08: Poisson and More

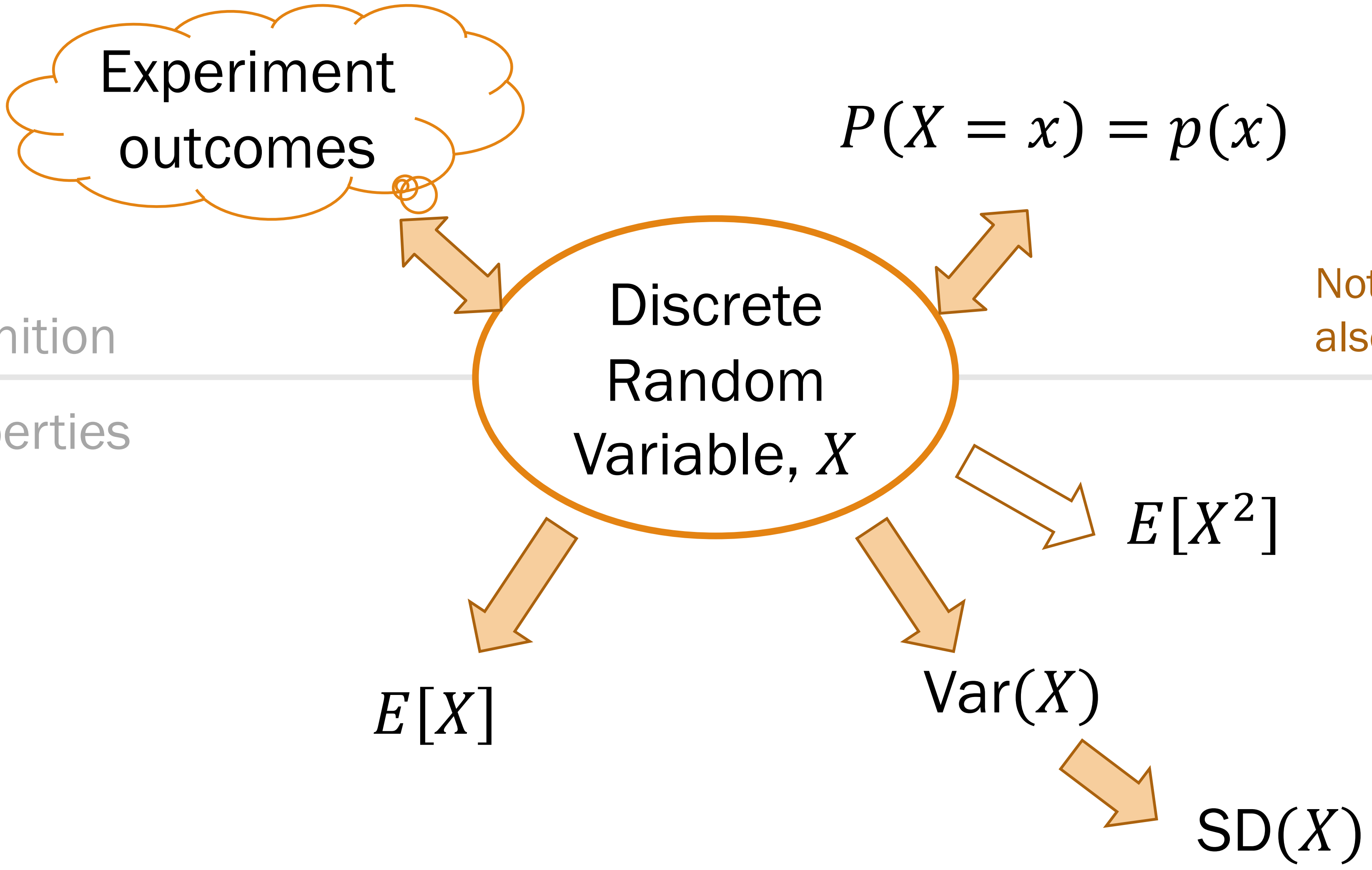
Lisa Yan

October 9, 2019

Discrete random variables

Definition

Properties



Note: Random Variables also called distributions

Variance

The variance of a random variable X with mean $E[X] = \mu$ is

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

Why isn't variance defined as $E\big[X - E[X]\big]$?

$$E\big[X - E[X]\big] = E[X] - E[X] = 0$$

Linearity of expectation!

Binomial random variable

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

Range: $\{0, 1, \ldots, n\}$
(aka support)

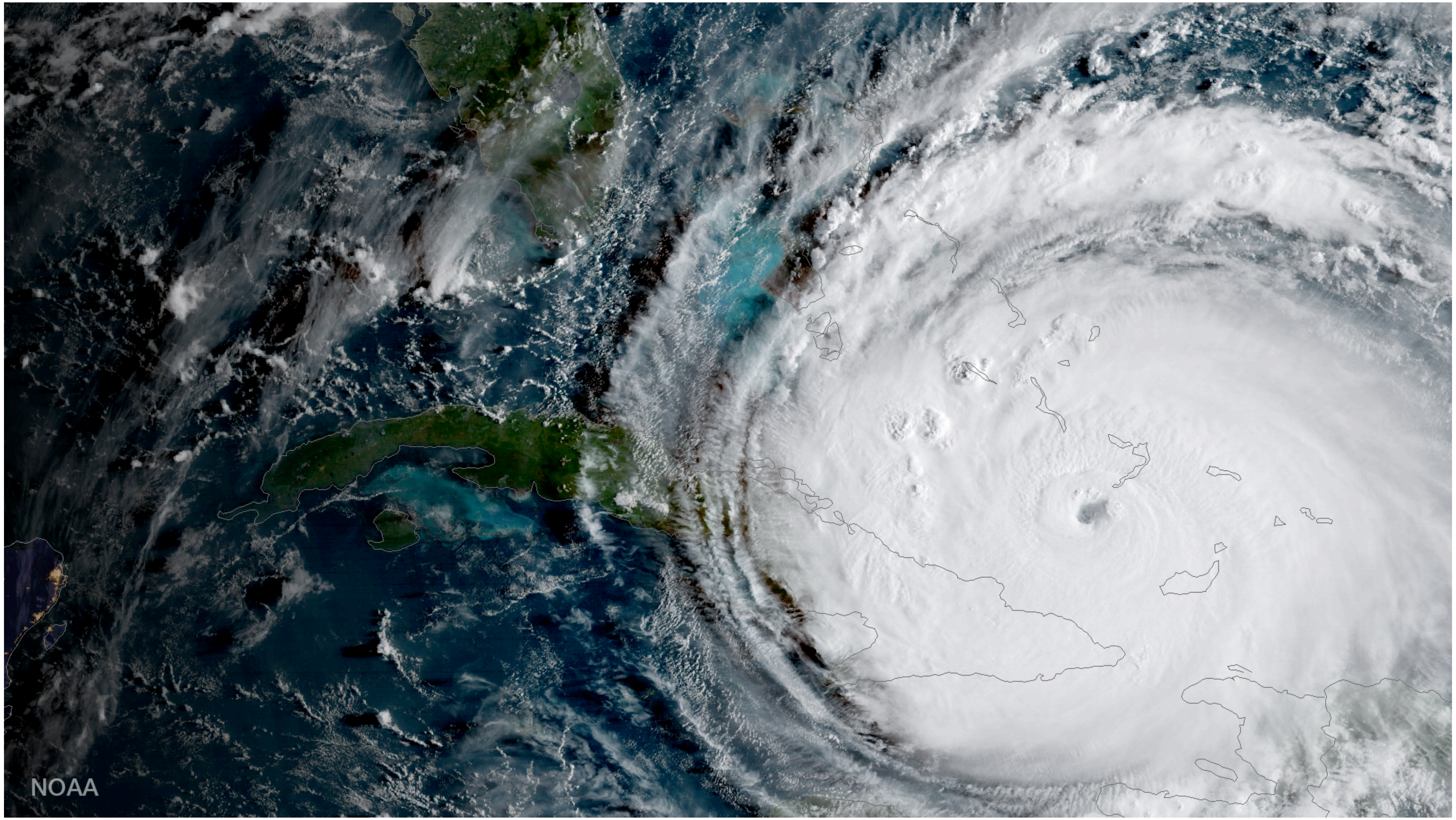
PMF	$P(X = k) = p(k) = \binom{n}{k} p^k (1-p)^{n-k}$
Expectation	$E[X] = np$
Variance	$\text{Var}(X) = np(1-p)$

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

- n: # independent trials
- p: P(success) on each trial

1. The random variable
2. is distributed as a
3. Binomial
4. with parameters

Today's plan: Hurricanes



What is the probability of an extreme weather event?

Today's plan

- Poisson

Poisson Paradigm

Some more Discrete RVs (if time)

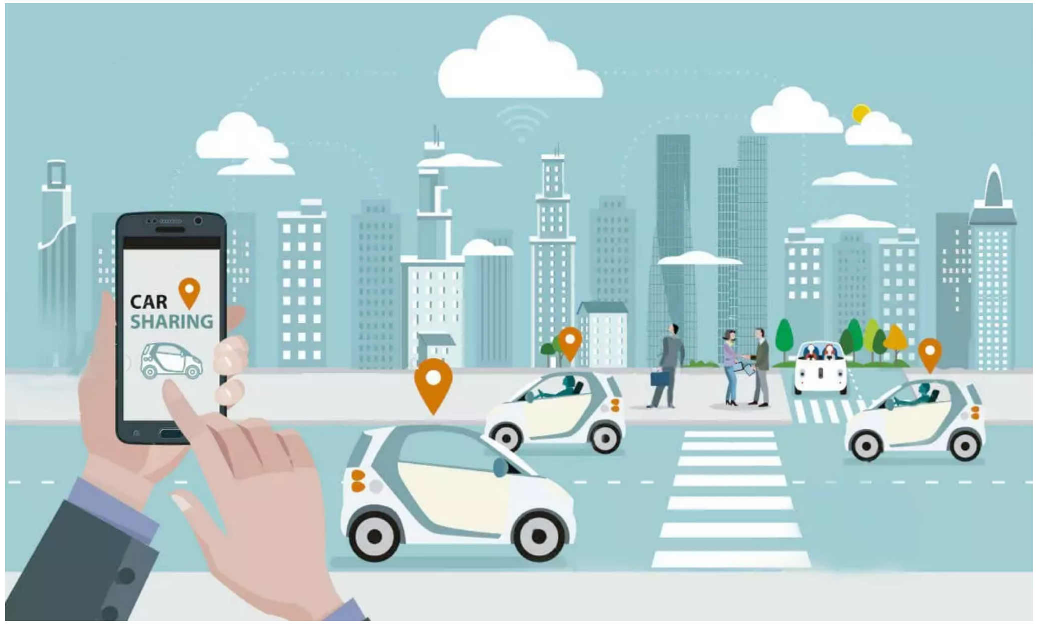
Before we start

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

https://en.wikipedia.org/wiki/E_(mathematical_constant)

Jacob Bernoulli
while studying
compound interest
in 1683

Algorithmic ride sharing



Probability of k requests from this area in the next 1 minute?

Suppose we know: On average, $\lambda = 5$ requests per minute

Algorithmic ride sharing, approximately

Probability of k requests from this area in the next 1 minute?

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

Break a minute down into 60 seconds:

1	2	3	4	5						60
0	0	1	0	1	…	0	0	0	0	1

At each second:

- Independent trial
- You get a request (1) or you don't (0).

Let X = # of requests in minute.

$E[X] = \lambda = 5$

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

$$P(X = k) = \binom{60}{k} \left(\frac{5}{60}\right)^k \left(1 - \frac{5}{60}\right)^{n-k}$$

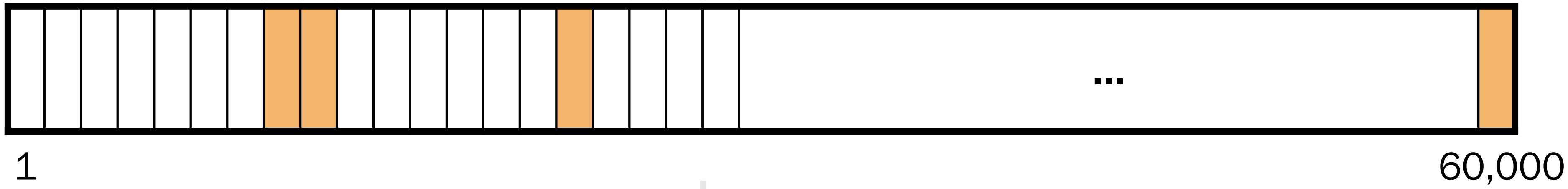
But what if there are *two* requests in the same second?

Algorithmic ride sharing, approximately

Probability of k requests from this area in the next 1 minute?

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

Break a minute down into 60,000 milliseconds:



At each millisecond:

- Independent trial
- You get a request (1) or you don't (0).

Let $X =$ # of requests in minute.

$E[X] = \lambda = 5$

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

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

But what if there are *two* requests in the same millisecond?

Algorithmic ride sharing, approximately

Probability of k requests from this area in the next 1 minute?

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

Break a minute down into infinitely small buckets:

OMG so small

1 ∞

For each time bucket:

- Independent trial
- You get a request (1) or you don't (0).

Let X = # of requests in minute.

$E[X] = \lambda = 5$

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

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

Who wants to see some cool math?

Binomial in the limit

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

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

Expand

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

Rearrange

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

Def natural exponent

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

Expand

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

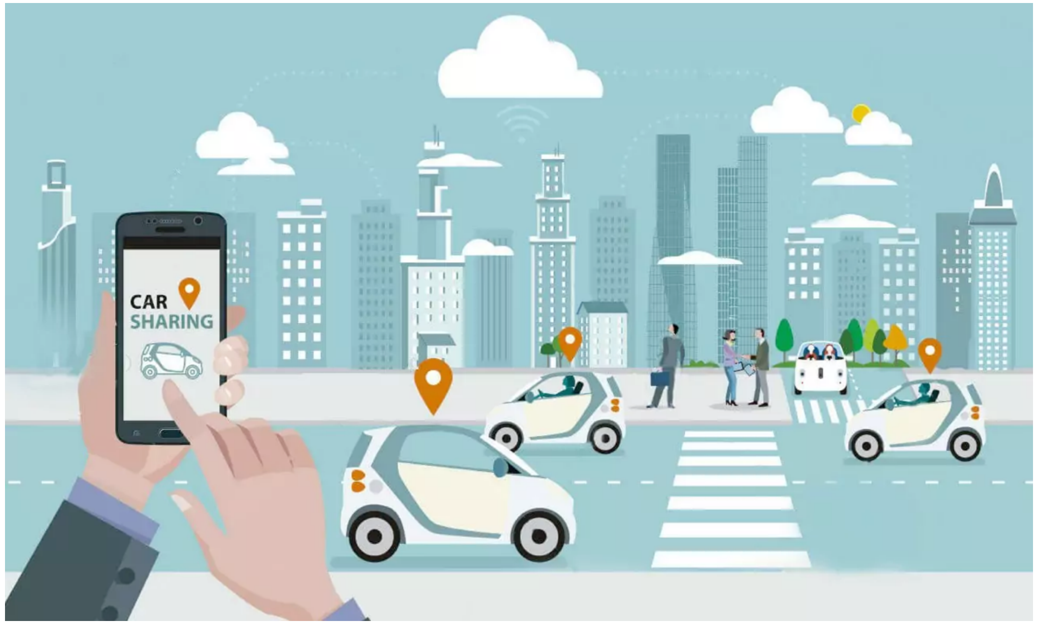
Limit analysis + cancel

$$= \lim_{n\to\infty}\frac{n^k}{n^k}\frac{\lambda^k}{k!}\frac{e^{-\lambda}}{1}$$

Simplify

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

Algorithmic ride sharing



Probability of k requests from this area in the next 1 minute?

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

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

Simeon-Denis Poisson

French mathematician (1781 – 1840)

- Published his first paper at age 18
- Professor at age 21
- Published over 300 papers

“Life is only good for two things: doing mathematics and teaching it.”

Poisson Random Variable

Consider an experiment that lasts a fixed interval of time.

def A **Poisson** random variable X is the number of successes over the experiment duration.

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

Range: $\{0, 1, 2, \ldots\}$

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

Expectation $E[X] = \lambda$

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

Examples:

- # earthquakes per year
- # server hits per second
- # of emails per day

Yes, expectation and variance of Poisson are the same (shown later)

Poisson process

$X \sim \text{Poi}(\lambda)$ $\quad p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$

$E[X] = \lambda$

1. Consider events that occur over time.
   - Event: earthquakes, radioactive decay, web server hits, etc.
   - Time interval: 1 year, 1 sec, whatever
   - Events arrive at average rate λ events/time interval

2. Split time interval into $n \to \infty$ subintervals.
   - Assume at most one event per sub-interval.
   - Event occurrences in sub-intervals are independent.
   - With many sub-intervals, probability of event occurring in any given sub-interval is small

3. Let $X = \#$ events in original time interval. $X \sim \text{Poi}(\lambda)$

Use Poisson if you:
- have a rate
- care about # occurrences

Earthquakes

$X \sim \text{Poi}(\lambda)$ $\quad p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$

$E[X] = \lambda$

There are an average of 2.79 major earthquakes in the world each year.

What is the probability of 3 major earthquakes happening next year?

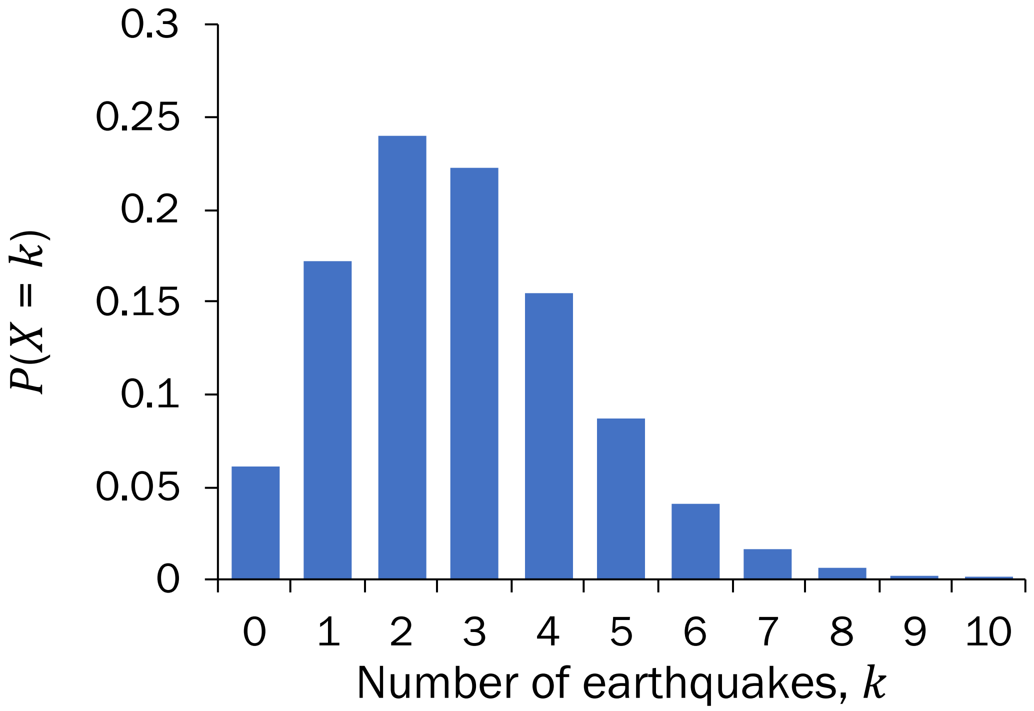
1. Define RVs

$$X \sim \text{Poi}(2.79)$$

2. Solve

$$P(X=3) = e^{-\lambda}\frac{\lambda^k}{k!}, \text{ where } k = 3, \lambda = 2.79$$

$$= e^{-2.79}\frac{(2.79)^3}{3!} \approx 0.23$$



Are earthquakes really Poissonian?

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.

Web server load

$X \sim \text{Poi}(\lambda)$ $\quad p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$
$E[X] = \lambda$

Consider requests to a web server in 1 second.

- In the past, server load averages 2 hits/second.
- Let X = # hits the server receives in a second.

What is $P(X < 5)$?

1. Define RVs

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

2. Solve

$$P(X < 5) = \sum_{k=0}^{4} P(X = k) = \sum_{k=0}^{4} e^{-\lambda}\frac{\lambda^k}{k!}$$, where $\lambda = 2$

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

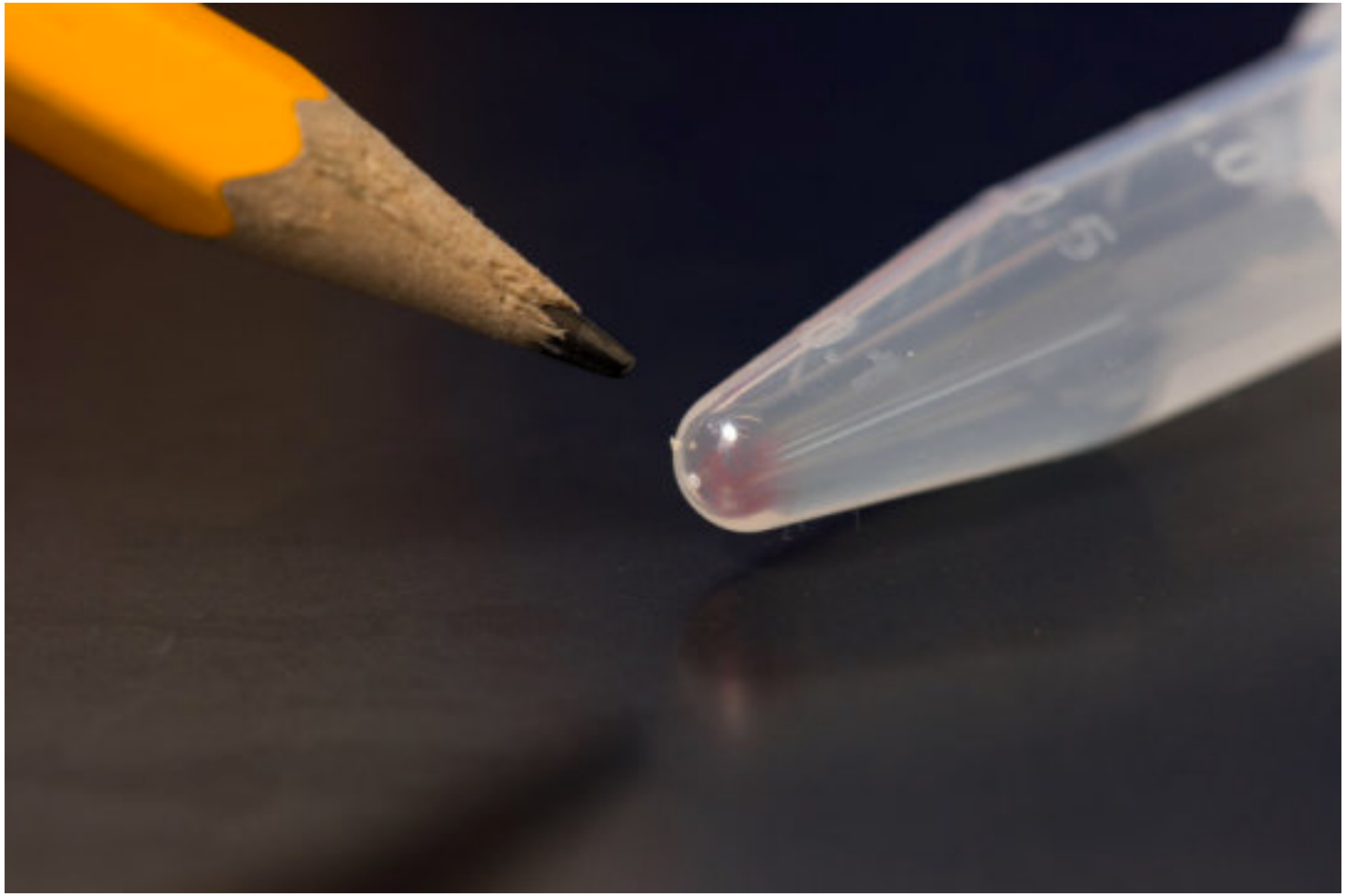
Today's plan

Poisson

Poisson Paradigm

Some more Discrete RVs

DNA



All the movies, images, emails and other digital data from more than 600 smartphones (10,000 GB) can be stored in the faint pink smear of DNA at the end of this test tube.

What is the probability that DNA storage stays uncorrupted?

DNA

What is the probability that DNA storage stays uncorrupted?

- In DNA (and real networks), we store large strings.
- Let string length be long, e.g., $n \approx 10^4$
- Probability of corruption of each base pair is very small, e.g., $p = 10^{-6}$
- Let X = # of corruptions.

What is P(DNA storage is uncorrupted) = $P(X = 0)$?

1. Approach 1:

$$X \sim \text{Bin}(n = 10^4, p = 10^{-6})$$

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

unwieldy! ⚠️ $= \binom{10^4}{0} 10^{-6 \cdot 0} (1 - 10^{-6})^{10^6 - 0}$

≈ 0.99049829

2. Approach 2:

$$X \sim \text{Poi}(\lambda = 10^4 \cdot 10^{-6} = 0.01)$$

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} = e^{-0.01} \frac{0.01^0}{0!}$$

$$= e^{-0.01}$$

≈ 0.99049834 a good approximation! ✅

The Poisson Paradigm, part 1

Poisson approximates Binomial when n is large, p is small, and $\lambda = np$ is "moderate."

Different interpretations of "moderate":

- $n > 20$ and $p < 0.05$
- $n > 100$ and $p < 0.1$

Poisson is Binomial in the limit:

- $\lambda = np$, where $n \to \infty, p \to 0$

👉 Poisson can approximate Binomial!

The Poisson Paradigm, part 1

$X \sim \text{Poi}(\lambda)$ $\quad E[X] = \lambda$

$Y \sim \text{Bin}(n, p)$ $\quad E[Y] = np$

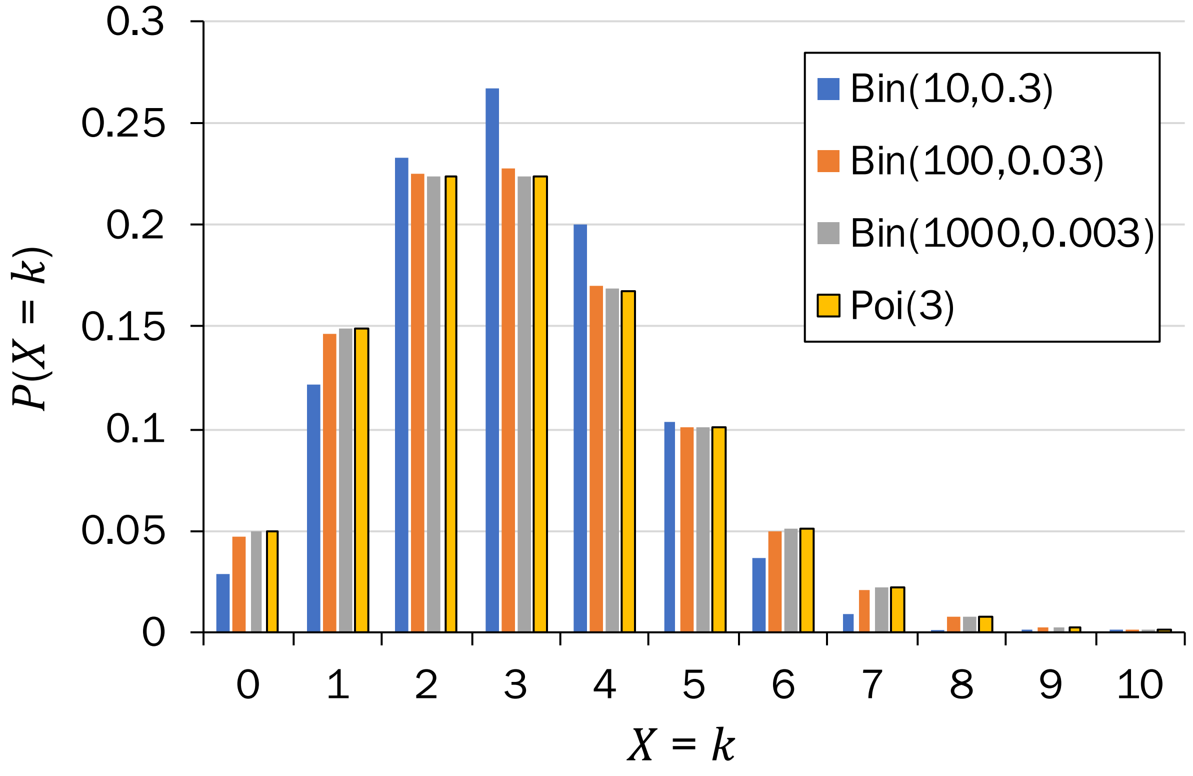
Poisson approximates Binomial when n is large, p is small, and $\lambda = np$ is "moderate."

Different interpretations of "moderate":

- $n > 20$ and $p < 0.05$
- $n > 100$ and $p < 0.1$

Poisson is Binomial in the limit:

- $\lambda = np$, where $n \to \infty, p \to 0$



Poisson can approximate Binomial!

Can these Binomial RVs be approximated?

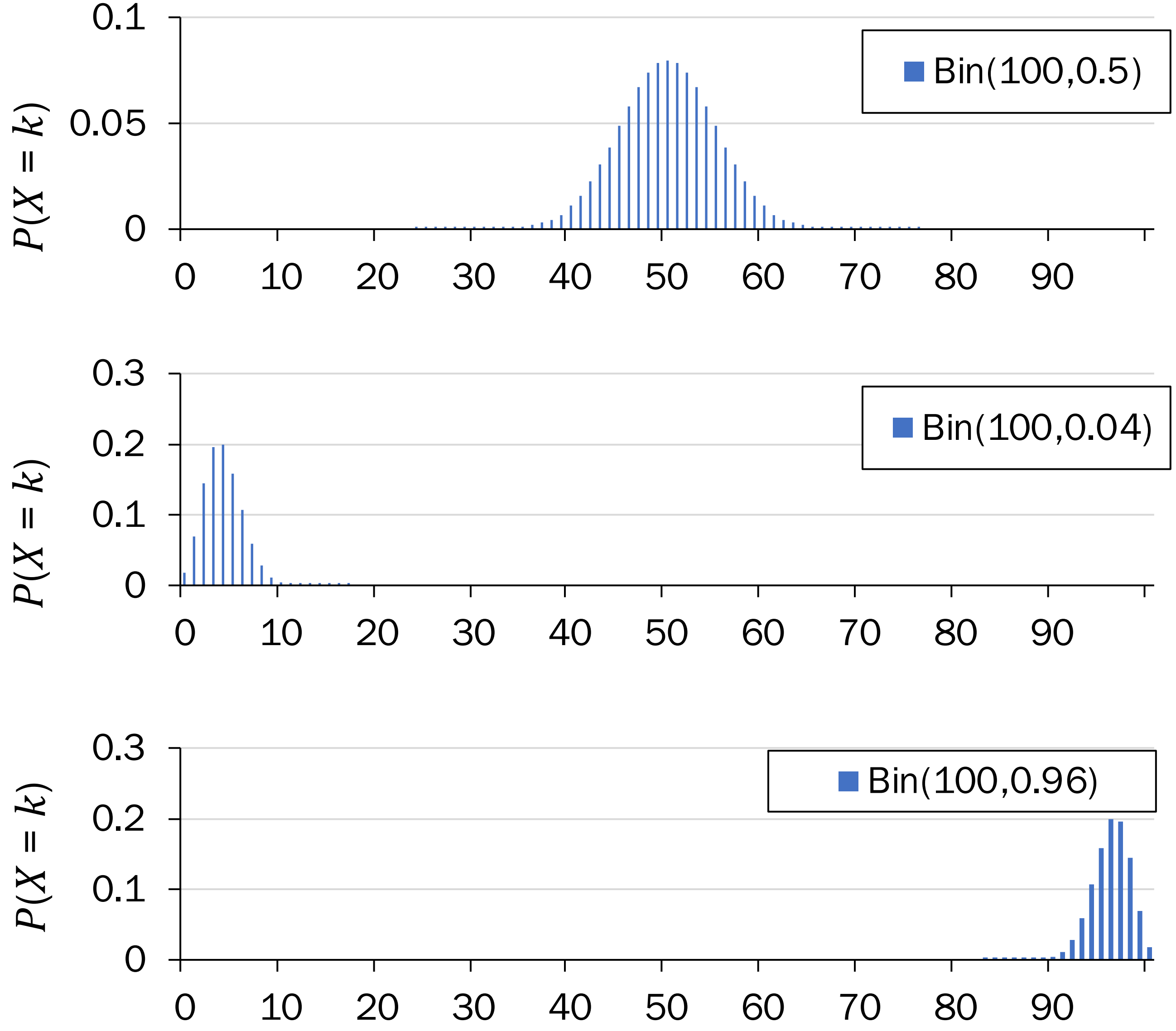
Poisson approximates Binomial when n is large, p is small, and $\lambda = np$ is "moderate."

Different interpretations of "moderate":

- $n > 20$ and $p < 0.05$
- $n > 100$ and $p < 0.1$

Poisson is Binomial in the limit:

- $\lambda = np$, where $n \to \infty, p \to 0$



Can these Binomial RVs be approximated?

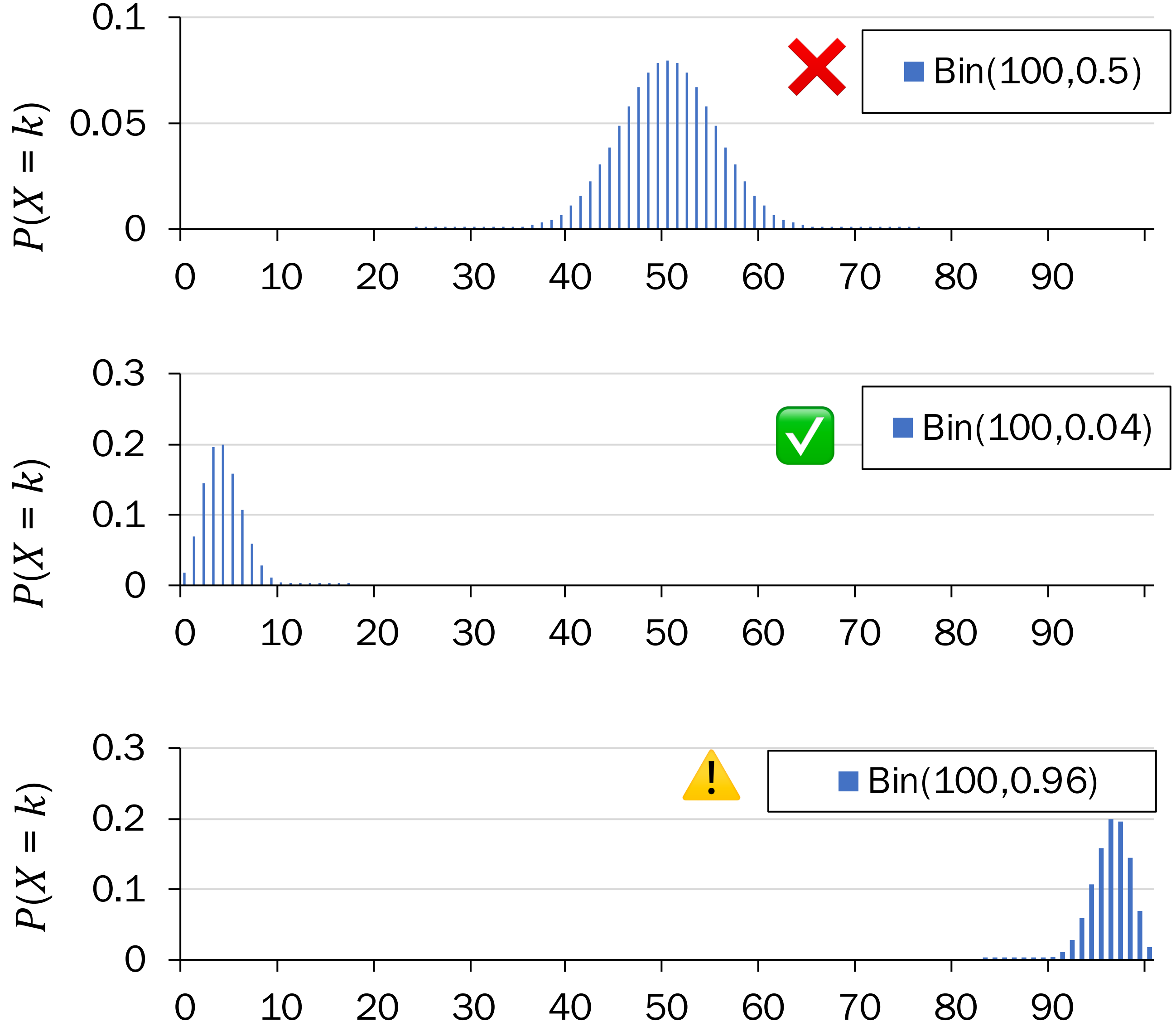
Poisson approximates Binomial when n is large, p is small, and $\lambda = np$ is "moderate."

Different interpretations of "moderate":

- $n > 20$ and $p < 0.05$
- $n > 100$ and $p < 0.1$

Poisson is Binomial in the limit:

- $\lambda = np$, where $n \to \infty, p \to 0$



Can these Binomial RVs be approximated?

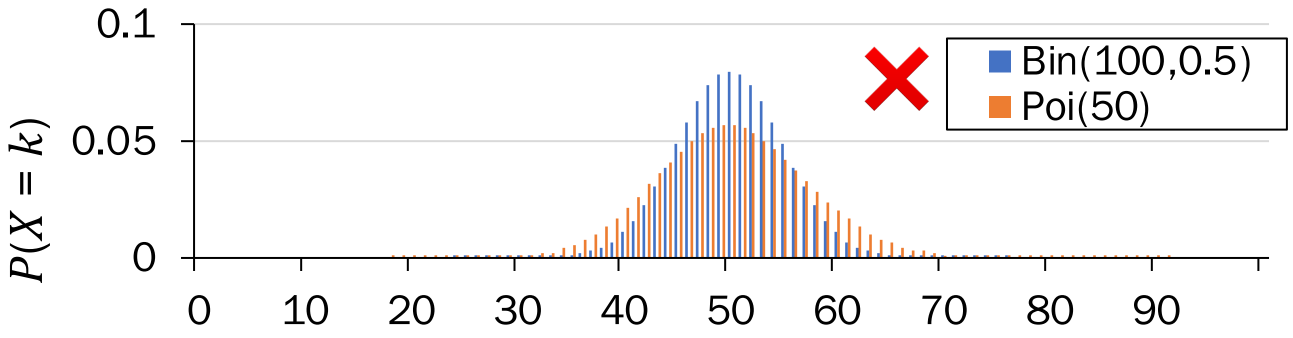
Poisson approximates Binomial when n is large, p is small, and $\lambda = np$ is "moderate."

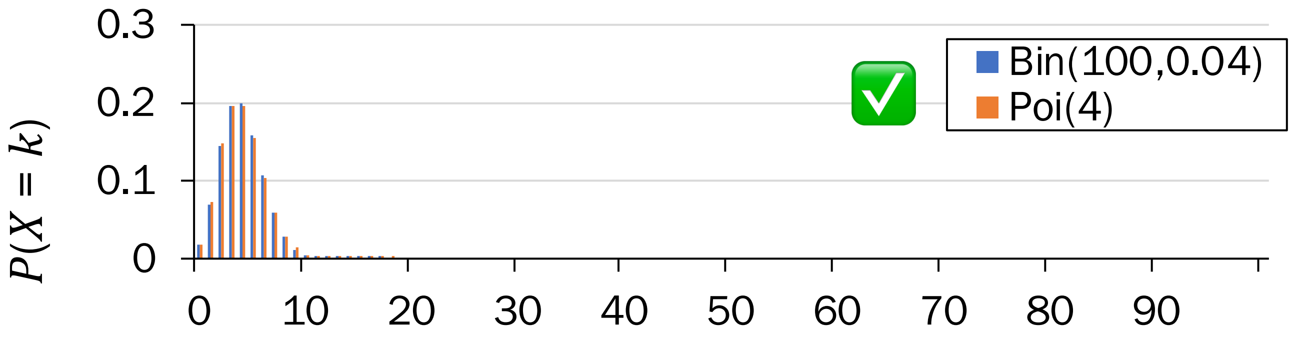
Different interpretations of "moderate":

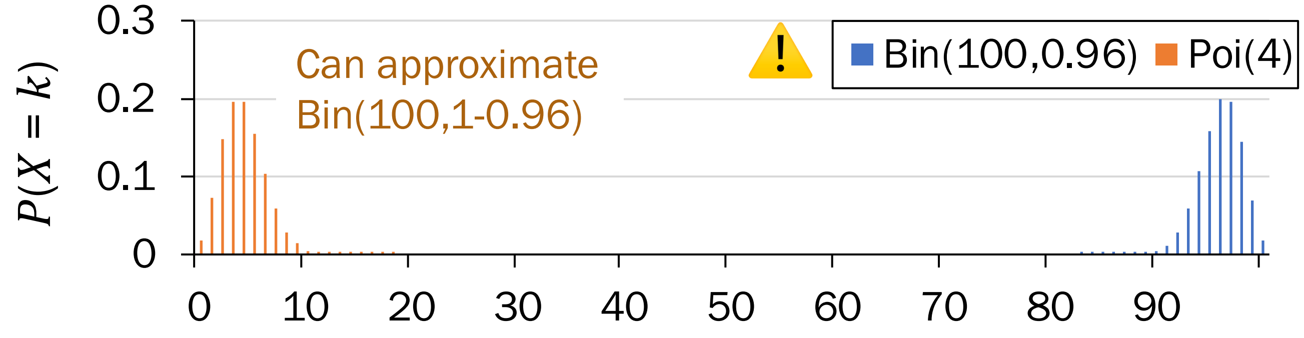
- $n > 20$ and $p < 0.05$
- $n > 100$ and $p < 0.1$

Poisson is Binomial in the limit:

- $\lambda = np$, where $n \to \infty, p \to 0$







Break for jokes/ announcements

Poisson Random Variable

Consider an experiment that lasts a fixed interval of time.

def A **Poisson** random variable X is the number of occurrences over the experiment duration.

$X \sim \text{Poi}(\lambda)$	PMF	$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$
	Expectation	$E[X] = \lambda$
Range: $\{0, 1, 2, \ldots\}$	Variance	$\text{Var}(X) = \lambda$

Examples:

- # earthquakes per year
- # server hits per second
- # of emails per day

👉 Yes, expectation and variance of Poisson are the same (intuition now)

Properties of Poi(λ) with the Poisson paradigm

Recall the Binomial:

$Y \sim \text{Bin}(n, p)$	Expectation	$E[Y] = np$
	Variance	$\text{Var}(Y) = np(1-p)$

Consider $X \sim \text{Poi}(\lambda)$, where $\lambda = np$ $(n \to \infty, p \to 0)$:

$X \sim \text{Poi}(\lambda)$	Expectation	$E[X] = \lambda$
	Variance	$\text{Var}(X) = \lambda$

Proof:

$$E[X] = np = \lambda$$
$$\text{Var}(X) = np(1-p) \to \lambda(1-0) = \lambda$$

A Real License Plate Seen at Stanford

No, it's not mine…
but I kind of wish it was.

Poisson Paradigm, part 2

Poisson can still provide a good approximation of the Binomial, even when assumptions are "mildly" violated.

You can apply the Poisson approximation when:

- "Successes" in trials are not entirely independent
  e.g.: # entries in each bucket in large hash table.
- Probability of "Success" in each trial varies (slightly), like a small relative change in a very small p
  e.g.: Average # requests to web server/sec may fluctuate slightly due to load on network

Today's plan

Poisson

Poisson Paradigm

⇨ Some more Discrete RVs

More discrete RVs

Part of CS109 learning goals:

- Translate a problem statement into a random variable
- Understand new random variables

We focus primarily on Binomial, Bernoulli, and Poisson.

Here are a few more to get a sense of how random variables work.

Focus on understanding how and when to use RVs, not on memorizing PMFs.

Geometric RV

Consider an experiment: independent trials of Ber(p) random variables.

<u>def</u> A **Geometric** random variable X is the # of trials until the <u>first</u> success.

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

Range: $\{1, 2, \dots\}$

PMF	$P(X = k) = (1 - p)^{k-1} p$
Expectation	$E[X] = \frac{1}{p}$
Variance	$\text{Var}(X) = \frac{1-p}{p^2}$

Examples:

- Flipping a coin ($P(\text{heads}) = p$) until first heads appears
- Generate bits with $P(\text{bit} = 1) = p$ until first 1 generated

Negative Binomial RV

Consider an experiment: independent trials of Ber(p) random variables.

def A **Negative Binomial** random variable X is the # of trials until r successes.

(fixed lecture error)

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

Range: $\{r, r+1, \dots\}$

PMF	$P(X = k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r$
Expectation	$E[X] = \frac{r}{p}$
Variance	$\text{Var}(X) = \frac{r(1-p)}{p^2}$

Examples:

- Flipping a coin until r^{th} heads appears
- # of strings to hash into table until bucket 1 has r entries

👉 $\text{Geo}(p) = \text{NegBin}(1, p)$

Grid of random variables

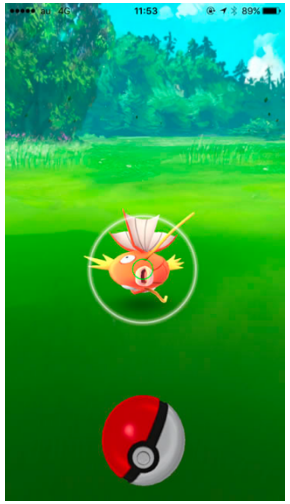
	Number of successes	Time until success	
One trial	$\text{Ber}(p)$	$\text{Geo}(p)$	One success
Several trials	$n = 1$ $\text{Bin}(n, p)$	$r = 1$ $\text{NegBin}(r, p)$	Several successes
Interval of time	$\text{Poi}(\lambda)$	(tomorrow)	Interval of time to first success

Catching Pokemon

Wild Pokemon are captured by throwing Pokeballs at them.

- Each ball has probability $p = 0.1$ of capturing the Pokemon.
- Each ball is an independent trial.

What is the probability that you catch the Pokemon on the 5th try?



1. Define events/ RVs & state goal

   $X \sim$ some distribution

   Want: $P(X = 5)$

2. Solve

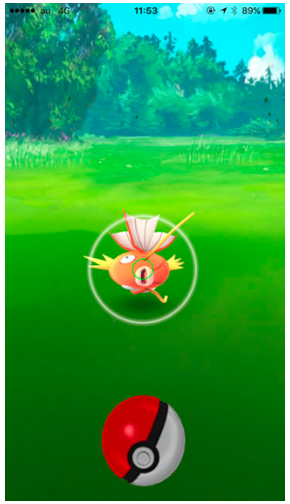
   A. $X \sim \text{Bin}(5, 0.1)$
   B. $X \sim \text{Poi}(0.5)$
   C. $X \sim \text{NegBin}(5, 0.1)$
   D. $X \sim \text{NegBin}(1, 0.1)$
   E. $X \sim \text{Geo}(0.1)$
   F. None/other

Catching Pokemon

Wild Pokemon are captured by throwing Pokeballs at them.

- Each ball has probability p = 0.1 of capturing the Pokemon.
- Each ball is an independent trial.

What is the probability that you catch the Pokemon on the 5th try?



1. Define events/ RVs & state goal

$X \sim$some distribution

Want: $P(X = 5)$

2. Solve

A. $X \sim \text{Bin}(5, 0.1)$
B. $X \sim \text{Poi}(0.5)$
C. $X \sim \text{NegBin}(5, 0.1)$
D. $X \sim \text{NegBin}(1, 0.1)$
E. $X \sim \text{Geo}(0.1)$
F. None/other

Be clear about what is *variable* (unknown) in the problem setup.

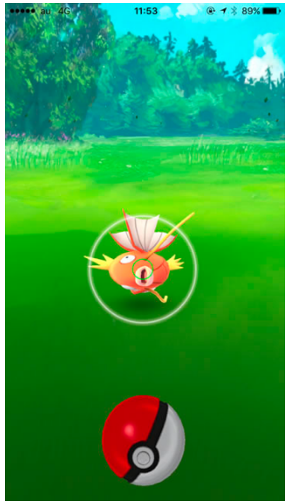
Catching Pokemon

$X \sim \text{Geo}(p) \quad p(k) = (1-p)^{k-1} p$

Wild Pokemon are captured by throwing Pokeballs at them.

- Each ball has probability p = 0.1 of capturing the Pokemon.
- Each ball is an independent trial.

What is the probability that you catch the Pokemon on the 5th try?



1. Define events/ RVs & state goal

$X \sim \text{Geo}(0.1)$

Want: $P(X = 5)$

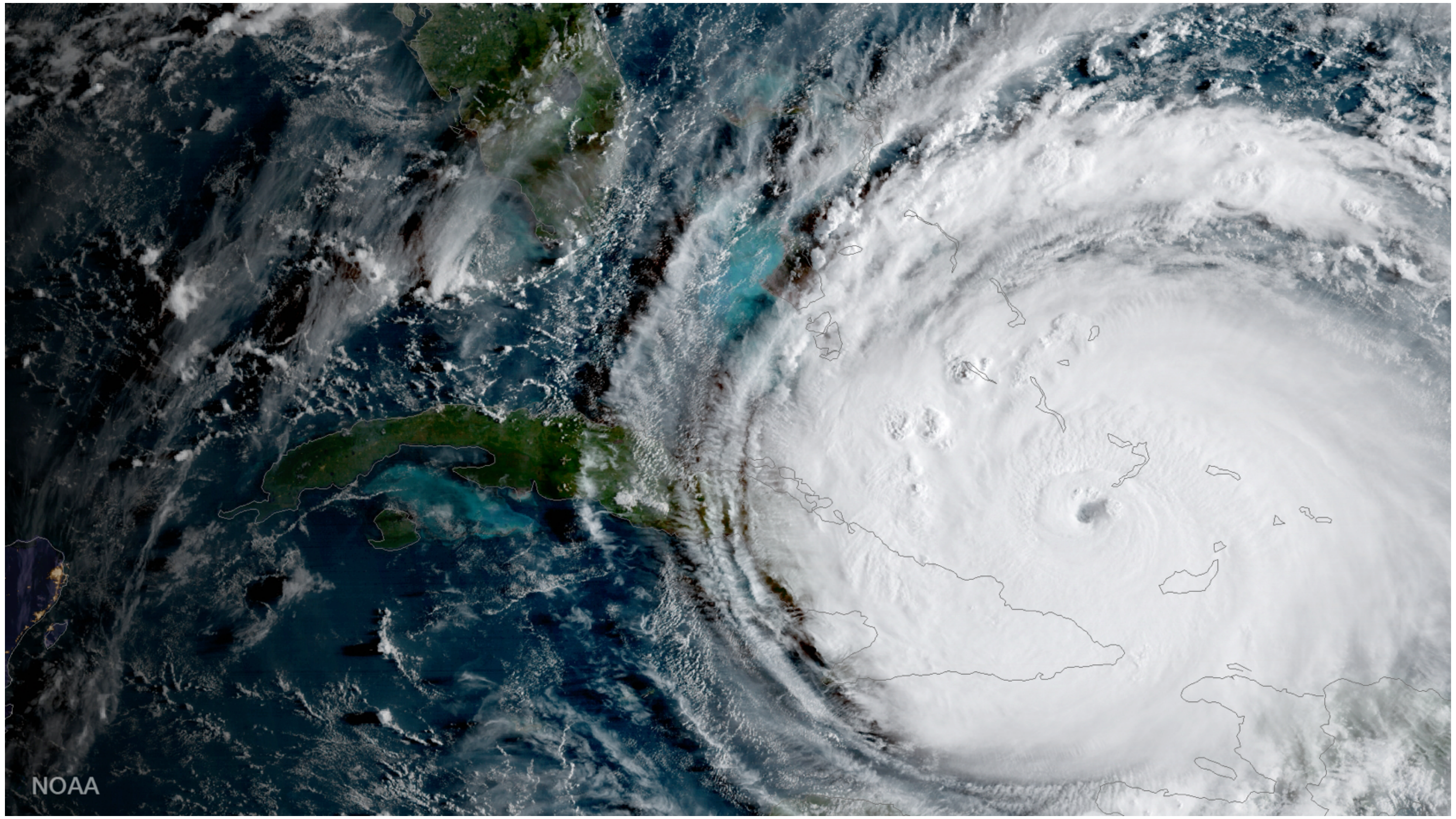
2. Solve

$$P(X=5) = (1-p)^{k-1} p, \text{ where } k = 5, p = 0.1$$

$$= (0.9)^4 (0.1)$$

$$\approx 0.066$$

Hurricanes



What is the probability of an extreme weather event?

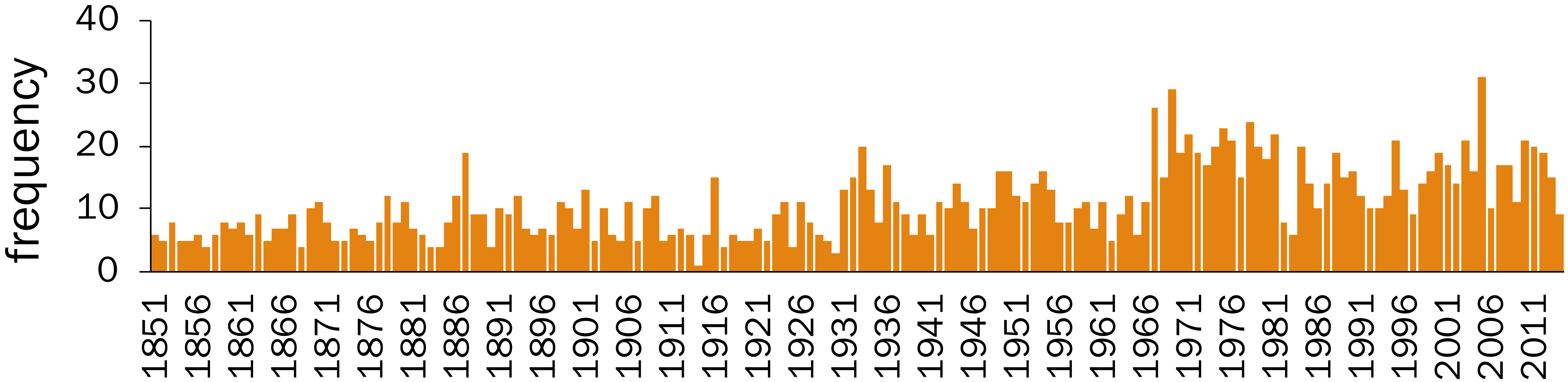
How do we model the number of hurricanes in a season (year)?

Step 1. graph your distribution.

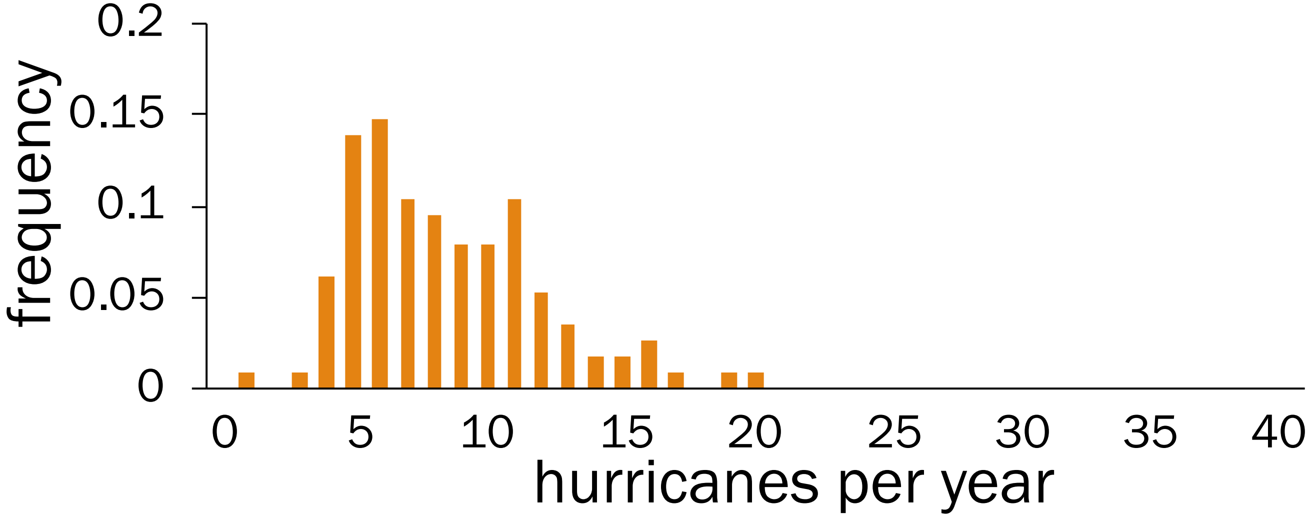
Hurricanes per year since 1851

Which graph is a histogram (i.e., distribution) of frequency (# of hurricanes per year)?

A.

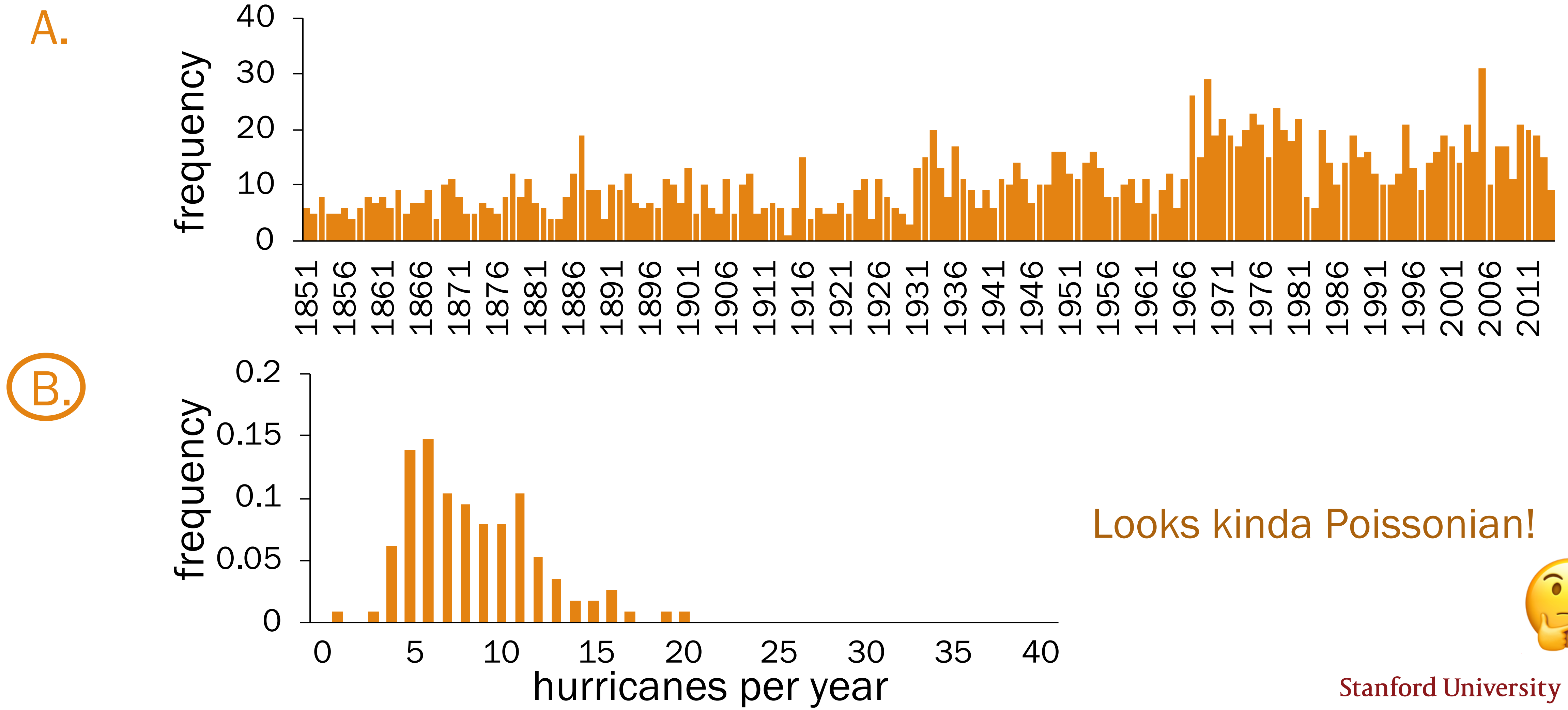


B.



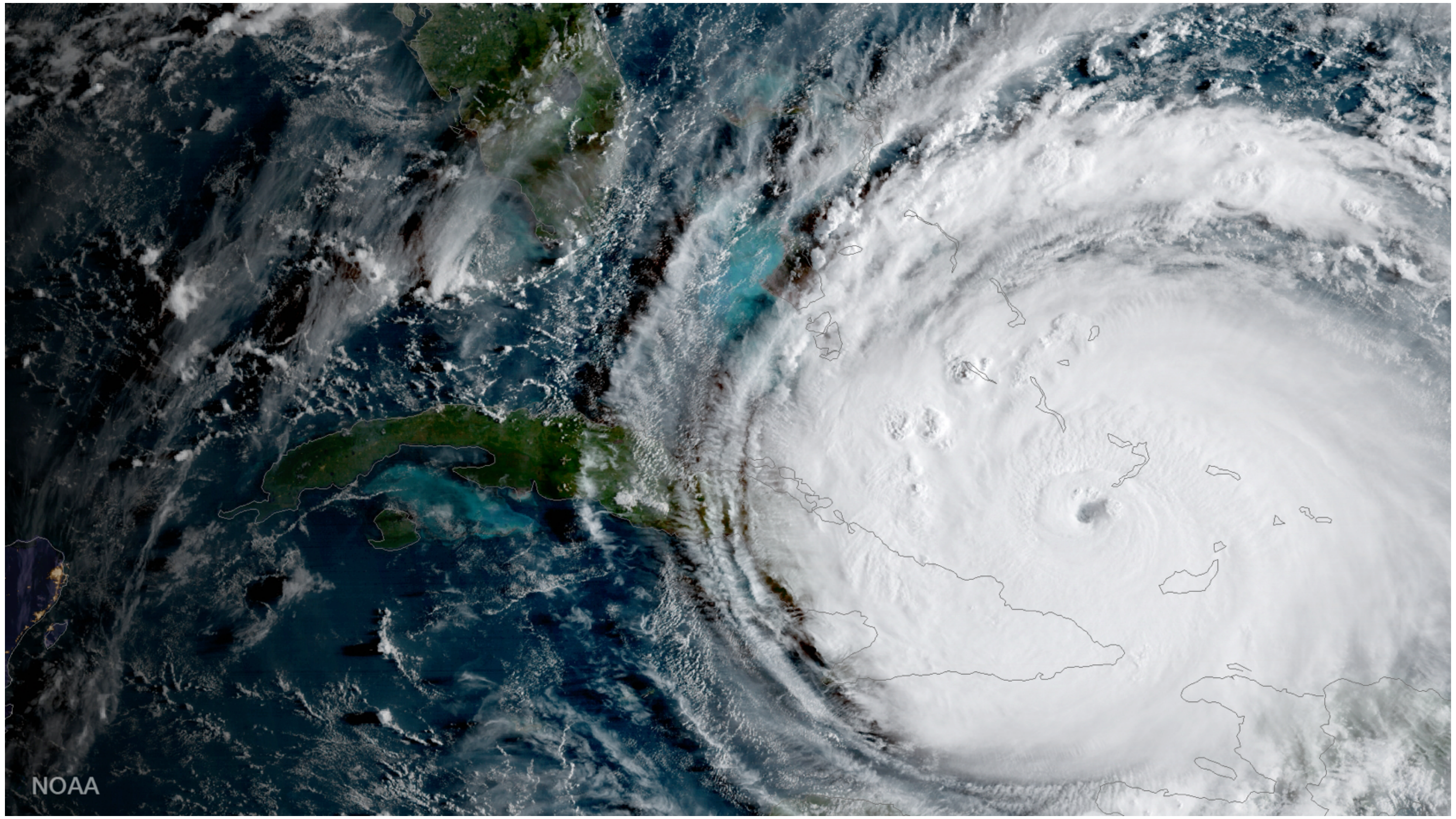
Hurricanes per year since 1851

Which graph is a histogram (i.e., distribution) of frequency (# of hurricanes per year)?



Looks kinda Poissonian!

Hurricanes



How do we model the number of hurricanes in a season (year)?

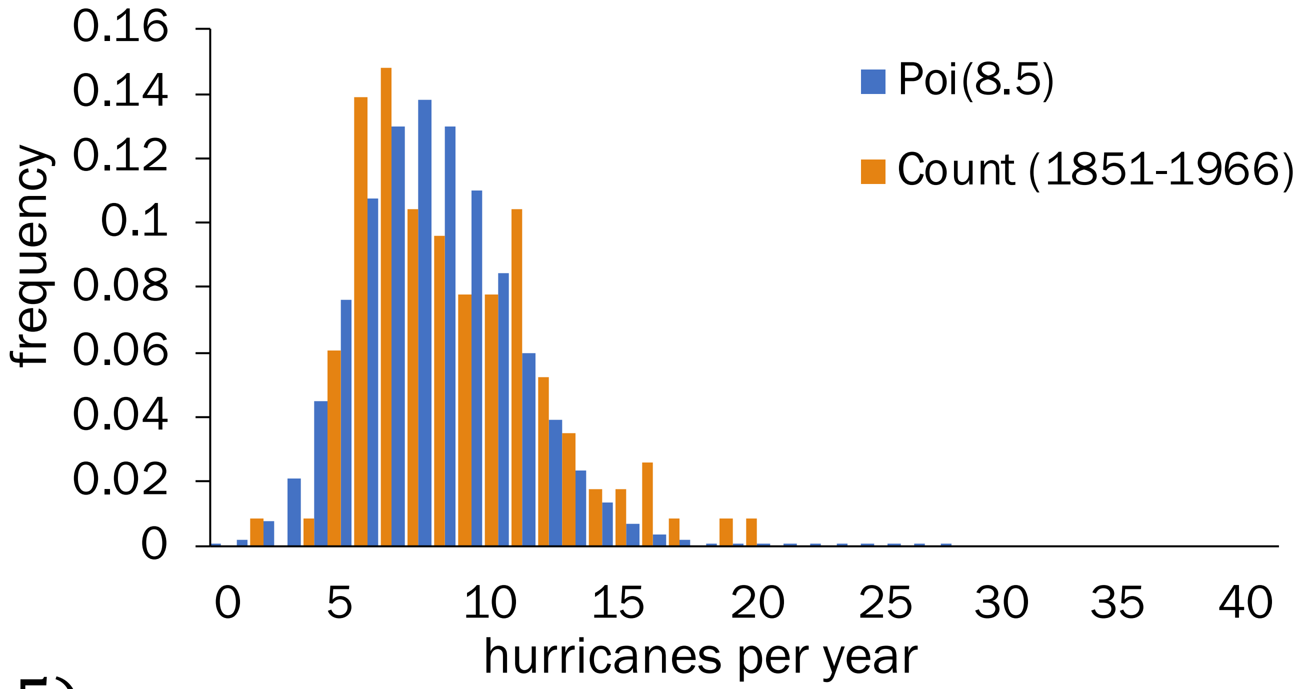
👉 Step 2. Find a reasonable distribution (Poisson) and compute parameters.

To the code!!

Improbability

$X \sim \text{Poi}(\lambda) \quad p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$

Until 1966, things look pretty Poisson.



What is the probability of over 15 hurricanes in a season (year) given that the distribution doesn't change?

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

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

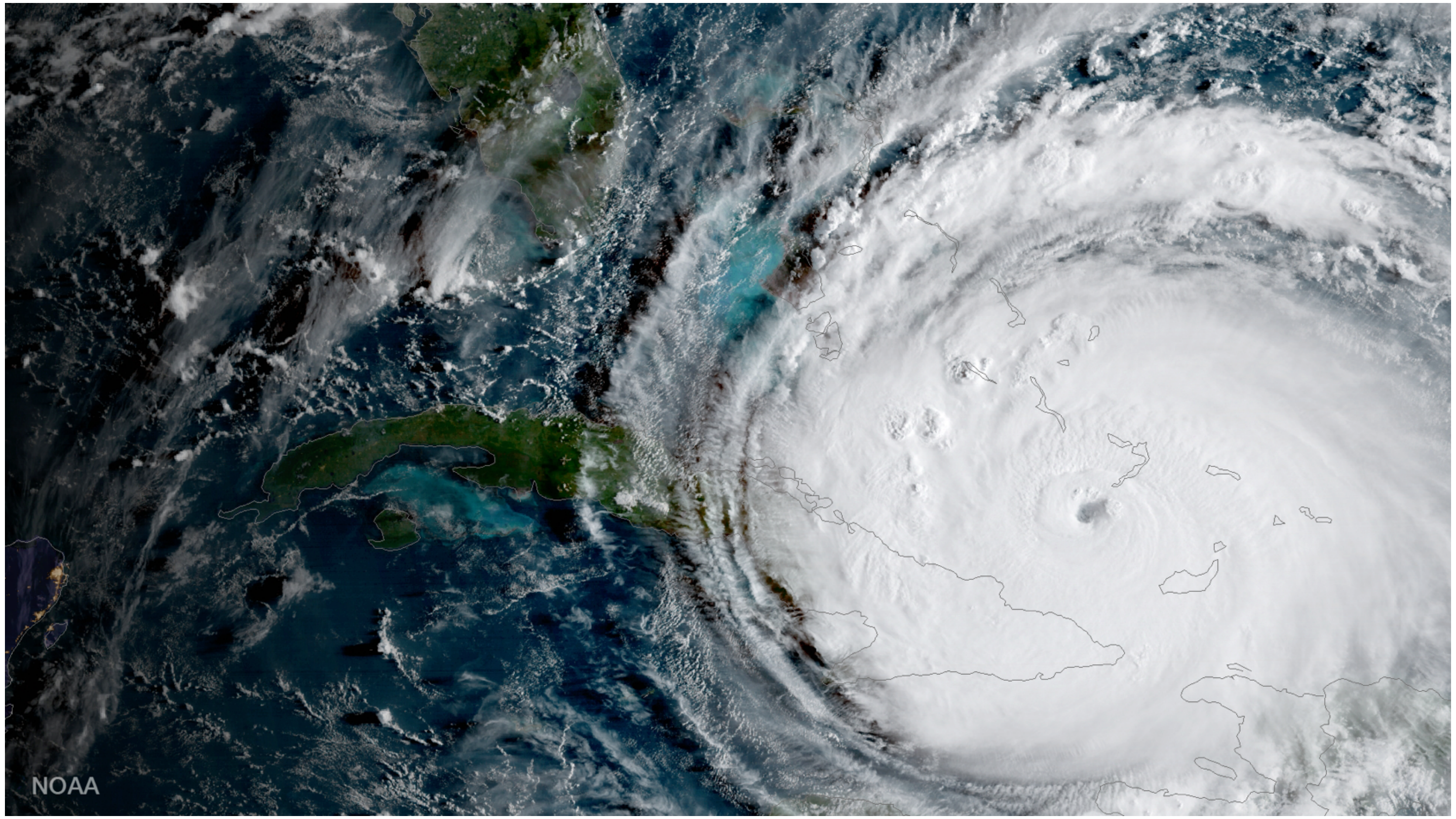
$$= 1 - 0.986 = 0.014$$

This is the PMF of a Poisson. Your favorite programming language has a function for it.

In Python 3:

```
from scipy import stats
X = stats.poisson(8.5)
X.pmf(k)
```

Hurricanes



How do we model the number of hurricanes in a season (year)?

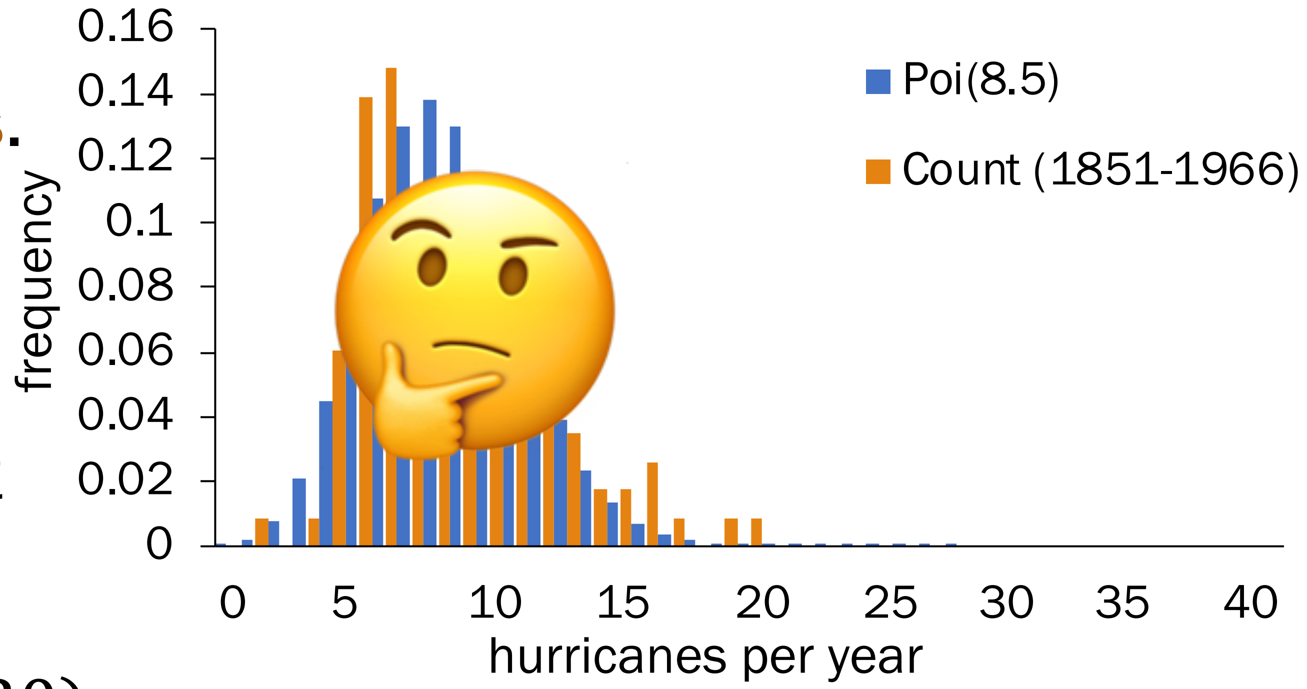
👉 Step 3. See if there are outliers

Improbability

$X \sim \text{Poi}(\lambda) \qquad p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$

Since 1966, there have been **<u>two years</u>** with over 30 hurricanes.

What is the probability of over 30 hurricanes in a season (year) given that the distribution doesn't change?



$$P(X > 30) = 1 - P(X \leq 30)$$

$$= 1 - \sum_{k=0}^{30} P(X = k)$$

$$= 2.2\text{E} - 09$$

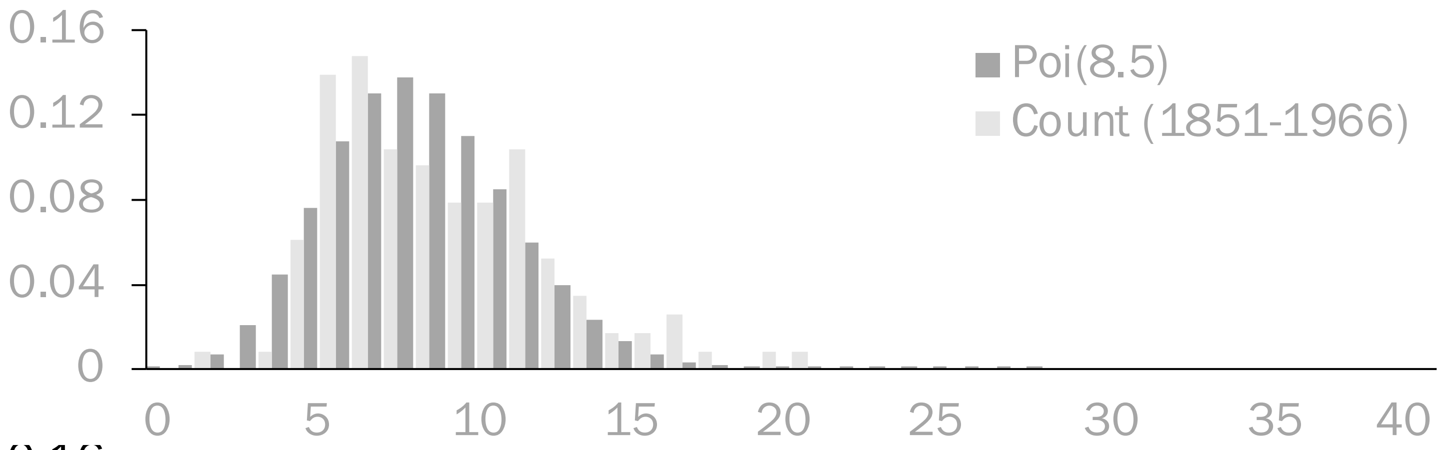
This is the PMF of a Poisson.
Your favorite programming language has a function for it.

In Python 3:

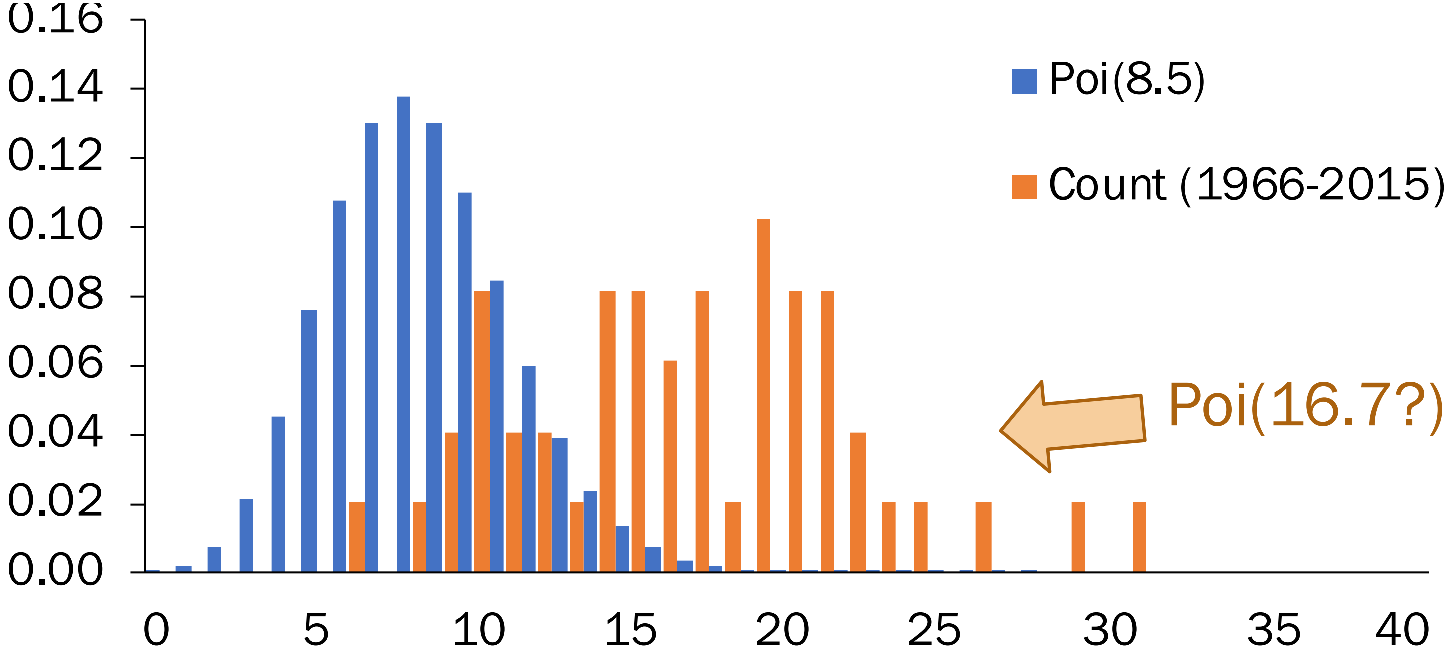
```
from scipy import stats
X = stats.poisson(8.5)
X.pmf(k)
```

The distribution has changed.

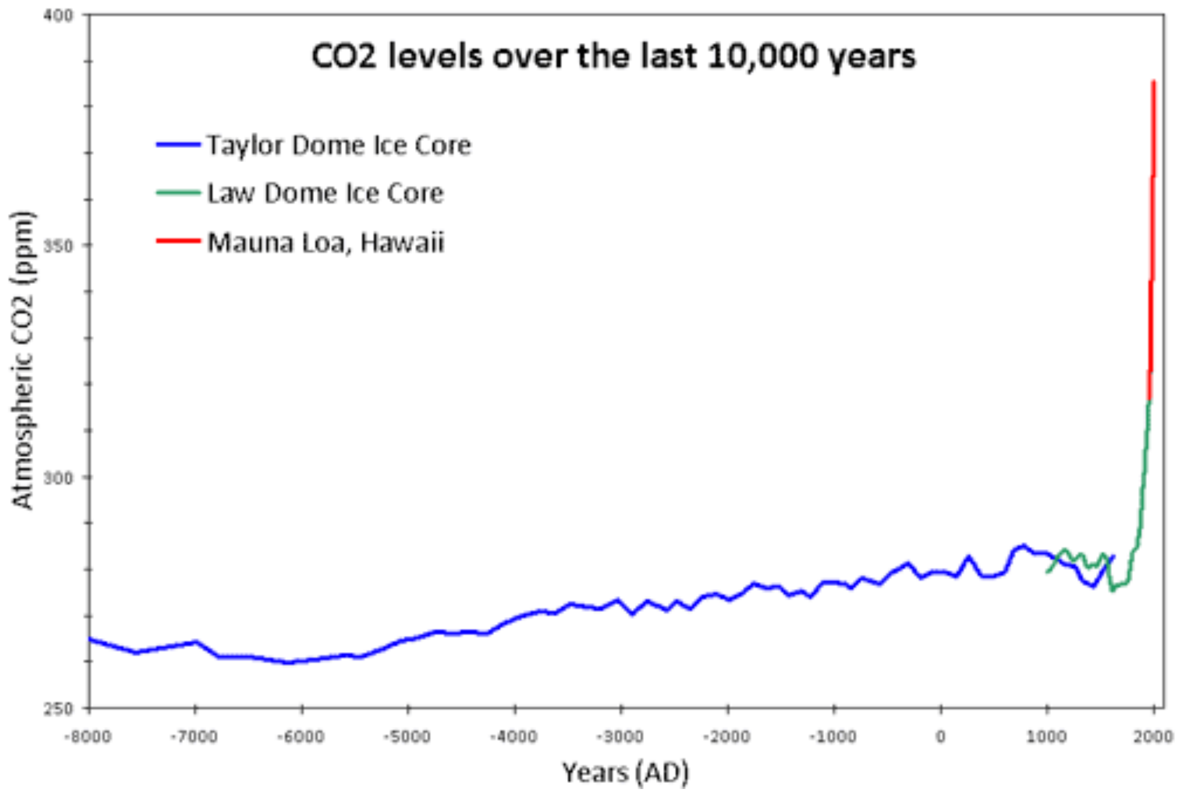
1851–
1966

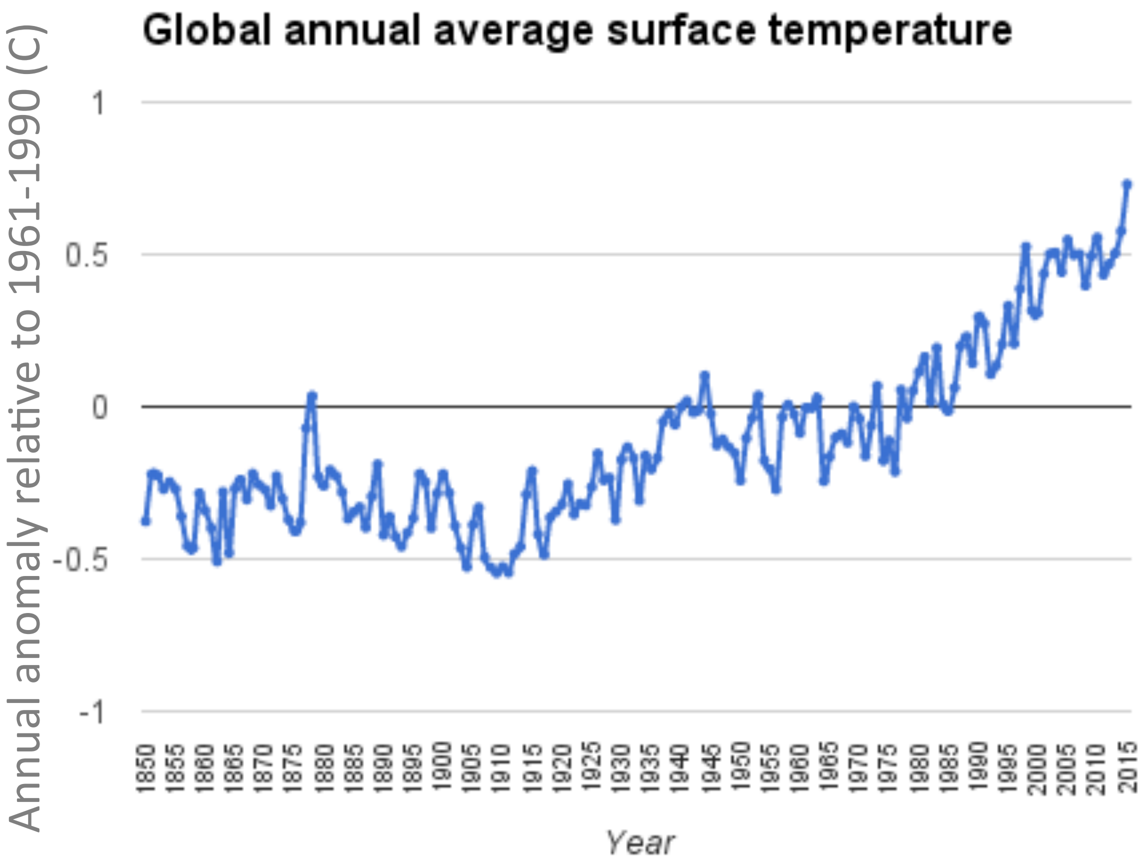


Since
1966

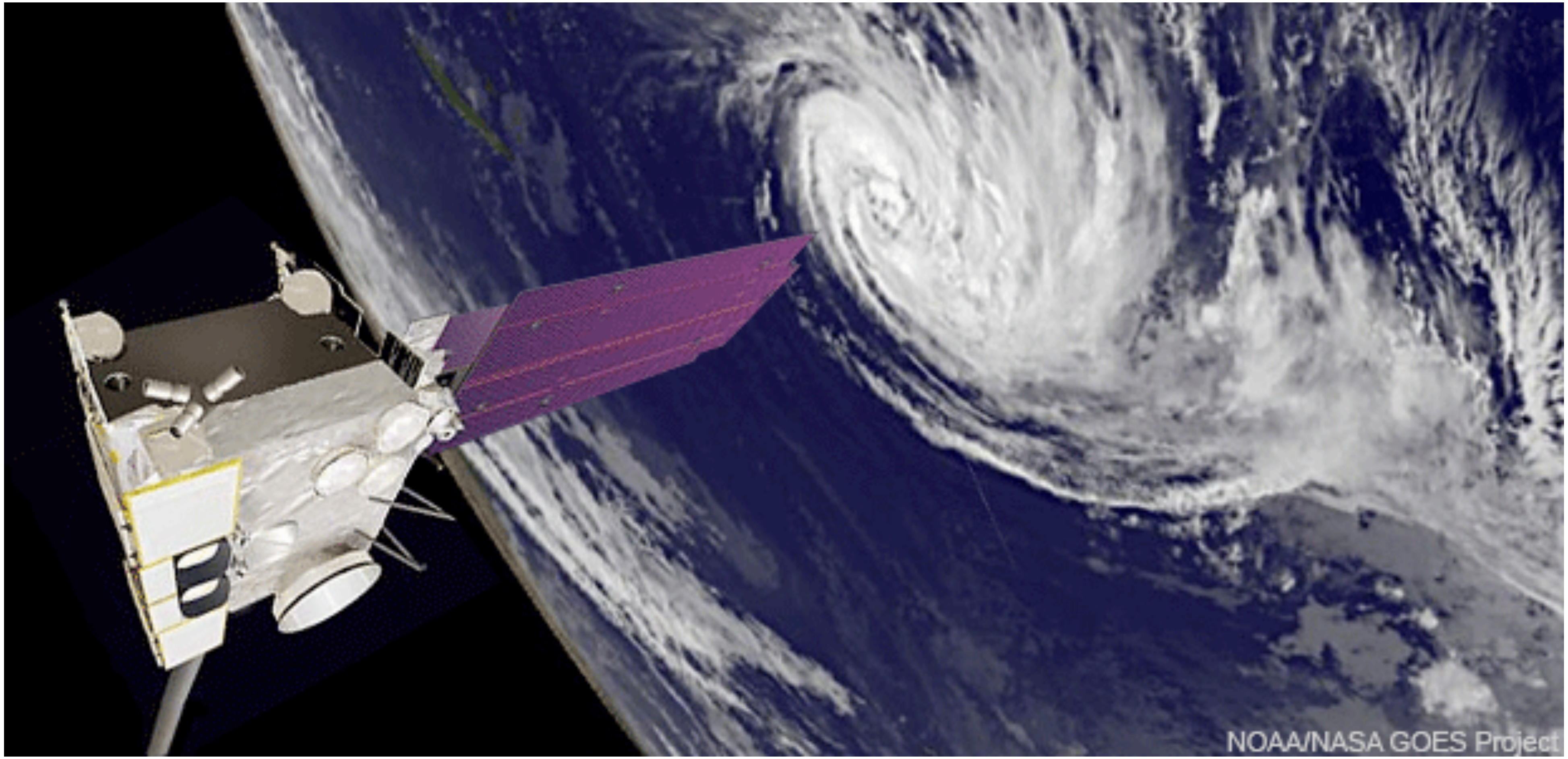


What changed?





What changed?



It's not just climate change. We also have better data collection now.

Python SciPy RV methods

$X \sim \text{Poi}(\lambda) \quad p(k) = e^{-\lambda} \frac{\lambda^k}{k!}$

```
from scipy import stats        # great package
X = stats.poisson(8.5)         # X ~ Poi(λ = 8.5)
X.pmf(2)                       # P(X = 2)
```

Function	Description
`X.pmf(k)`	$P(X = k)$
`X.cdf(k)`	$P(X \leq k)$
`X.mean()`	$E[X]$
`X.var()`	$\text{Var}(X)$
`X.std()`	$\text{SD}(X)$

SciPy reference:
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.poisson.html