling a 6 sided dice.

What is Var(X)?

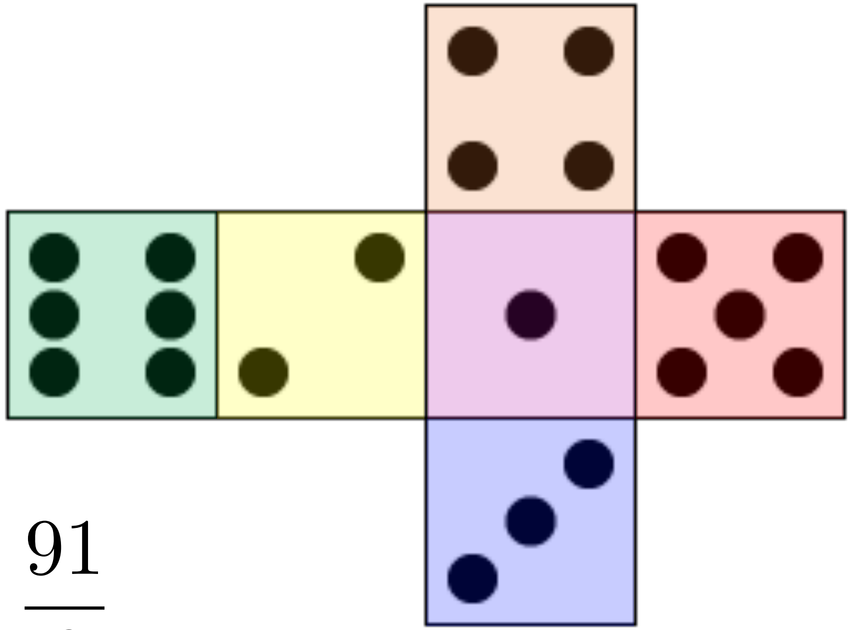


Example: Variance of a Dice Roll

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

Let X be the result of rolling a 6 sided dice.

What is Var(X)?



$$E[X] = 3.5$$

$$E[X^2] = 1^2\frac{1}{6} + 2^2\frac{1}{6} + 3^2\frac{1}{6} + 4^2\frac{1}{6} + 5^2\frac{1}{6} = \frac{91}{6}$$

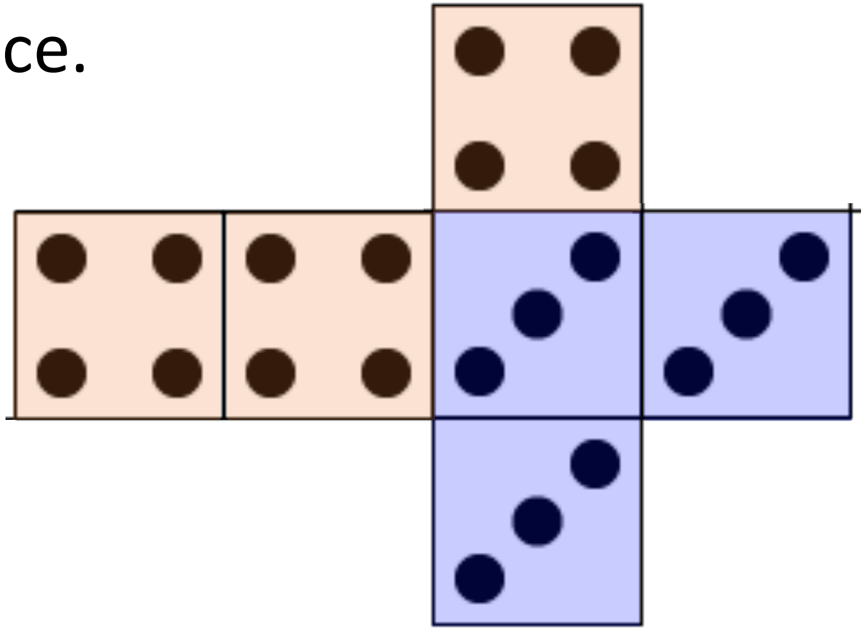
$$\begin{aligned}\text{Var}(X) &= E[X^2] - E[X]^2 \\ &= \frac{91}{6} - (3.5)^2 = 2.91\end{aligned}$$

Example: Variance of a Dice Roll

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

Let X be the result of rolling **this weird** 6 sided dice.

What is Var(X)?

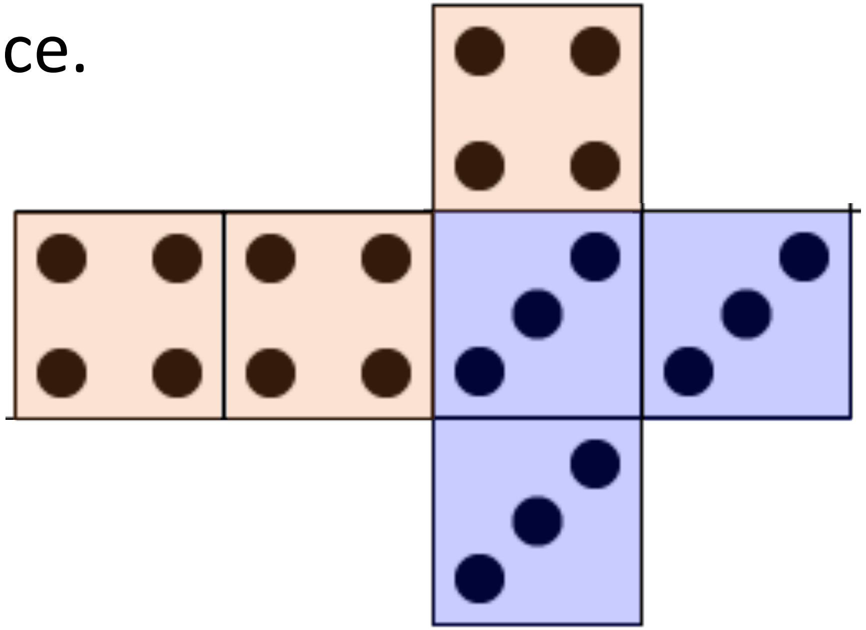


Example: Variance of a Dice Roll

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

Let X be the result of rolling **this weird** 6 sided dice.

What is Var(X)?



$$E[X] = 3.5$$

$$E[X^2] = 3^2 \cdot \frac{3}{6} + 4^2 \cdot \frac{3}{6} = 12.5$$

$$\begin{aligned}\mathrm{Var}(X) &= E[X^2] - E[X]^2 \\ &= 12.5 - (3.5)^2 = 0.25\end{aligned}$$

What About Standard Deviation?

$$\mathrm{Std}(X) = \sqrt{\mathrm{Var}(X)}$$

Units are the same as your random variable

Units are squared

Variance of Classic Random Variables

$X \sim \text{Geo}(p)$

$$Var(X) = \frac{1-p}{p^2}$$

$X \sim \text{Bern}(p)$

$$Var(X) = p(1-p)$$

$Y \sim \text{NegBin}(r, p)$

$$Var(X) = \frac{r \cdot (1-p)}{p^2}$$

$Y \sim \text{Bin}(n, p)$

$$Var(Y) = n \cdot p(1-p)$$

Random Variables: You Get Even More For Free!

Binomial Random Variable

Notation: $X \sim \text{Bin}(n, p)$

Description: Number of "successes" in n identical, independent experiments each with probability of success p.

Parameters: $n \in \{0, 1, \ldots\}$, the number of experiments.
$p \in [0, 1]$, the probability that a single experiment gives a "success".

Support: $x \in \{0, 1, \ldots, n\}$

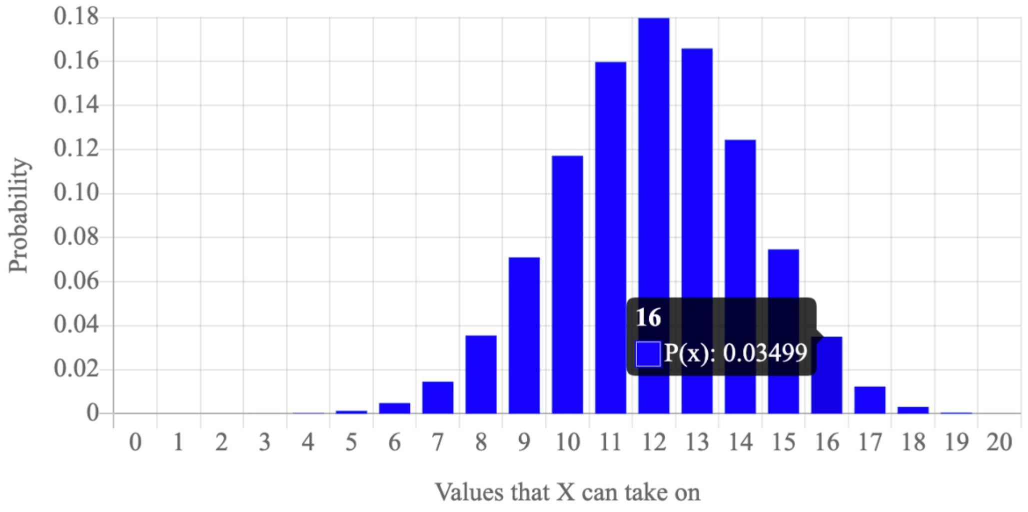
PMF equation: $\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

Expectation: $\text{E}[X] = n \cdot p$

Variance: $\text{Var}(X) = n \cdot p \cdot (1-p)$

PMF graph:

Parameter n: 20 Parameter p: 0.60



Bernoulli Random Variable

Notation: $X \sim \text{Bern}(p)$

Description: A boolean variable that is 1 with probability p

Parameters: p, the probability that $X = 1$.

Support: x is either 0 or 1

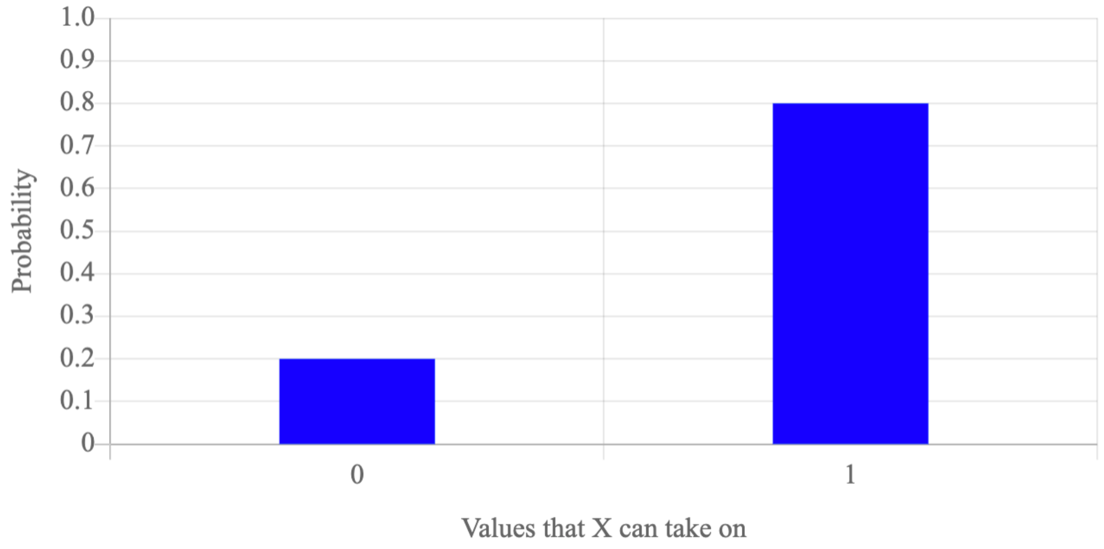
PMF equation: $\Pr(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$

Expectation: $\text{E}[X] = p$

Variance: $\text{Var}(X) = p(1-p)$

PMF graph:

Parameter p: 0.80



Pokémon™

Gotta catch 'em all!™

(The Last Discrete Random Variable)

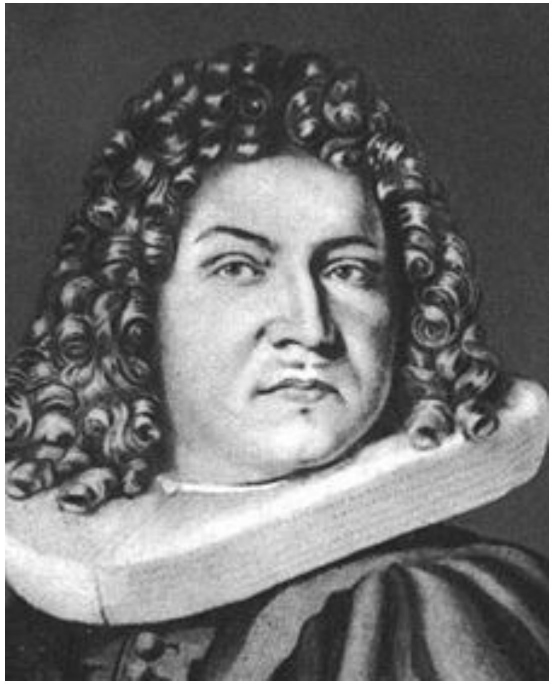
Ready?

It’s Time
To Talk About Time

Random Fun Fact: *e*

How the "natural exponent" e is defined:

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$$

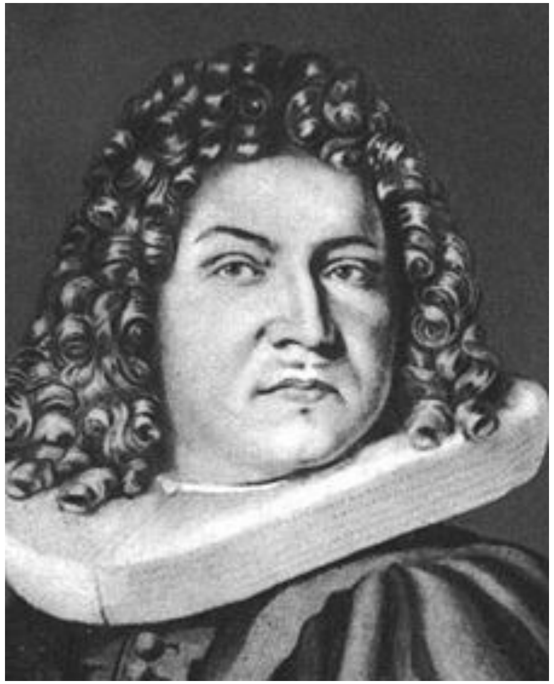


Also invented by
Jacob Bernoulli!

Random Fun Fact: *e*

How the “natural exponent” e is defined:

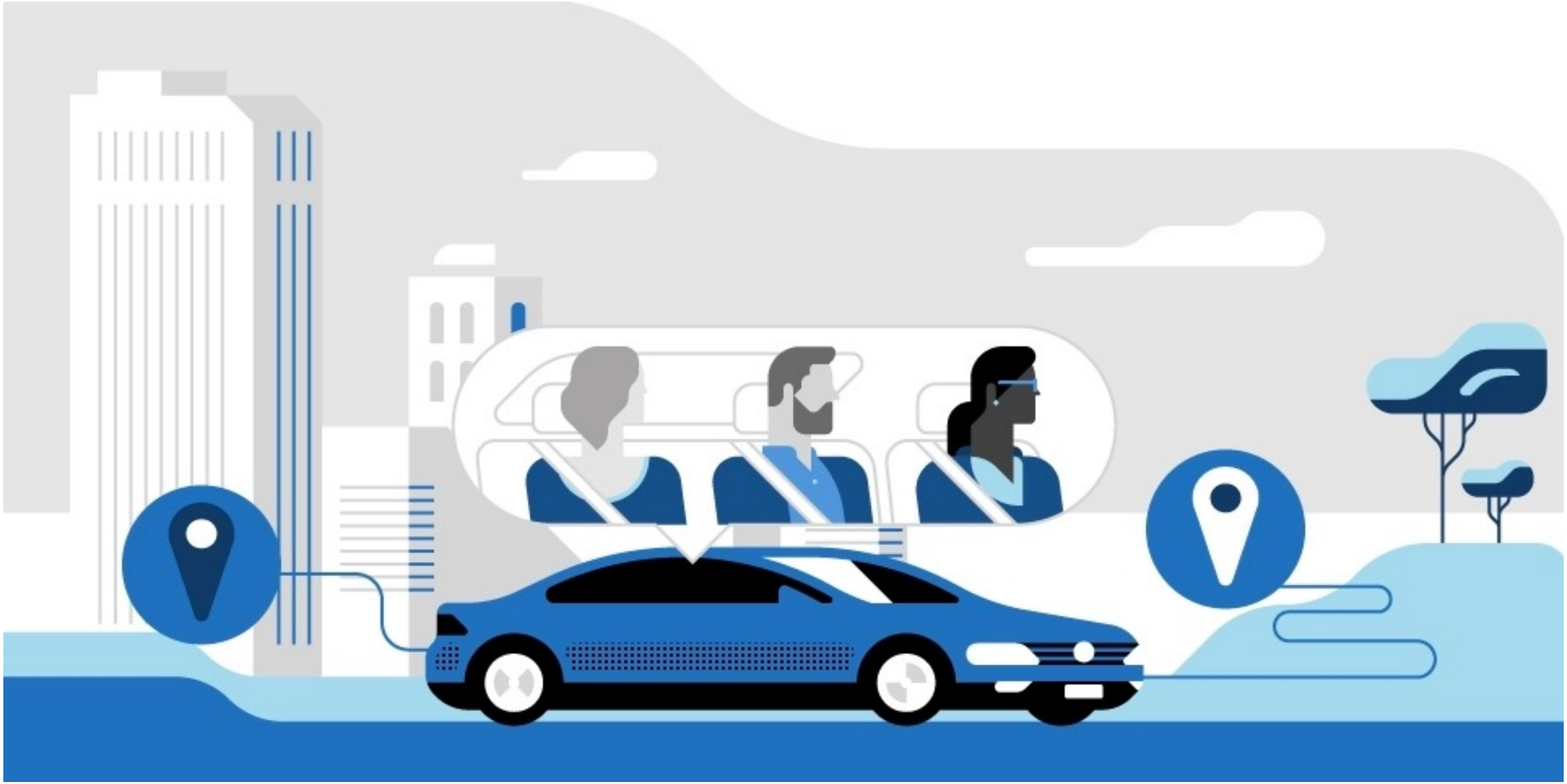
$$\lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$



Also invented by Jacob Bernoulli!

https://github.com/eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee/eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee

Case Study: Ride Sharing Apps



Probability of *k* **Requests** From This Area Each Minute

Probability of ***k* Requests** From This Area Each Minute

On average, λ = 5 requests per minute

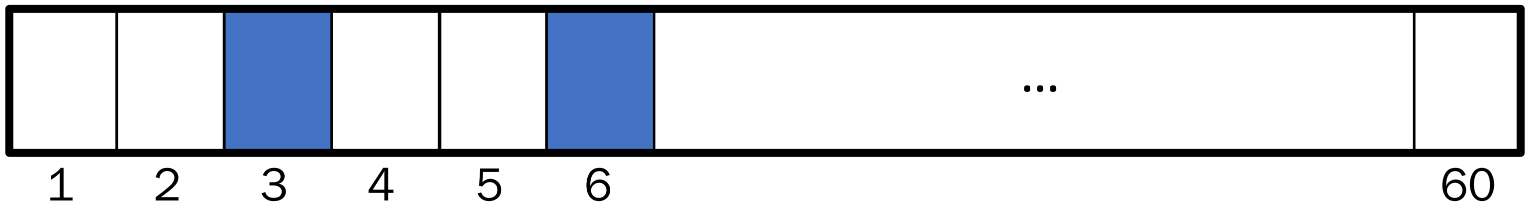
Probability of ***k* Requests** From This Area Each Minute

Idea: we can break a minute down into 60 seconds…

On average, λ = 5 requests per minute

Probability of *k* Requests From This Area Each Minute

Idea: we can break a minute down into 60 seconds…

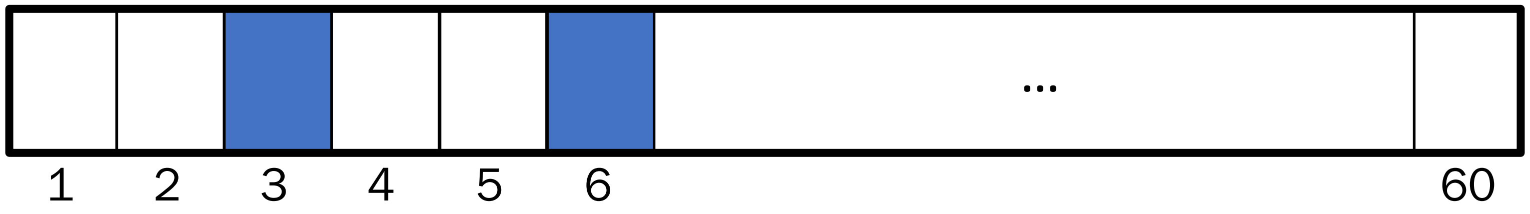


At each second, you either get a request or don't.

On average, λ = 5 requests per minute

Probability of ***k* Requests** From This Area Each Minute

Idea: we can break a minute down into 60 seconds…



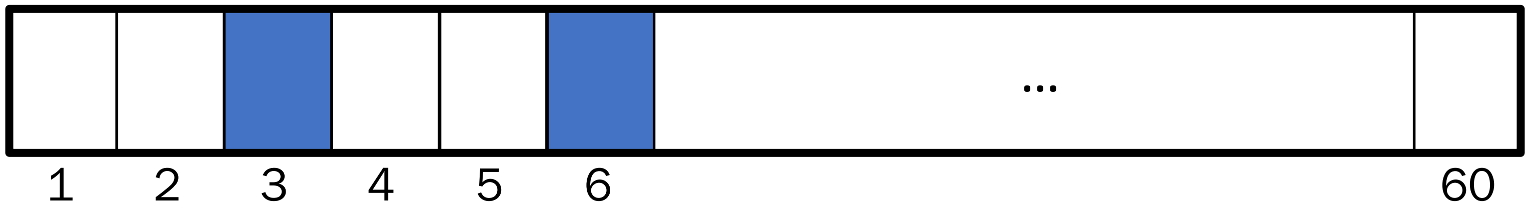
At each second, you either get a request or don't.
Let *X* be the number of requests in a minute.

On average, $\lambda = 5$ requests per minute

$$X \sim \text{Bin}(n = 60, p = \ ? \ )$$

Probability of ***k* Requests** From This Area Each Minute

Idea: we can break a minute down into 60 seconds…



At each second, you either get a request or don't.
Let *X* be the number of requests in a minute.

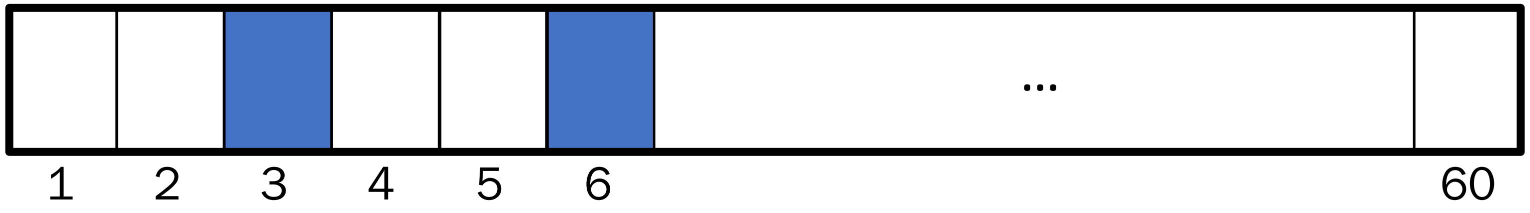
On average, λ = 5 requests per minute

$$X \sim \mathrm{Bin}(n = 60, p = 5/60)$$

$$p = \frac{\lambda}{n}$$

Probability of ***k* Requests** From This Area Each Minute

Idea: we can break a minute down into 60 seconds…



At each second, you either get a request or don't.
Let *X* be the number of requests in a minute.

On average, λ = 5 requests per minute

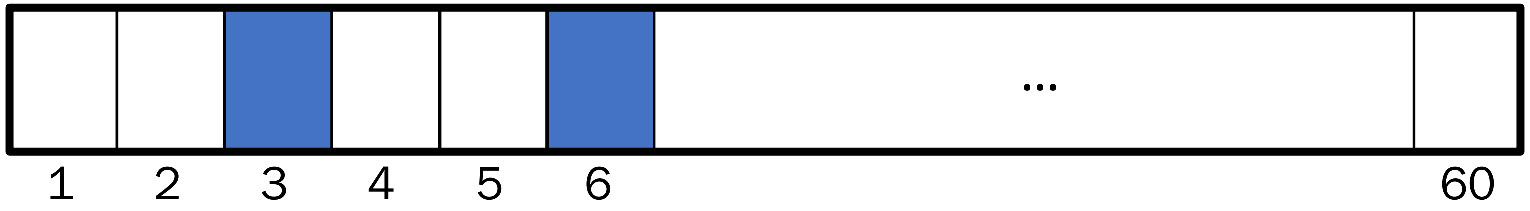
$$X \sim \text{Bin}(n = 60, p = 5/60)$$

$$p = \frac{\lambda}{n}$$

$$P(X = 3) = \binom{60}{3}(5/60)^3(1 - 5/60)^{57}$$

Probability of ***k* Requests** From This Area Each Minute

Idea: we can break a minute down into 60 seconds…



At each second, you either get a request or don't.
Let *X* be the number of requests in a minute.

On average, λ = 5 requests per minute

$$X \sim \text{Bin}(n = 60, p = 5/60)$$

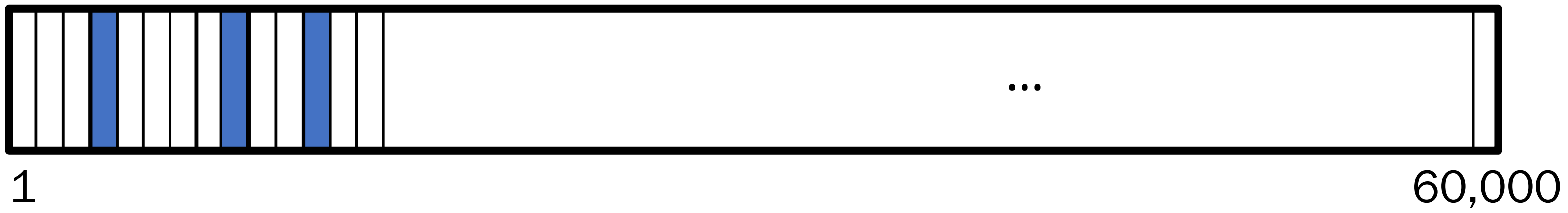
$$p = \frac{\lambda}{n}$$

$$P(X = 3) = \binom{60}{3}(5/60)^3(1 - 5/60)^{57}$$

But what if there are two requests in the same second?

Probability of *k* **Requests** From This Area Each Minute

Idea: we can break a minute down into 60,000 milliseconds…



At each ms, you either get a request or don't.
Let X be the number of requests in a minute.

On average, $\lambda = 5$ requests per minute

Probability of ***k* Requests** From This Area Each Minute

Idea: we can break a minute down into 60,000 milliseconds…

At each ms, you either get a request or don't.
Let *X* be the number of requests in a minute.

On average, λ = 5 requests per minute

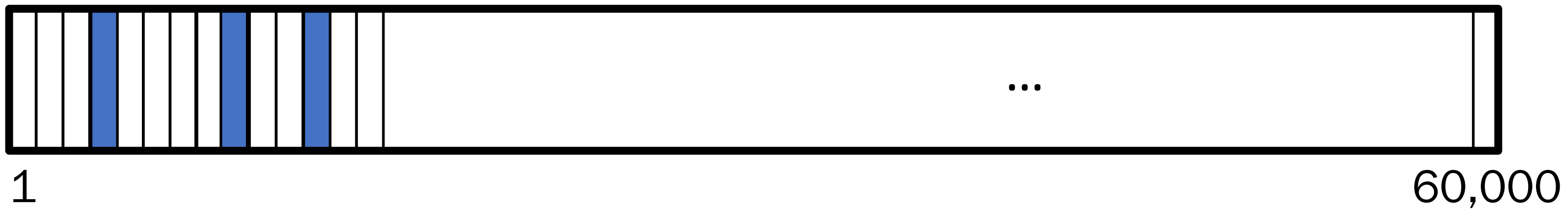
$$X \sim \text{Bin}(n = 60000, p = \lambda/n)$$

$$p = \frac{\lambda}{n}$$

$$P(X = k) = \binom{n}{k} (\lambda/n)^k (1 - \lambda/n)^{n-k}$$

Probability of ***k* Requests** From This Area Each Minute

Idea: we can break a minute down into 60,000 milliseconds…



At each ms, you either get a request or don't.
Let *X* be the number of requests in a minute.

On average, λ = 5 requests per minute

$$X \sim \text{Bin}(n = 60000, p = \lambda/n)$$

$$p = \frac{\lambda}{n}$$

$$P(X = k) = \binom{n}{k} (\lambda/n)^k (1 - \lambda/n)^{n-k}$$

Can we do even better?

Probability of ***k* Requests** From This Area Each Minute

Idea: we can break a minute down into *infinitely small* buckets

too small to draw ☹

1 ∞

On average, λ = 5 requests per minute

In each bucket, you either get a request or don't.
Let *X* be the number of requests in a minute.

$$p = \frac{\lambda}{n}$$

$$X \sim \text{Bin}(n = \infty, p = \lambda/n)$$

$$P(X = k) = \binom{n}{k}(\lambda/n)^k(1 - \lambda/n)^{n-k}$$

Is this impossible to work with? No?! Time for cool math!

Probability of ***k* Requests** From This Area Each Minute

$$P(X=k) = \lim_{n\to\infty} \binom{n}{k} (\lambda/n)^k (1-\lambda/n)^{n-k}$$

$$= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} \cdot \frac{\lambda^k}{n^k} \cdot \frac{(1-\lambda/n)^n}{(1-\lambda/n)^k} \quad \text{By expanding each term}$$

$$= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} \cdot \frac{\lambda^k}{n^k} \cdot \frac{e^{-\lambda}}{1} \quad \text{By definition of natural exp}$$

$$= \lim_{n\to\infty} \frac{n!}{(n-k)!n^k} \cdot \frac{\lambda^k}{k!} \cdot \frac{e^{-\lambda}}{1} \quad \text{Rearranging terms}$$

$$= \lim_{n\to\infty} \frac{n^k}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \frac{e^{-\lambda}}{1} \quad \text{Limit analysis}$$

$$= \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{Simplifying}$$

The Poisson Random Variable

A **Poisson** random variable models the number of occurrences that happen in a *fixed* interval of time.

$$X \sim \text{Poi}(\lambda)$$

PMF:

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

X takes on values 0, 1, 2...up to infinity.

Simeon-Denis Poisson

Prolific French mathematician (1781-1840)

He published his first paper at 18?

Became a professor at 21???

And published over 300 papers in his life?????

He reportedly said, *"Life is good for only two things: discovering mathematics and teaching mathematics."*

Simeon-Denis Poisson

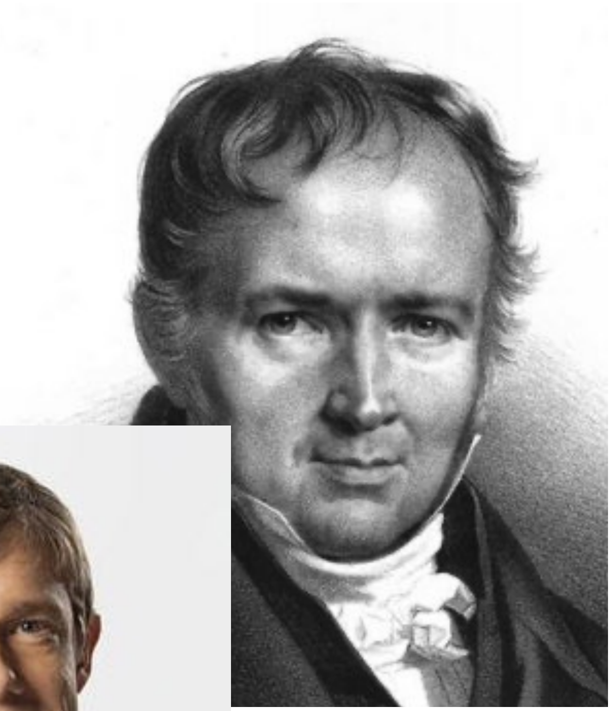
Prolific French mathematician (1781-1840)

He published his first paper at 18?

Became a professor at 21???

And published over 300 papers in his life?????

He reportedly said, *"Life is good for only two things: discovering mathematics and teaching mathematics."*



Looks like Martin Freeman, but…Frenchier

Problem Solving with The Poisson

Say you want to model events occurring over a given time interval.

- Earthquakes, radioactive decay, queries to a web server, etc.

Problem Solving with The Poisson

Say you want to model events occurring over a given time interval.

- Earthquakes, radioactive decay, queries to a web server, etc.

The events you're modeling must follow a **Poisson Process**:

1. Events happen *independently* of one another
2. Events arrive at a *fixed* rate: λ events per interval of time

Problem Solving with The Poisson

Say you want to model events occurring over a given time interval.

- Earthquakes, radioactive decay, queries to a web server, etc.

The events you're modeling must follow a **Poisson Process**:

1. Events happen *independently* of one another
2. Events arrive at a *fixed* rate: λ events per interval of time

If those conditions are met:

Let X be the number of events that happen in the time interval.

$$X \sim \text{Poi}(\lambda)$$

Is Lambda All You Need? Yes

Let X be the number of Uber requests from Times Square each minute.

$$X \sim \text{Poi}(\lambda = 5)$$

What is $\text{E}[X]$?

Hint: what is the definition of λ?

Poisson Random Variable

Notation: $X \sim \text{Poi}(\lambda)$

Description: Number of events in a fixed time frame if (a) the events occur with a constant mean rate and (b) they occur independently of time since last event.

Parameters: $\lambda \in \{0, 1, \ldots\}$, the constant average rate.

Support: $x \in \{0, 1, \ldots\}$

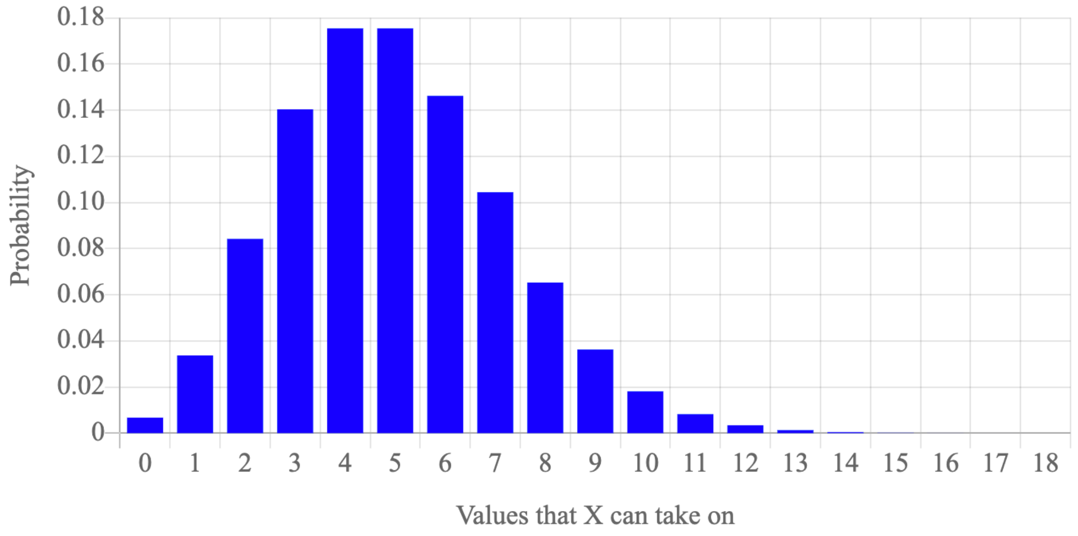
PMF equation: $\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$

Expectation: ?

Variance:

PMF graph:

Parameter λ: 5



Is Lambda All You Need? Yes

Let X be the number of Uber requests from Times Square each minute.

$$X \sim \text{Poi}(\lambda = 5)$$

What is $E[X]$?

$$E[X] = \lambda = \text{Var}(X)$$

The parameter λ is sufficient to fully define the whole Poisson distribution.

Poisson Random Variable

Notation: $X \sim \text{Poi}(\lambda)$

Description: Number of events in a fixed time frame if (a) the events occur with a constant mean rate and (b) they occur independently of time since last event.

Parameters: $\lambda \in \{0, 1, \ldots\}$, the constant average rate.

Support: $x \in \{0, 1, \ldots\}$

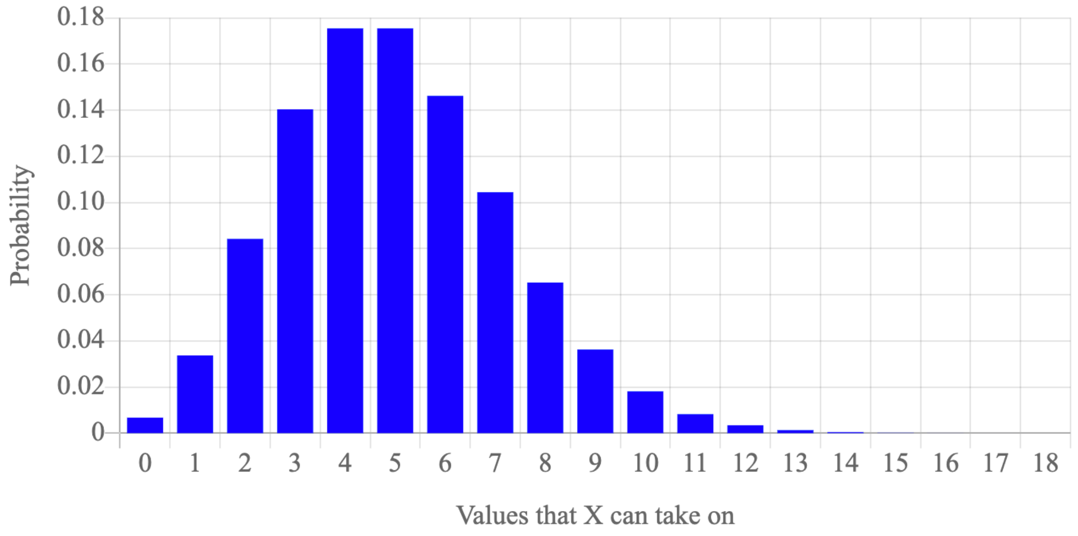
PMF equation: $\Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$

Expectation: $E[X] = \lambda$

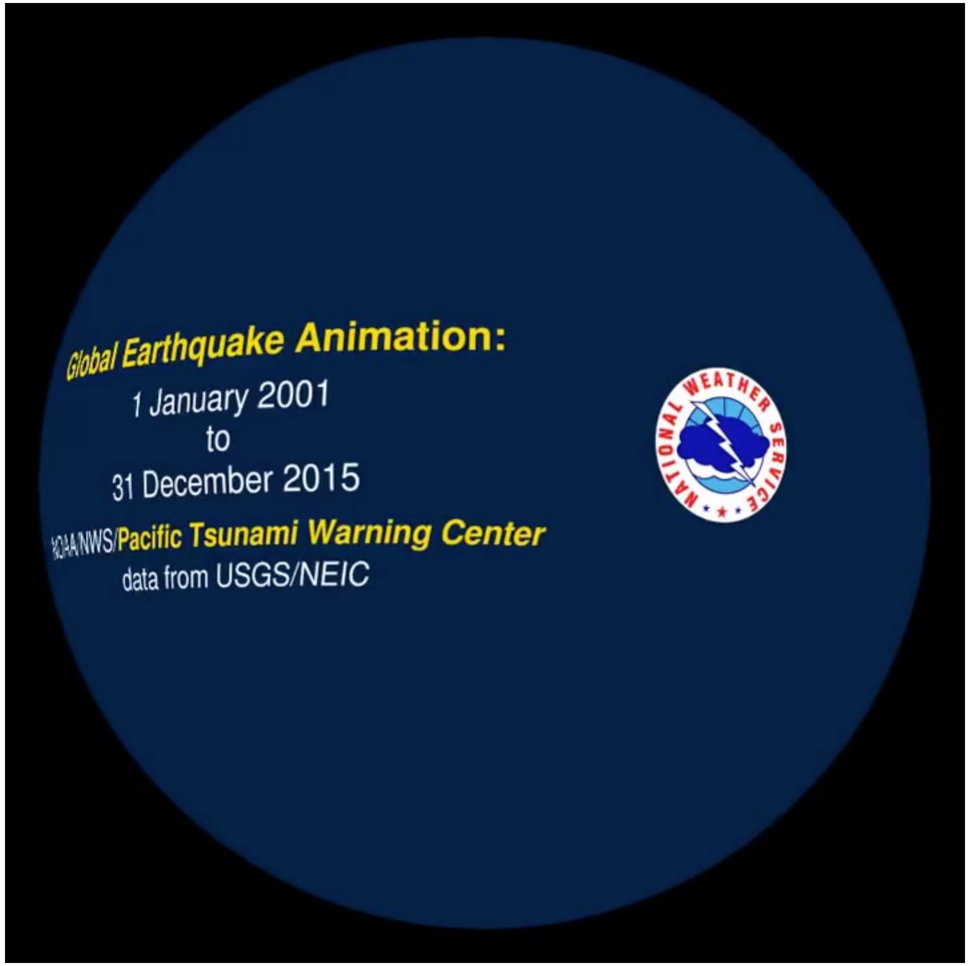
Variance: $\text{Var}(X) = \lambda$

PMF graph:

Parameter λ: 5



Example: Earthquakes



Bulletin of the
Seismological Society of America

Vol. 64 October 1974 No. 5

IS THE SEQUENCE OF EARTHQUAKES IN SOUTHERN CALIFORNIA, WITH AFTERSHOCKS REMOVED, POISSONIAN?

BY J. K. GARDNER and L. KNOPOFF

ABSTRACT

Yes.

You Now Know Where This PMF Comes From!

Let X be the number of earthquakes that happen in California every year.

Here's the PMF for X:

$$P(X = x) = \frac{69^x e^{-69}}{x!}$$

What is the probability that there are 60 earthquakes in California next year?

You Now Know Where This PMF Comes From!

Let X be the number of earthquakes that happen in California every year.

Here's the PMF for X:

$$P(X = x) = \frac{69^x e^{-69}}{x!}$$

X is a Poisson!

What is E[X] (λ)?

What is the probability that there are 60 earthquakes in California next year?

You Now Know Where This PMF Comes From!

Let X be the number of earthquakes that happen in California every year.

Here's the PMF for X: $$P(X = x) = \frac{69^x e^{-69}}{x!}$$

X is a Poisson!
What is E[X] (λ)?

What is the probability that there are 60 earthquakes in California next year?

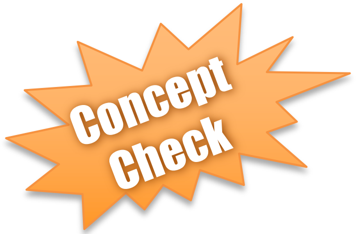
$$P(X = 60) = \frac{69^{60} e^{-69}}{60!} \approx 0.028$$

Just plug numbers into the PMF!

Practice: Web Server Load

Historically, a particular web server averages 120 requests each **minute**.

Let X be the number of hits this server receives in a **second**. What is $P(X < 5)$?



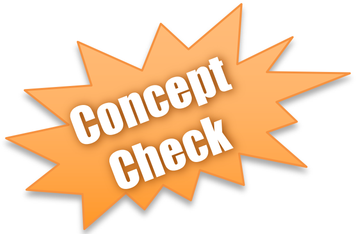
Practice: Web Server Load

Historically, a particular web server averages 120 requests each **minute**.

Let X be the number of hits this server receives in a **second**. What is $P(X < 5)$?

$$X \sim \mathrm{Poi}(\lambda = 2)$$

We have to use a value for λ that matches the time interval we want to model!



Practice: Web Server Load

Historically, a particular web server averages 120 requests each **minute**.

Let X be the number of hits this server receives in a **second**. What is $P(X < 5)$?

$$X \sim \text{Poi}(\lambda = 2)$$

$$P(X < 5) = \sum_{i=0}^{4} P(X = i)$$

$$= \sum_{i=0}^{4} e^{-\lambda} \frac{\lambda^i}{i!}$$

$$= \sum_{i=0}^{4} e^{-2} \frac{2^i}{i!} \approx 0.95$$

We have to use a value for λ that matches the time interval we want to model!

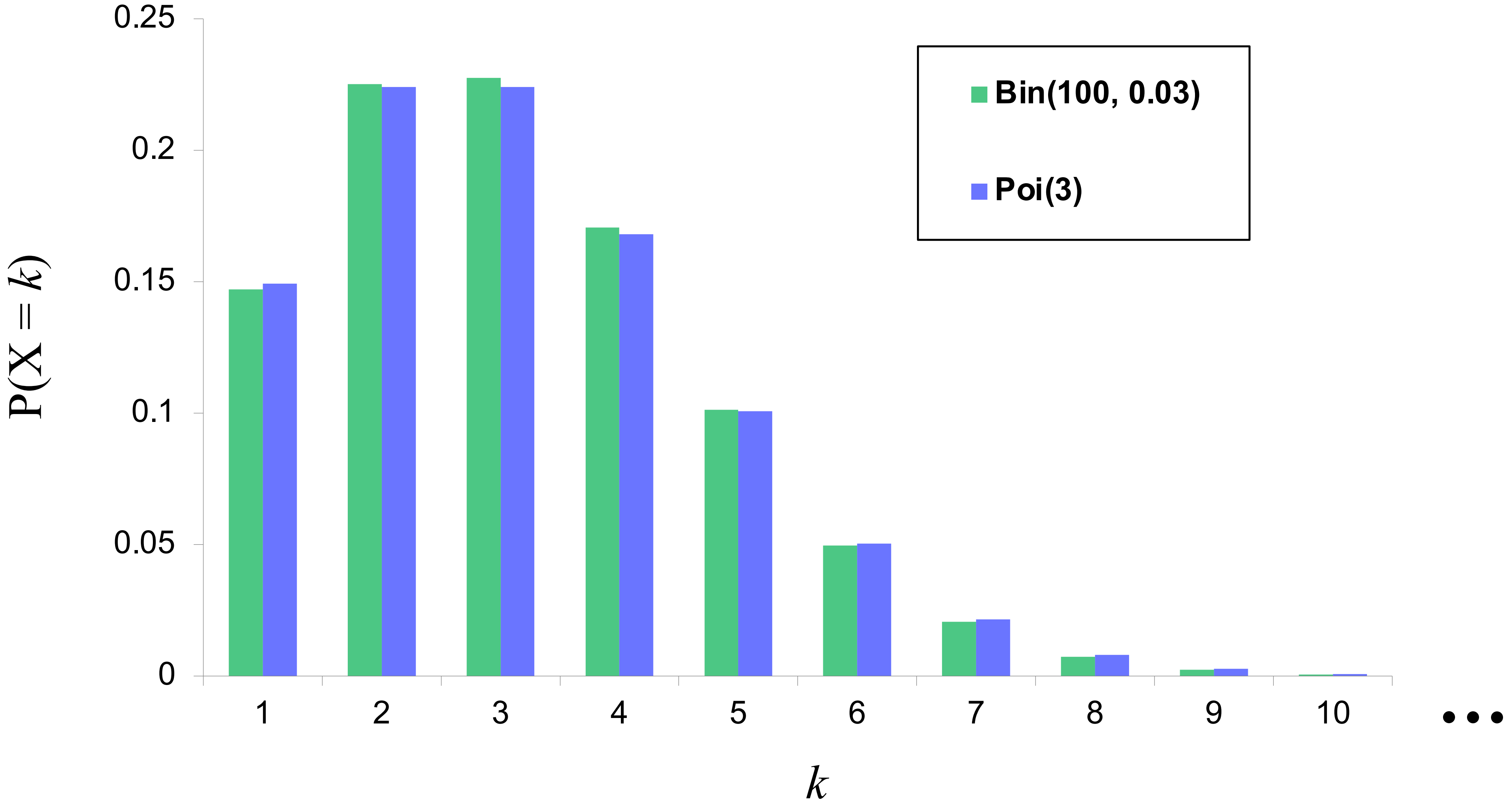
Concept Check

Another Fun Fact:

Another Fun Fact:

The Poisson can approximate The Binomial!

Why Can We Do This? Because The Shapes Are The Same

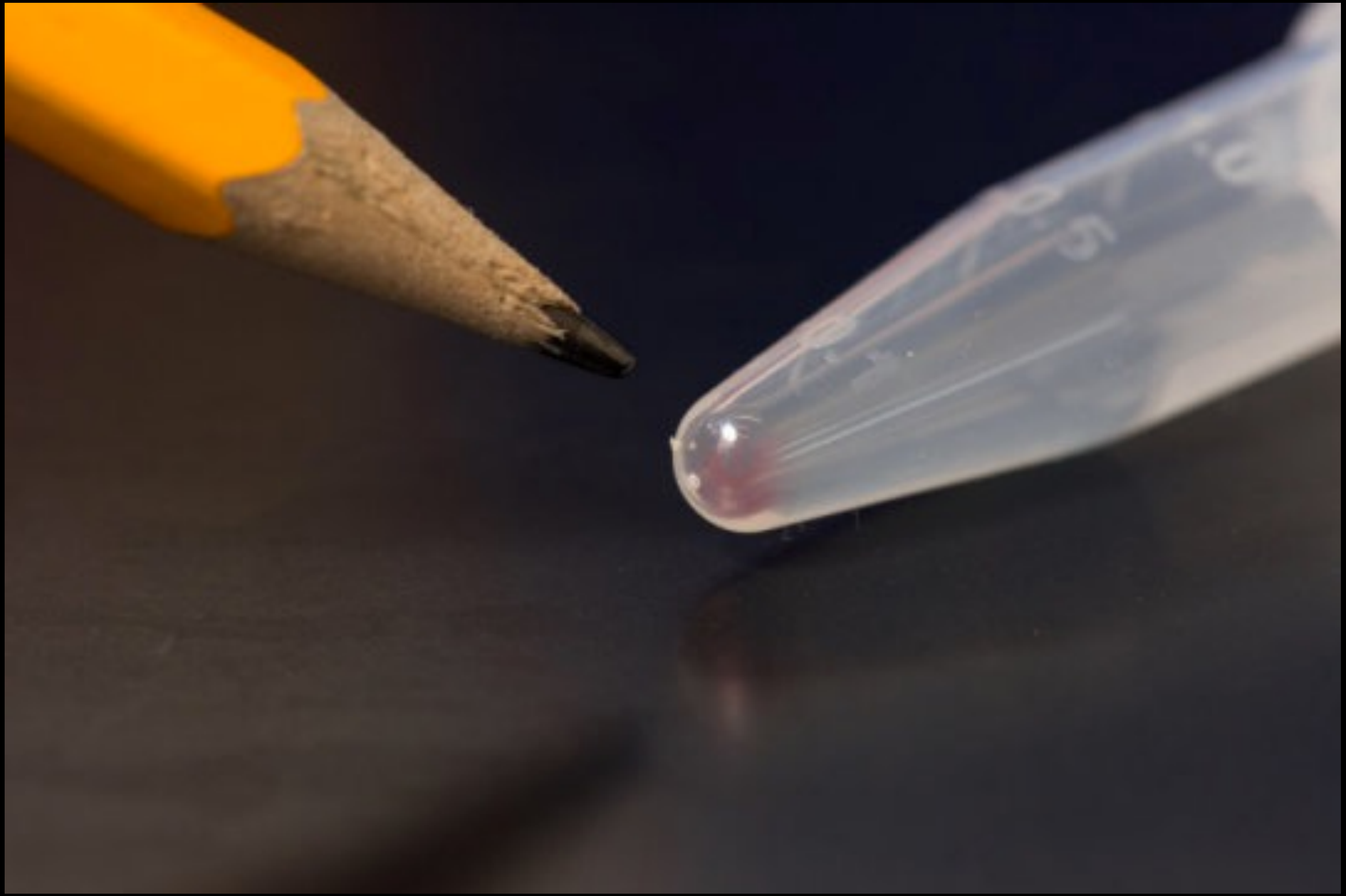


Another Fun Fact:

The Poisson can approximate The Binomial!

(Wait why would you want to do that?)

Storing Data in DNA: Super Promising Technology



The amount of data contained in ~ 600 smartphones (10,000 gigabytes) can be stored in just the faint pink smear of DNA at the end of this test tube.

Storing Data in DNA

Writing data to DNA is an imperfect process.

- Probability of corruption at each position (basepair) is very small: $p \approx 10^{-6}$.
- But we would want to store a LOT of data this way: say, $n \approx 10^{8}$ positions.

What's the probability that < 1% of DNA storage is corrupted?

Storing Data in DNA

Writing data to DNA is an imperfect process.

- Probability of corruption at each position (basepair) is very small: $p \approx 10^{-6}$.
- But we would want to store a LOT of data this way: say, $n \approx 10^{8}$ positions.

What's the probability that < 1% of DNA storage is corrupted?

Let X be the number of corrupted positions.

$$X \sim \text{Bin}(10^{8}, 10^{-6})$$

But the PMF for this would be unwieldy to compute :/

Storing Data in DNA

Writing data to DNA is an imperfect process.

- Probability of corruption at each position (basepair) is very small: $p \approx 10^{-6}$.
- But we would want to store a LOT of data this way: say, $n \approx 10^{8}$ positions.

What's the probability that < 1% of DNA storage is corrupted?

Let X be the number of corrupted positions.

$$X \sim \text{Bin}(10^{8}, 10^{-6})$$

But the PMF for this would be unwieldy to compute :/

There are lots of cases where extreme *n* and *p* values arise:

- Errors sending streams of bits over an imperfect network
- Server crashes per day in giant data center

Storing Data in DNA

Writing data to DNA is an imperfect process.

- Probability of corruption at each position (basepair) is very small: $p \approx 10^{-6}$.
- But we would want to store a LOT of data this way: say, $n \approx 10^{8}$ positions.

What's the probability that < 1% of DNA storage is corrupted?

Let X be the number of corrupted positions.

$$X \sim \text{Poi}(\lambda = 10^{8} * 10^{-6} = 100)$$

Storing Data in DNA

Writing data to DNA is an imperfect process.

- Probability of corruption at each position (basepair) is very small: $p \approx 10^{-6}$.
- But we would want to store a LOT of data this way: say, $n \approx 10^8$ positions.

What's the probability that < 1% of DNA storage is corrupted?

Let X be the number of corrupted positions.

$$X \sim \text{Poi}(\lambda = 10^8 * 10^{-6} = 100)$$

Where did we get λ from? E[X] for a binomial is $n * p$

Storing Data in DNA

Writing data to DNA is an imperfect process.

- Probability of corruption at each position (basepair) is very small: $p \approx 10^{-6}$.
- But we would want to store a LOT of data this way: say, $n \approx 10^8$ positions.

What's the probability that < 1% of DNA storage is corrupted?

Let X be the number of corrupted positions.

$$X \sim \mathrm{Poi}(\lambda = 10^8 * 10^{-6} = 100)$$

$$P(X < 0.01 \cdot 10^8) = P(X < 10^6) = \sum_{k=0}^{10^6-1} P(X = k) = \sum_{k=0}^{10^6-1} \frac{100^k \cdot e^{-100}}{k!}$$

Approximating Binomial With Poisson: General Rule

The Poisson approximates the Binomial well when:

1. n is large
2. p is small
3. Therefore, $\lambda = np$ is "moderate"

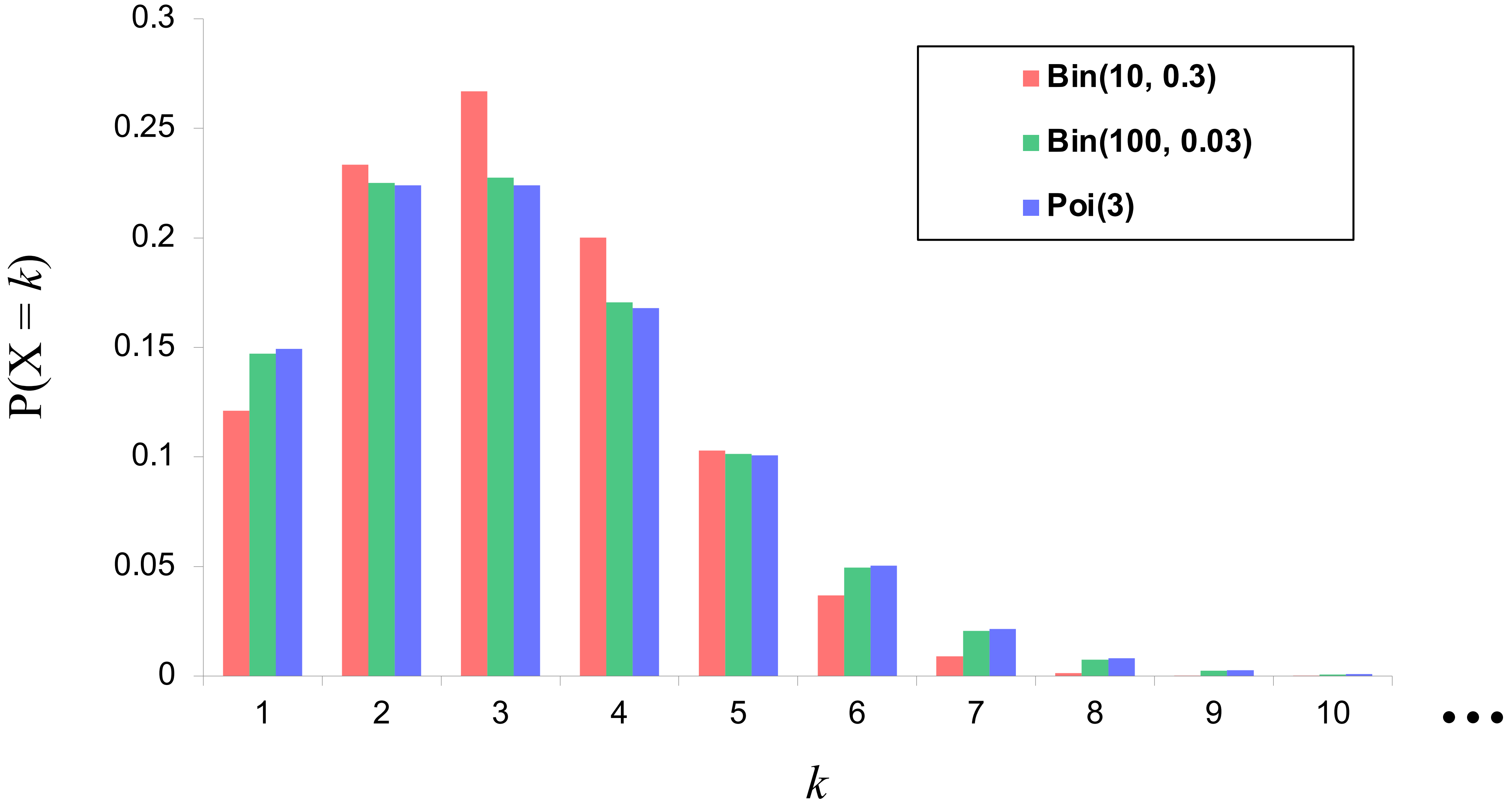
Different interpretations of "moderate":

$n > 20$ and $p < 0.05$

$n > 100$ and $p < 0.1$

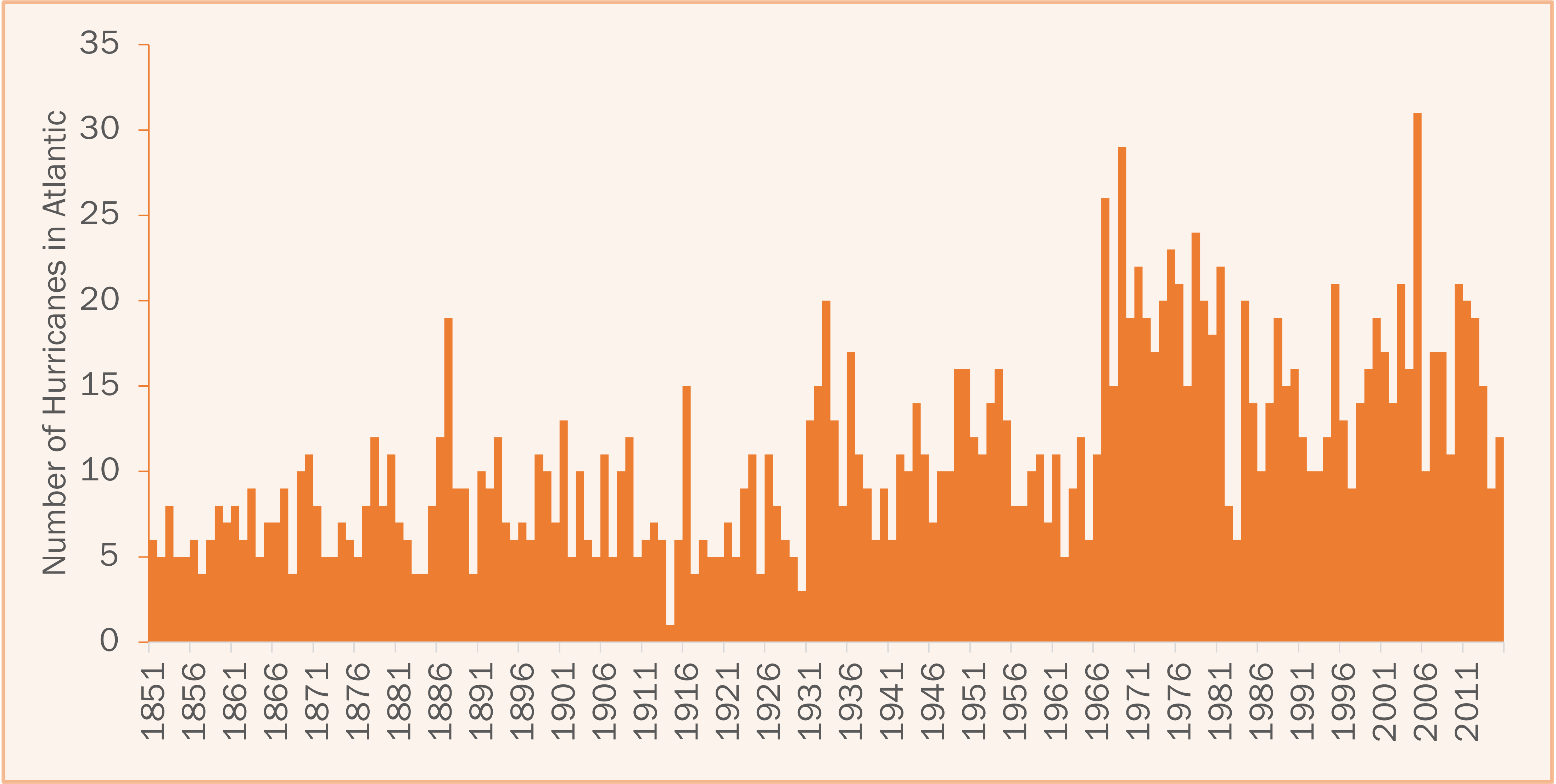
Really, Poisson is Binomial as $n \to \infty$ and $p \to 0$, where $np = \lambda$

How Similar Are The Shapes, With Different n and p?

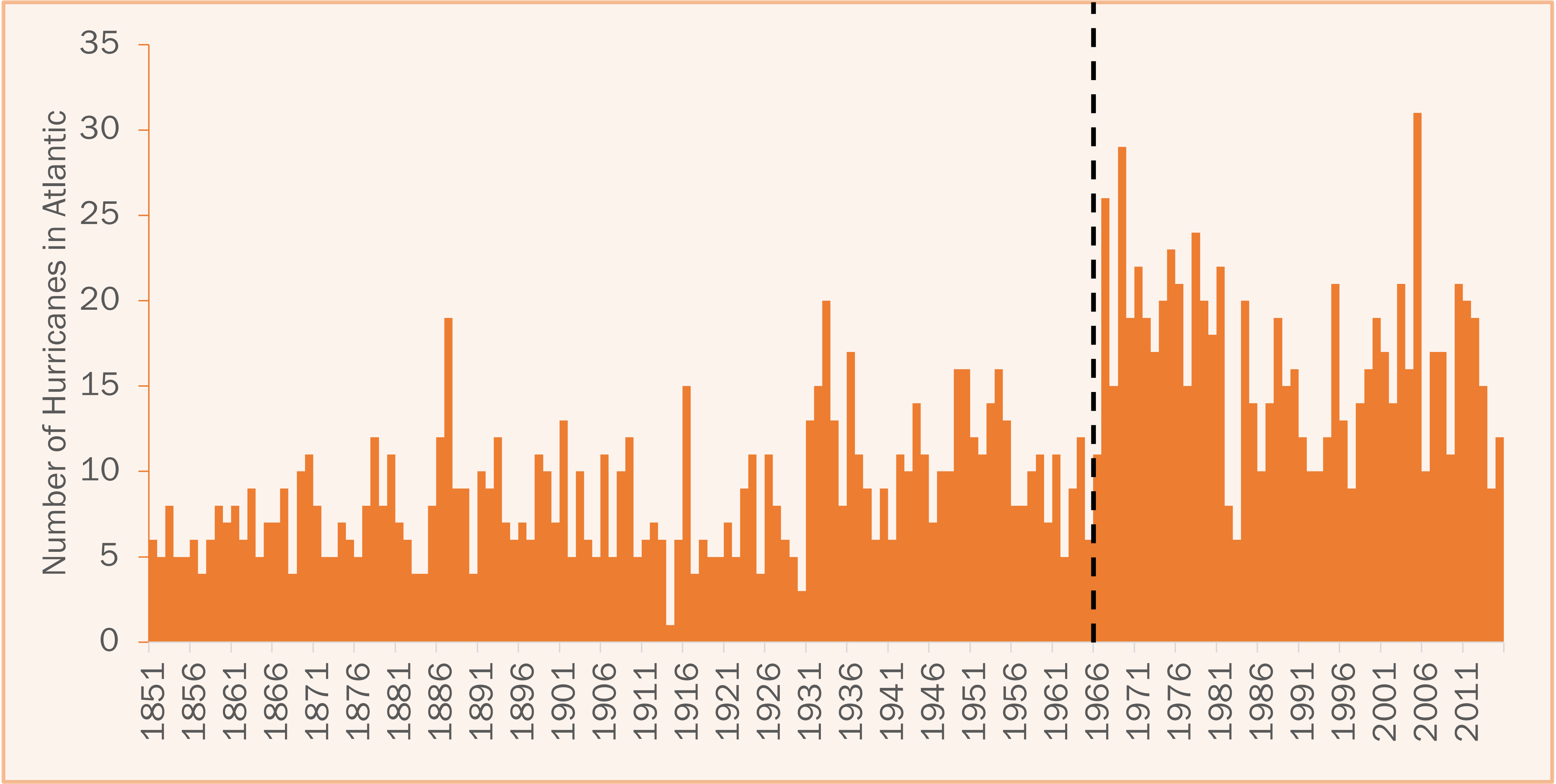


Probability of Extreme Weather?

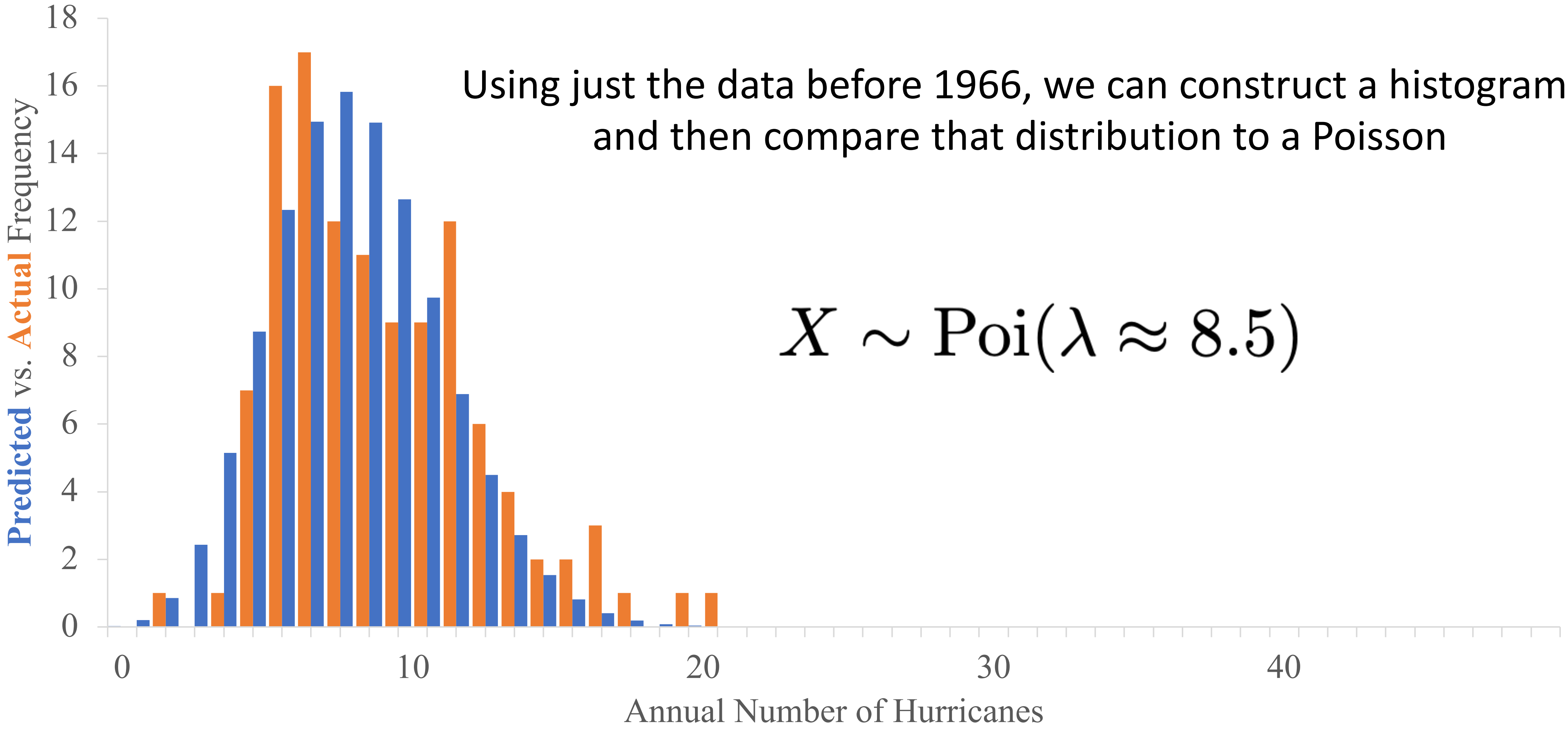
Let's Model Data: Hurricanes Per Year Since 1851



Let's Model Data: Hurricanes Per Year Since 1851



Let's Model Data: Observations vs. Poisson



Using just the data before 1966, we can construct a histogram, and then compare that distribution to a Poisson

$$X \sim \text{Poi}(\lambda \approx 8.5)$$

Let's Model Data As A Poisson

Based on our Poisson model from pre-1966 data, what is the probability of seeing more than 15 hurricanes in one year?

$$X \sim \text{Poi}(\lambda \approx 8.5)$$

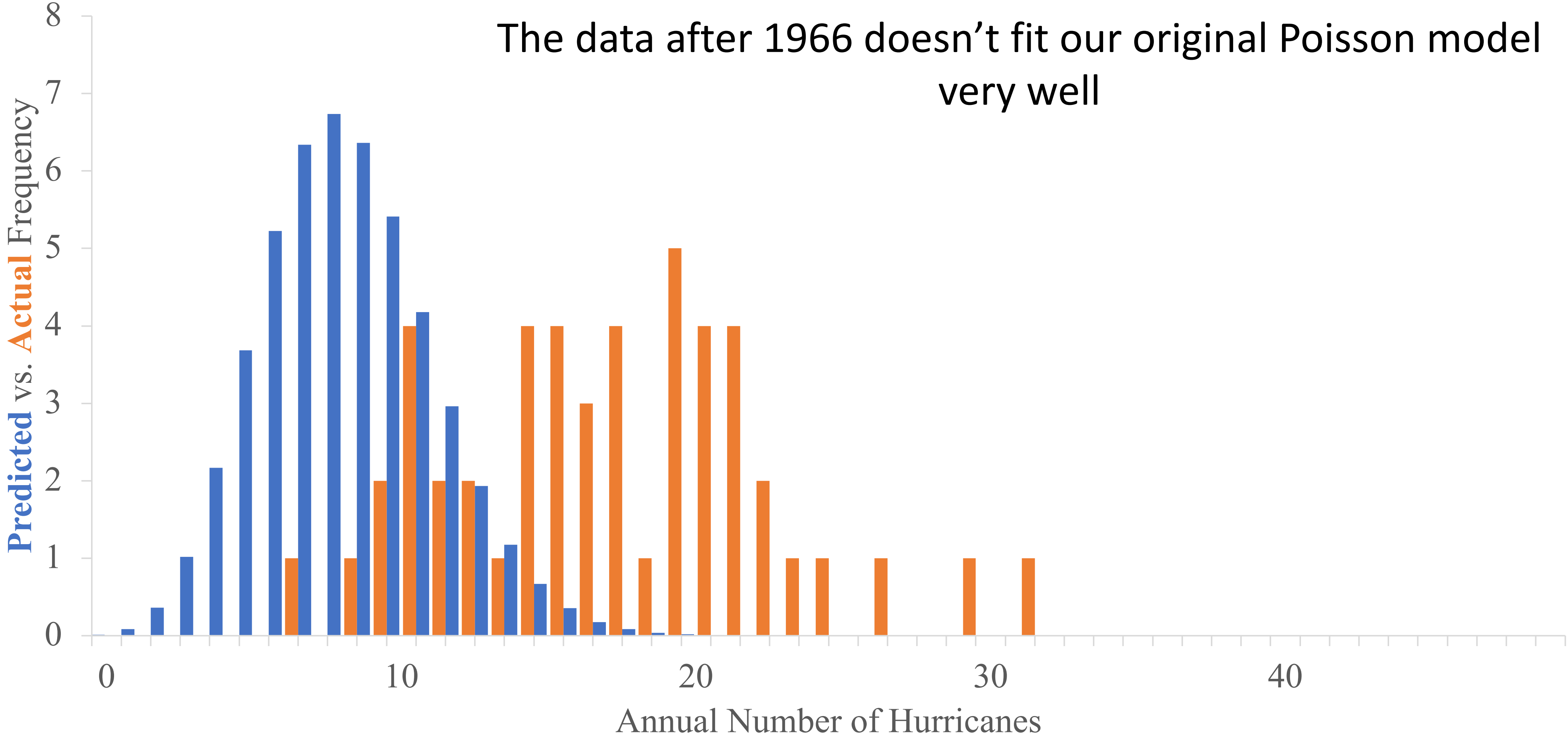
Let’s Model Data As A Poisson

Based on our Poisson model from pre-1966 data, what is the probability of seeing more than 15 hurricanes in one year?

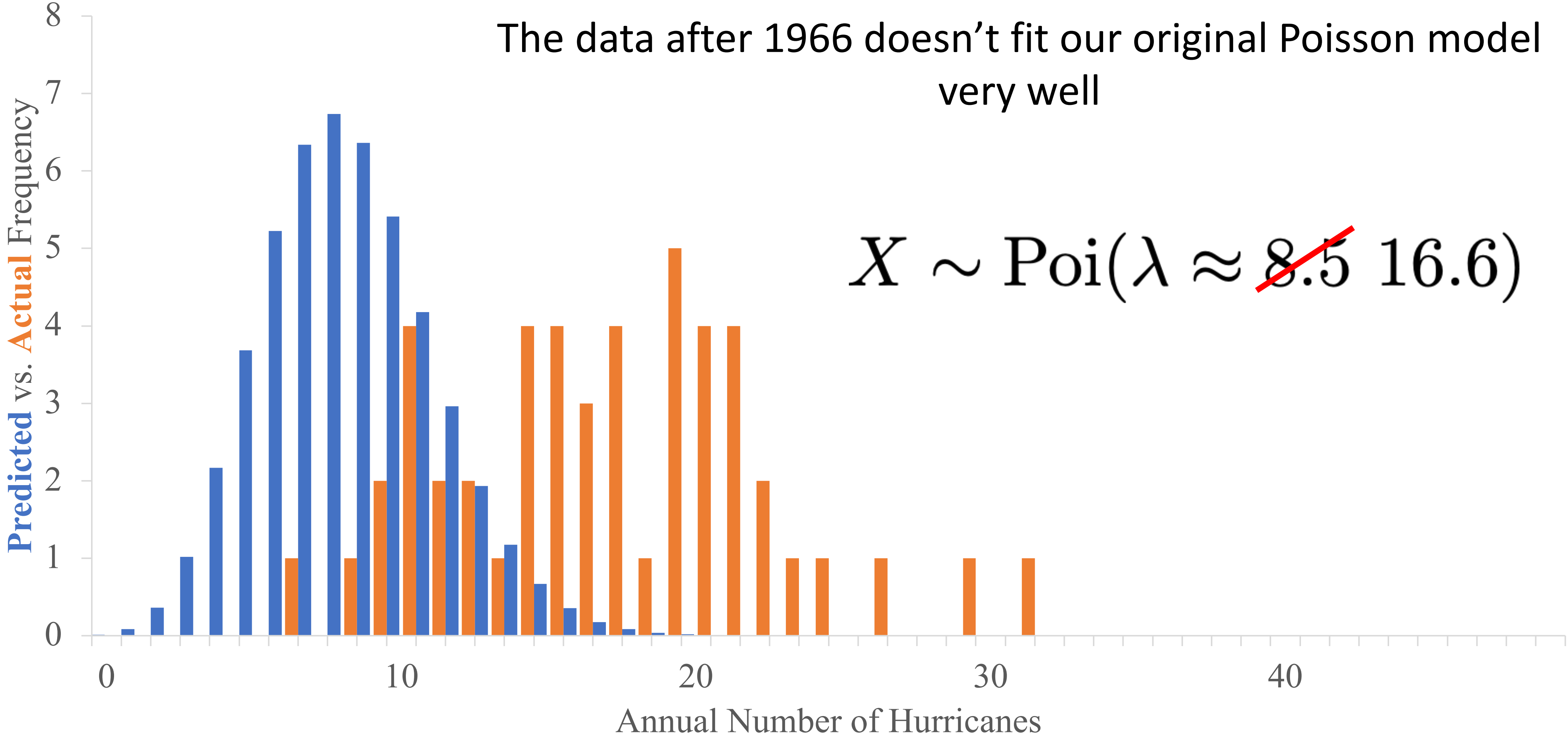
$$X \sim \text{Poi}(\lambda \approx 8.5)$$

$$\begin{aligned} P(X > 15) &= 1 - P(X \leq 15) \\ &= 1 - \sum_{i=0}^{15} P(X = i) \\ &= 0.0135 \end{aligned}$$

Since 1966, The Distribution Has Shifted



Since 1966, The Distribution Has Shifted



Let's Model Data As A Poisson, Round 2

Based on a post-1996 Poisson model, what is the probability of seeing more than 15 hurricanes in one year?

$$X \sim \text{Poi}(\lambda \approx 16.6)$$

$$P(X > 15) = 1 - P(X \leq 15)$$

$$= 1 - \sum_{i=0}^{15} P(X = i)$$

$$= \sout{0.0135} \quad 0.686$$

You can do so much with what you know already

Next Time: ~*Continuous*~ Random Variables