08: Poisson and More

Lisa Yan

April 22, 2020

Quick slide reference

3	Poisson	08a_poisson
15	Poisson Paradigm	08b_poisson_paradigm
23	Other Discrete RVs	08c_other_discrete
31	Exercises	LIVE

Poisson

Before we start

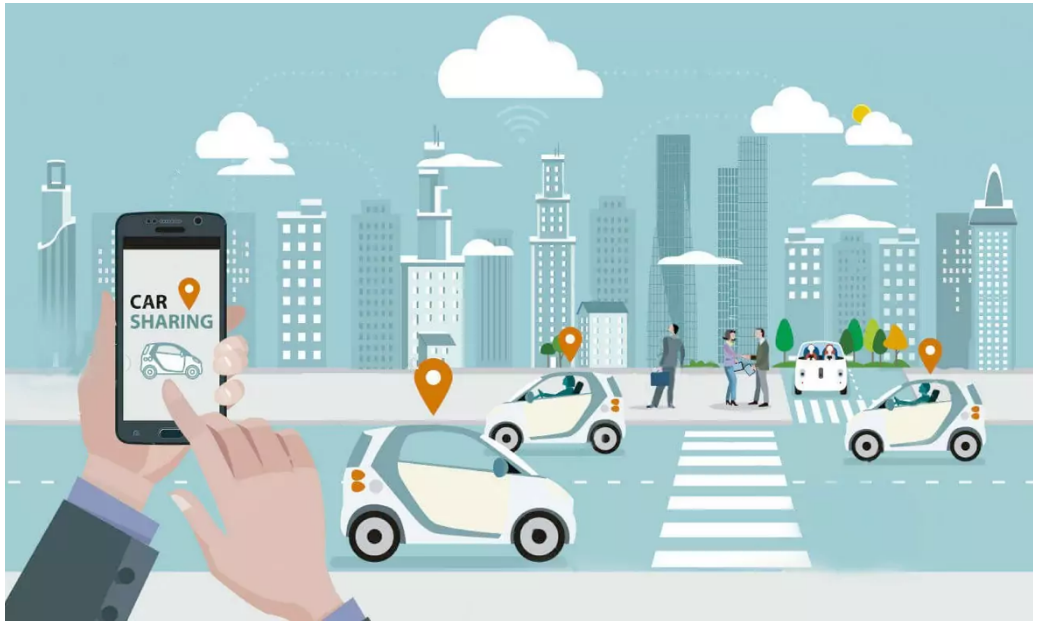
The natural exponent e:

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

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:

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

At each second:

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

Let X = # of requests in minute.

$E[X] = \lambda = 5$

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

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

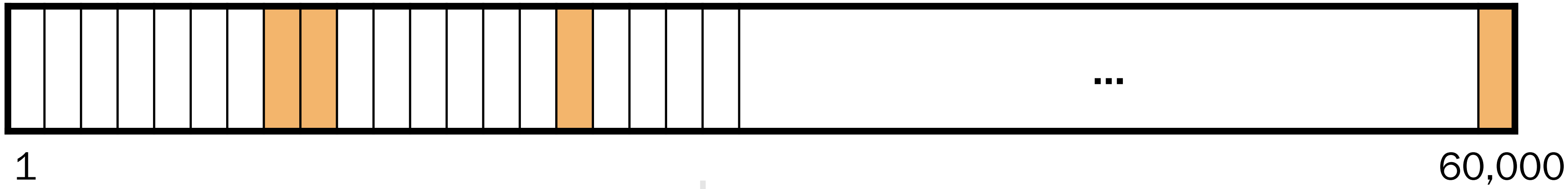
🤔 But what if there are *two* requests in the same second?

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

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

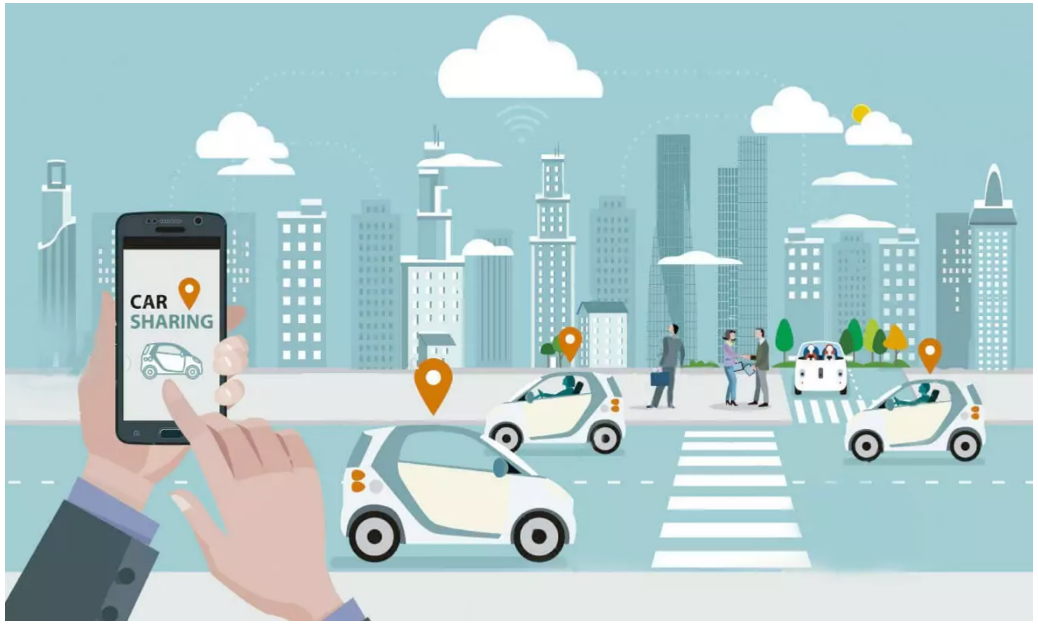
Limit analysis + cancel

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

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.

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

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

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 == variance for Poisson RV! More later.

Earthquakes

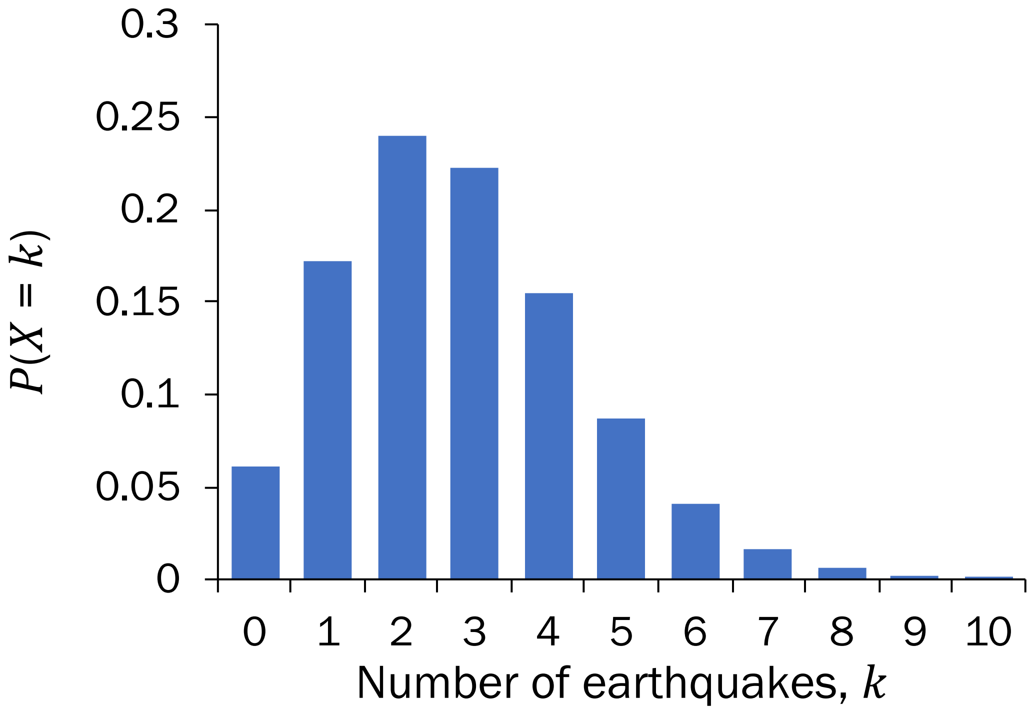
$X \sim \text{Poi}(\lambda)$
$E[X] = \lambda$
$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}$

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

2. Solve



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?

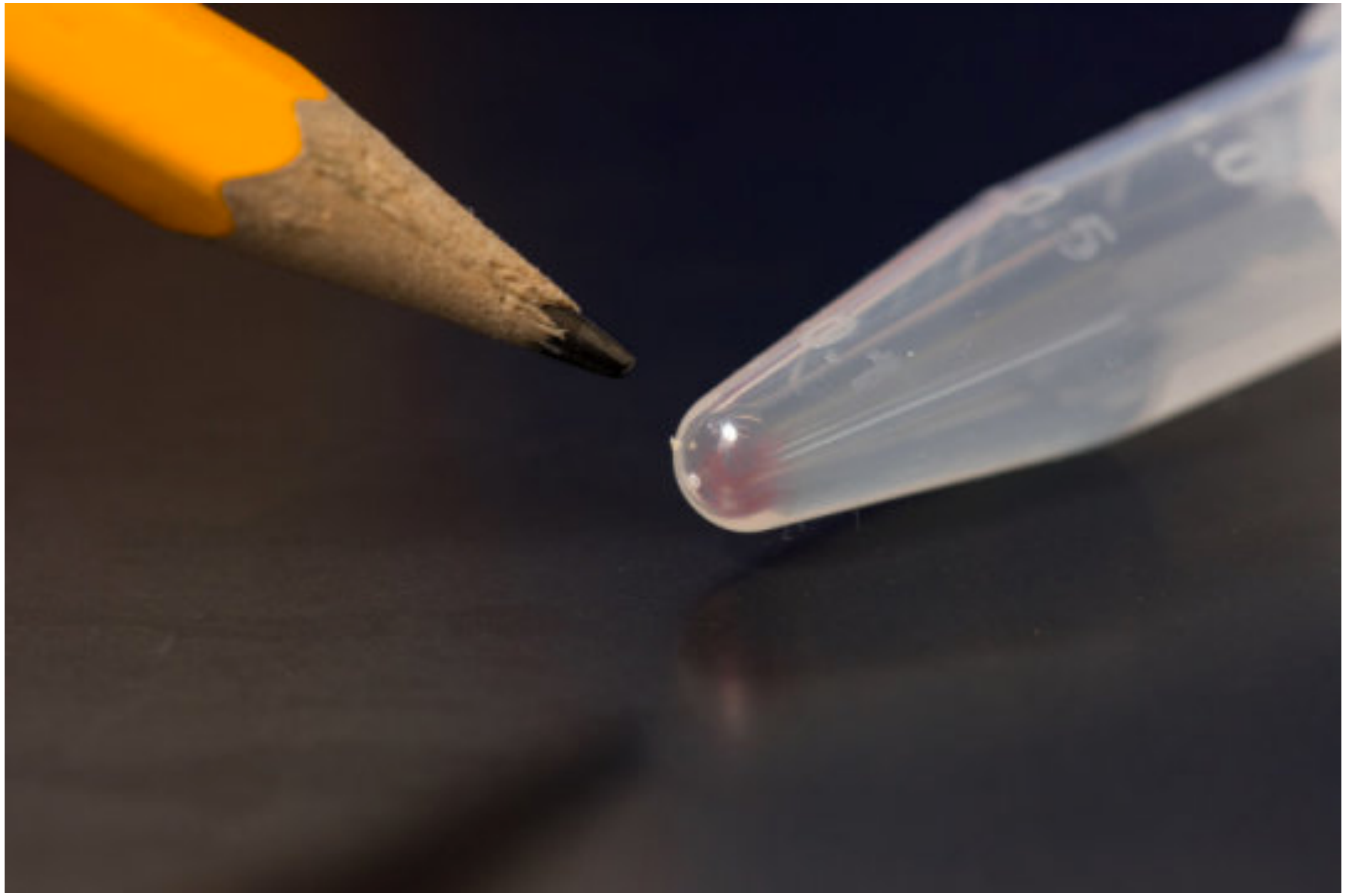
BY J. K. GARDNER and L. KNOPOFF

ABSTRACT

Yes.

Poisson Paradigm

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

≈ 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

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

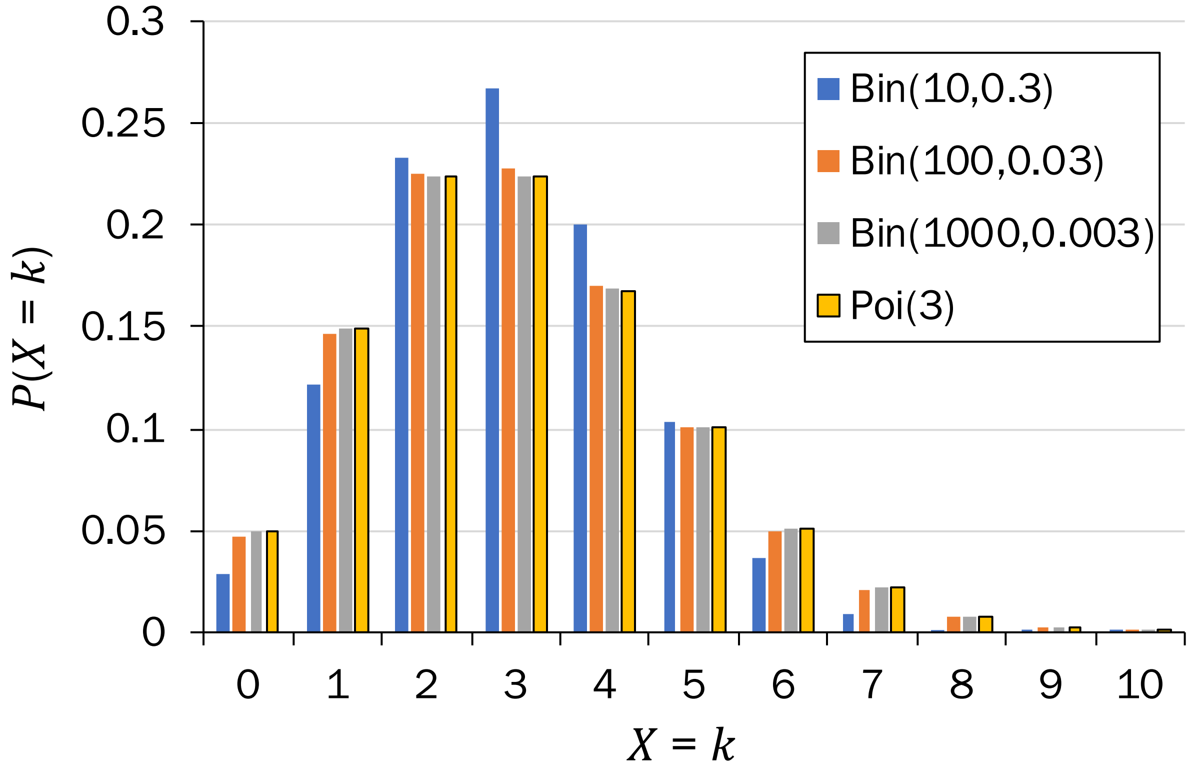
Poisson approximates Binomial when n is large, p is small, and $\lambda = np$ is "moderate."

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.

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.

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

Support: $\{0, 1, 2, \dots\}$

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

Time to show intuition for why expectation == variance!

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

We won't explore this too much, but I want you to know it exists.

Other Discrete RVs

Grid of random variables

	Number of successes	Time until success	
One trial	Ber(p)		One success
Several trials	$n = 1$ ↑ Bin(n, p)		Several successes
Interval of time	Poi(λ)	(tomorrow)	Interval of time to first success

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.

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

$X \sim \text{Geo}(p)$	PMF	$P(X = k) = (1-p)^{k-1}p$
	Expectation	$E[X] = \frac{1}{p}$
Support: $\{1, 2, \ldots\}$	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)

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

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

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

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

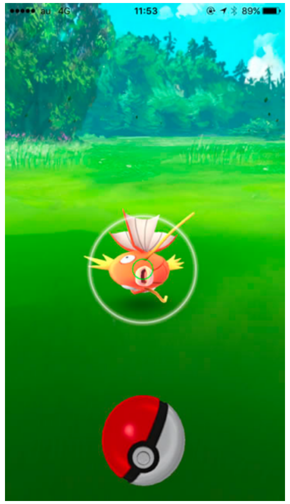
Grid of random variables

	Number of successes	Time until success	
One trial	$\text{Ber}(p)$	$\text{Geo}(p)$	One success
	$n = 1$	$r = 1$	
Several trials	$\text{Bin}(n, p)$	$\text{NegBin}(r, p)$	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

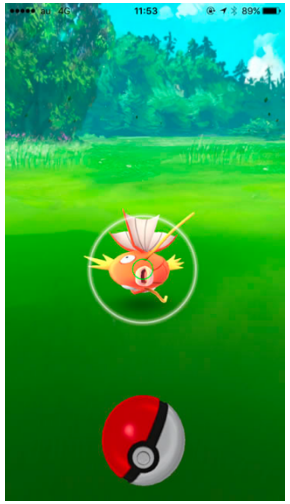
A. $X \sim \text{Bin}(5, 0.1)$
B. $X \sim \text{Poi}(0.5)$
C. $X \sim \text{NegBin}(5, 0.1)$
D. $X \sim \text{NegBin}(1, 0.1)$
E. $X \sim \text{Geo}(0.1)$
F. None/other

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

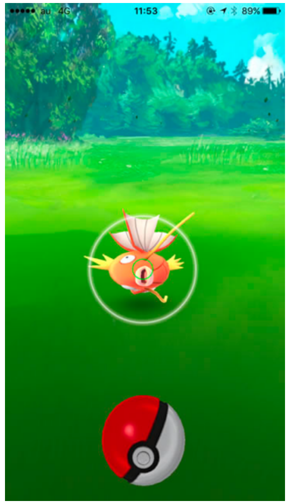
   Want: $P(X = 5)$

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

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
2. Solve

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

Want: $P(X = 5)$

(live)
08: Poisson and More

Lisa Yan

April 22, 2020

LIVE

Discrete RVs

The hardest part of problem-solving is determining what is a random)variable .

Grid of random variables

Review

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

Grid of random variables

Review

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

Breakout Rooms

Check out the question on the next slide (Slide 36). Post any clarifications here!

https://us.edstem.org/courses/109/discussion/39076

Breakout rooms: 5 min. Introduce yourself!

Kickboxing with RVs

Choose from:
A. $\text{Ber}(p)$
B. $\text{Bin}(n, p)$
C. $\text{Poi}(\lambda)$
D. $\text{Geo}(p)$
E. $\text{NegBin}(r, p)$

How would you model the following?

1. # of snapchats you receive in a day
2. # of children until the first one with brown eyes (same parents)
3. Whether stock went up or down in a day
4. # of probability problems you try until you get 5 correct (if you are randomly correct)
5. # of years in some decade with more than 6 Atlantic hurricanes

Kickboxing with RVs

Choose from:
A. $\text{Ber}(p)$
B. $\text{Bin}(n, p)$
C. $\text{Poi}(\lambda)$
D. $\text{Geo}(p)$
E. $\text{NegBin}(r, p)$

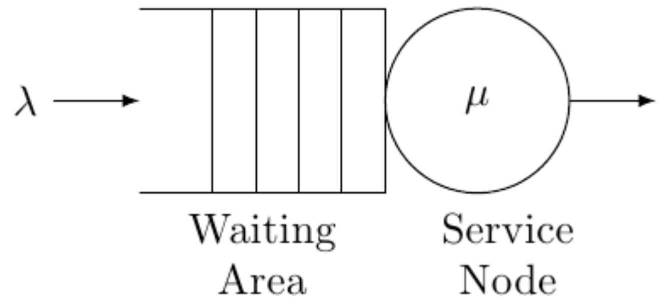
How would you model the following?

1. # of snapchats you receive in a day — C. $\text{Poi}(\lambda)$
2. # of children until the first one with brown eyes (same parents) — D. $\text{Geo}(p)$ or E. $\text{NegBin}(1, p)$
3. Whether stock went up or down in a day — A. $\text{Ber}(p)$ or B. $\text{Bin}(1, p)$
4. # of probability problems you try until you get 5 correct (if you are randomly correct) — E. $\text{NegBin}(r = 5, p)$
5. # of years in some decade with more than 6 Atlantic hurricanes — B. $\text{Bin}(n = 10, p)$, where $p = P(\geq 6 \text{ hurricanes in a year})$ calculated from C. $\text{Poi}(\lambda)$

CS109 Learning Goal: Use new RVs

Let's say you are learning about servers/networks.

You read about the M/D/1 queue:



"The service time busy period is distributed as a Borel with parameter $\mu = 0.2$."

Goal: You can recognize terminology and understand experiment setup.

Poisson RV

Poisson Random Variable

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

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

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

Expectation $\quad E[X] = \lambda$

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

In CS109, a Poisson RV $X \sim \text{Poi}(\lambda)$ most often models

- # of successes over a fixed interval of time.
  $\lambda = E[X]$, average success/interval
- Approximation of $Y \sim \text{Bin}(n, p)$ where n is large and p is small.
  $\lambda = E[Y] = np$
- Approximation of Binomial even when success in trials are not entirely independent.

(explored in problem set 3)

Breakout Rooms

Slide 42 has two questions to go over in groups.

Post any clarifications here!

https://us.edstem.org/courses/109/discussion/39076

Breakout rooms: 5 mins

Web server load

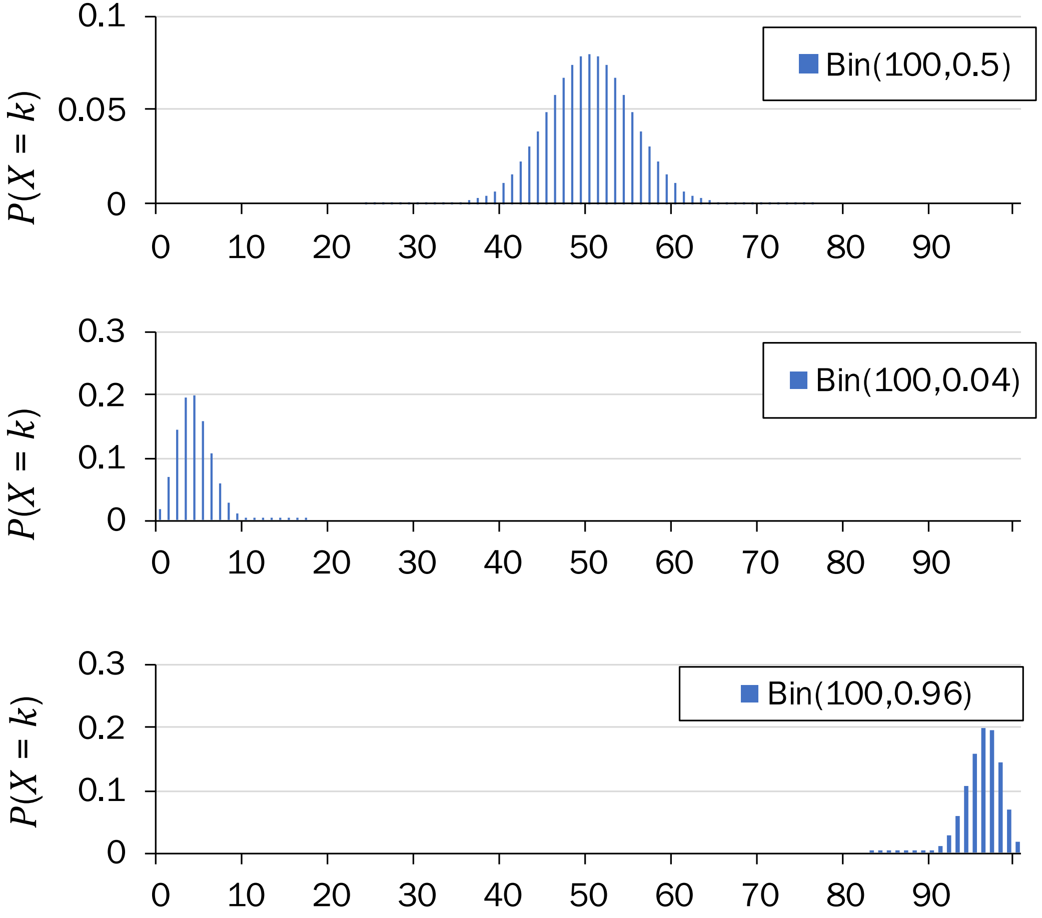
$X \sim \text{Poi}(\lambda)$ $\quad p(k) = e^{-\lambda} \frac{\lambda^k}{k!}$

$E[X] = \lambda$

1. 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 $P(X < 5)$?

2. Can the following Binomial RVs be approximated?



1. Web server load

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

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 $P(X < 5)$?

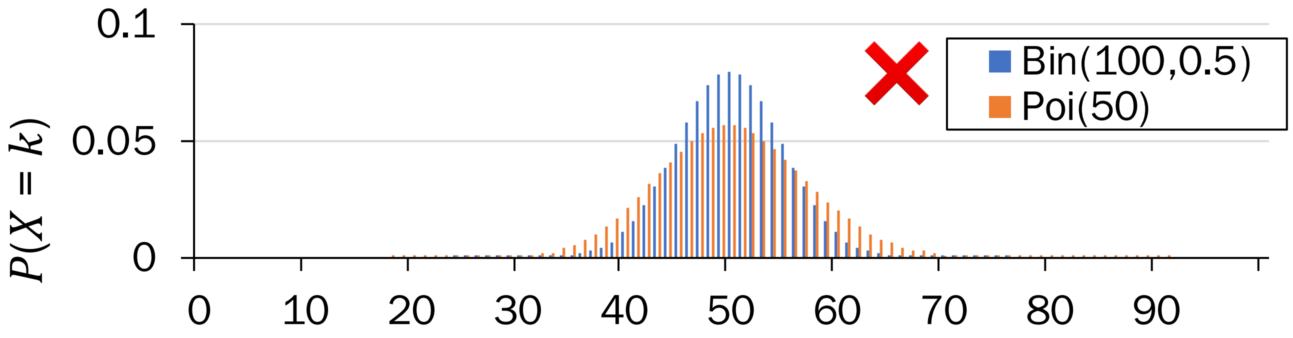
1. Define RVs
2. Solve

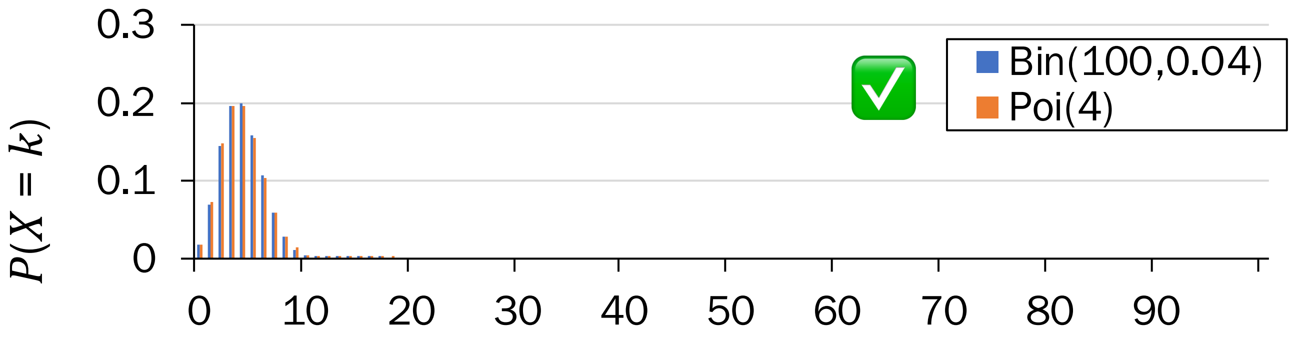
2. Can these Binomial RVs be approximated?

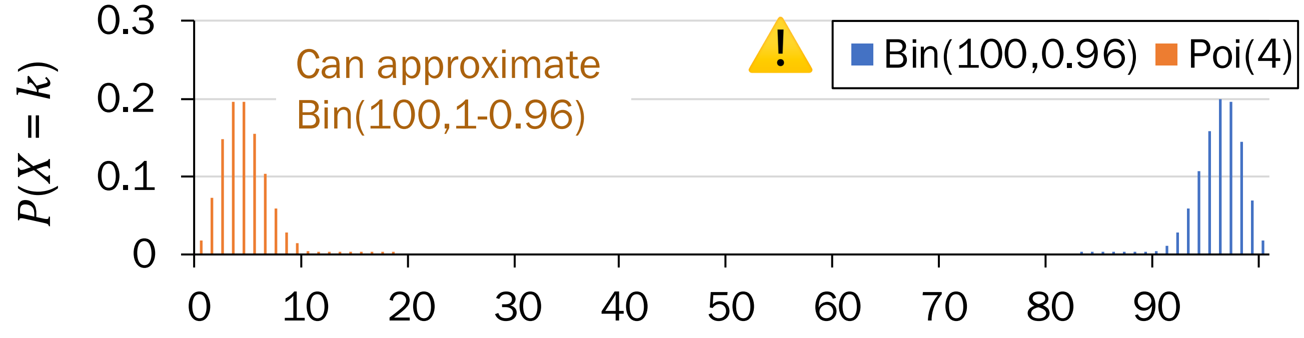
Poisson approximates Binomial when n is large, p is small, and $\lambda = np$ is "moderate."

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 $n \to \infty, p \to 0$







Interlude for jokes/announcements

Announcements

Quiz #1

Time frame: Thursday 4/30 12:00am-11:59pm PT
Covers: Up to end of Week 3 (including Lecture 9)
Tim's Review session: Tuesday 4/28 12-2pm PT
https://stanford.zoom.us/j/92275547392
Info and practice: https://web.stanford.edu/class/cs109/exams/quizzes.html

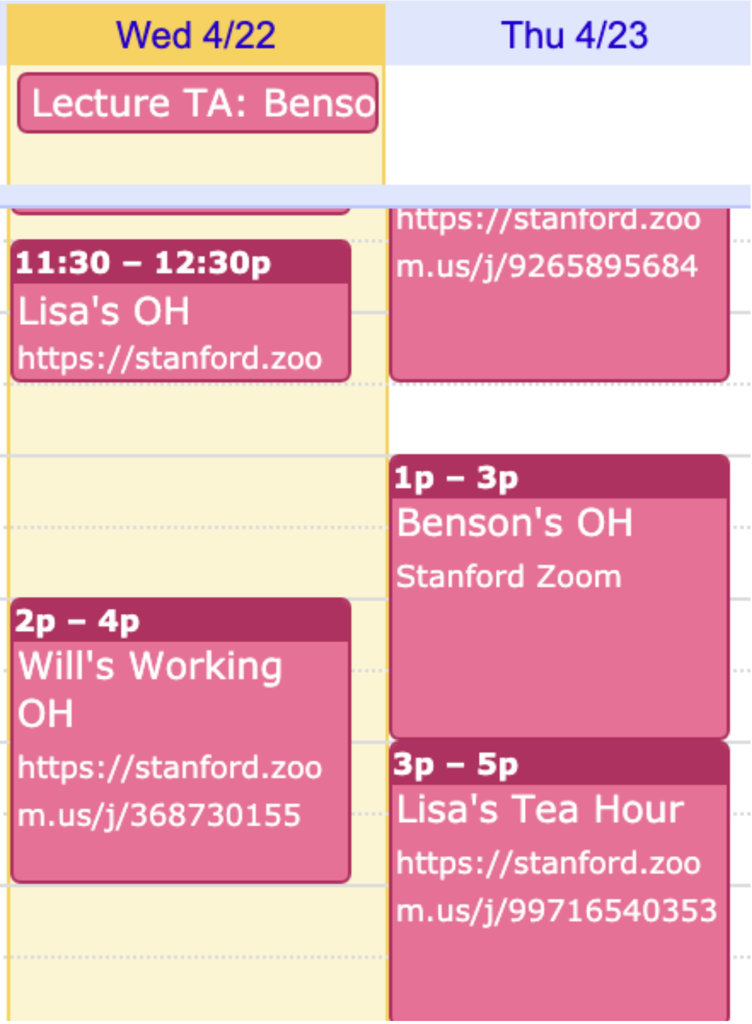
Python tutorial #2

When: Friday 4/24 5:00-6:00PT
Recorded? yes
Notes: to be posted online
Useful for: pset2, pset3

Problem Set 3

Due: Monday 5/8 (after Quiz)
Covers: Up to and including Lecture 11
Out: later today
(Note: early release for quiz practice)

Office Hour update



Working OH

- Sign up on QueueStatus,
- Join the group Zoom

Otherwise, by default:

- Sign up on QueueStatus
- Join 1on1 Zoom when pinged by TA

Lisa's Tea OH (Th 3-5pm PT):

- More casual, any CS109 or non-CS109 questions here
- Collaborate on jigsaw puzzle

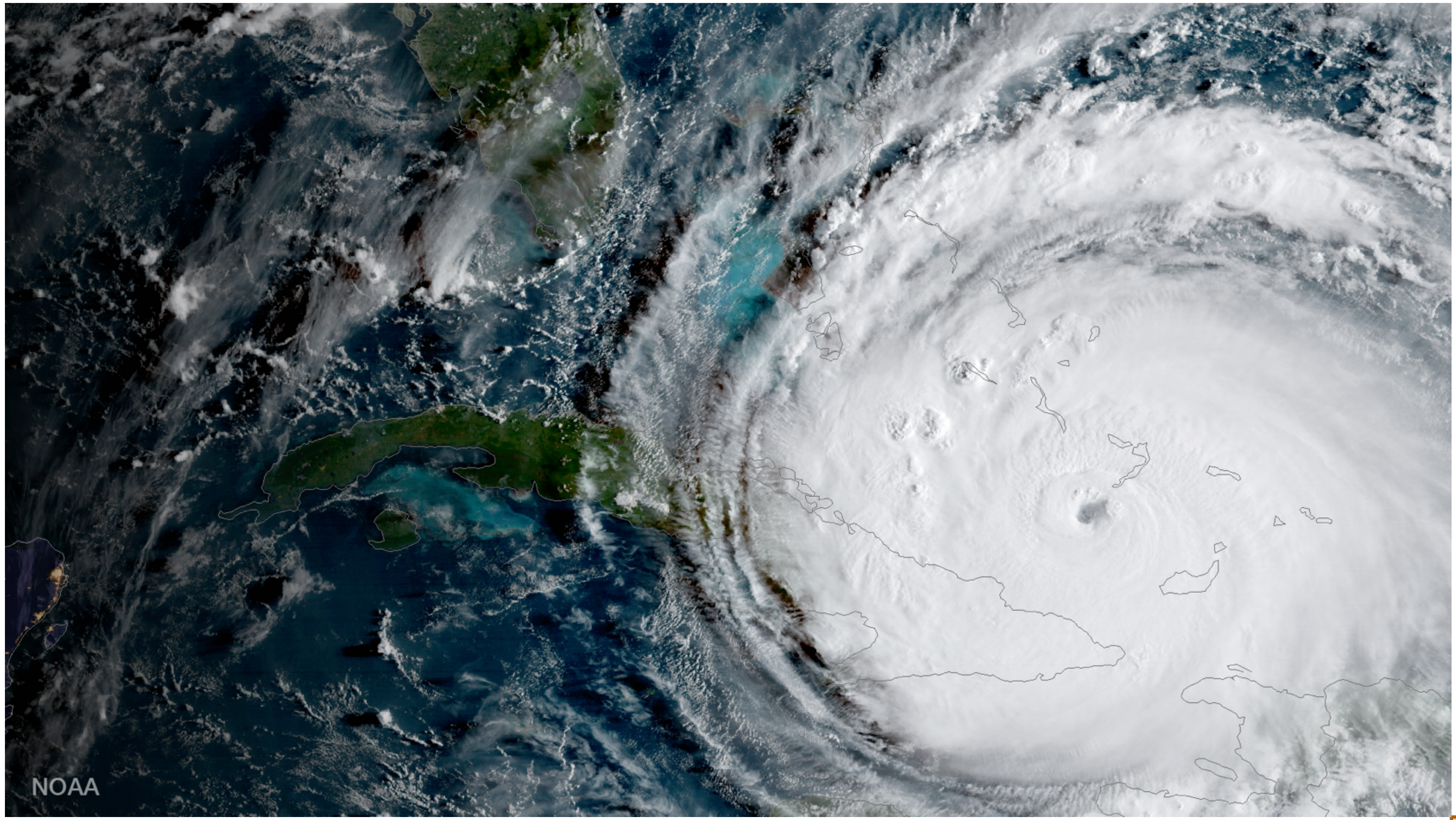
Interesting probability news

https://theconversation.com/polly-knows-probability-this-parrot-can-predict-the-chances-of-something-happening-132767

Find something cool, submit for extra credit on Problem Set #2 ☺

Modeling exercise: Hurricanes

Hurricanes



What is the probability of an extreme weather event?

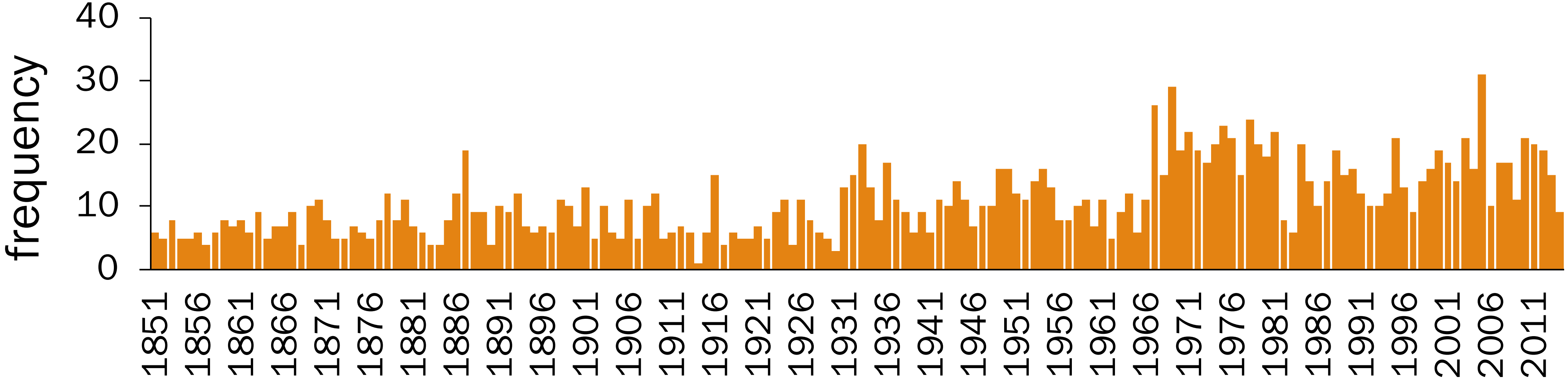
How do we model the number of hurricanes in a season (year)?

1. Graph your distribution.

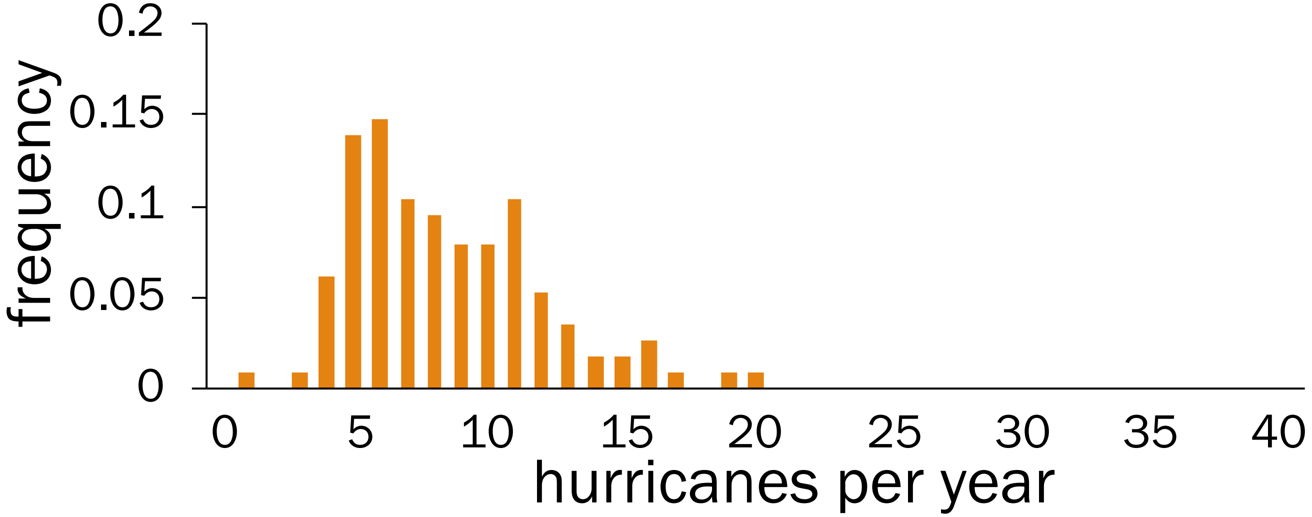
1. Graph: Hurricanes per year since 1851

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

A.



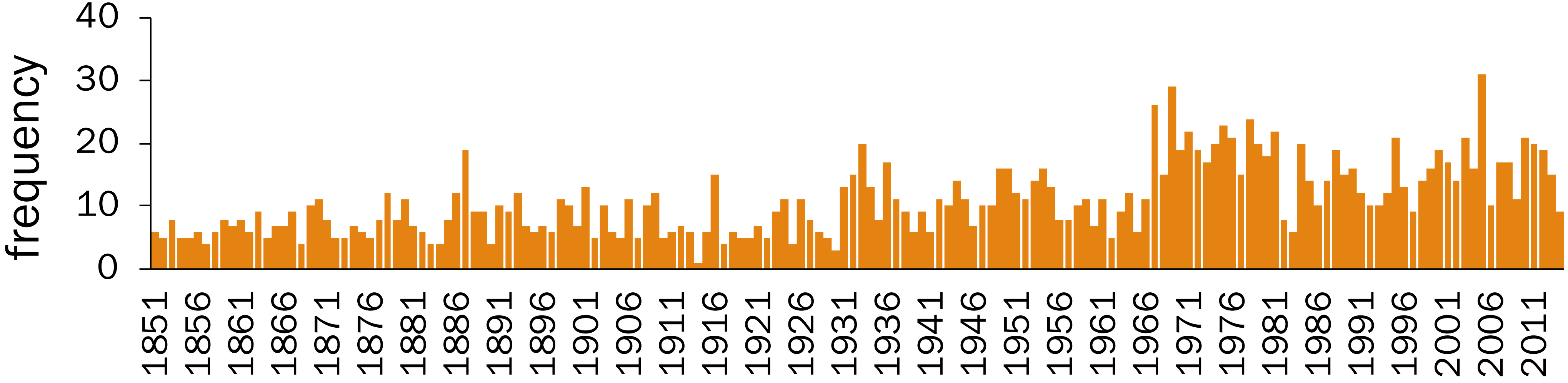
B.



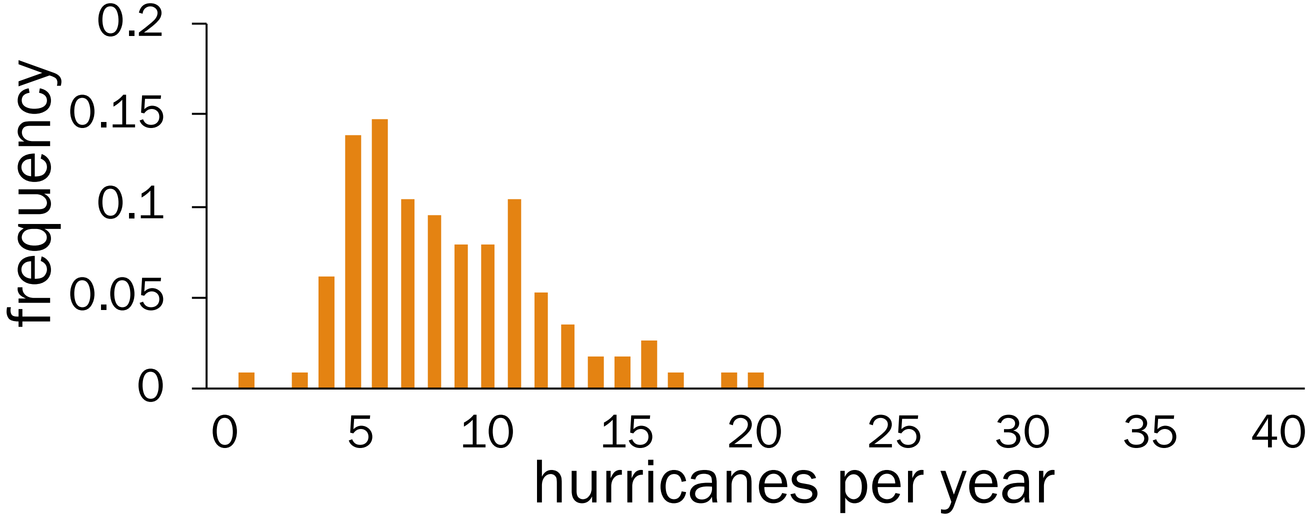
1. Graph: Hurricanes per year since 1851

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

A.

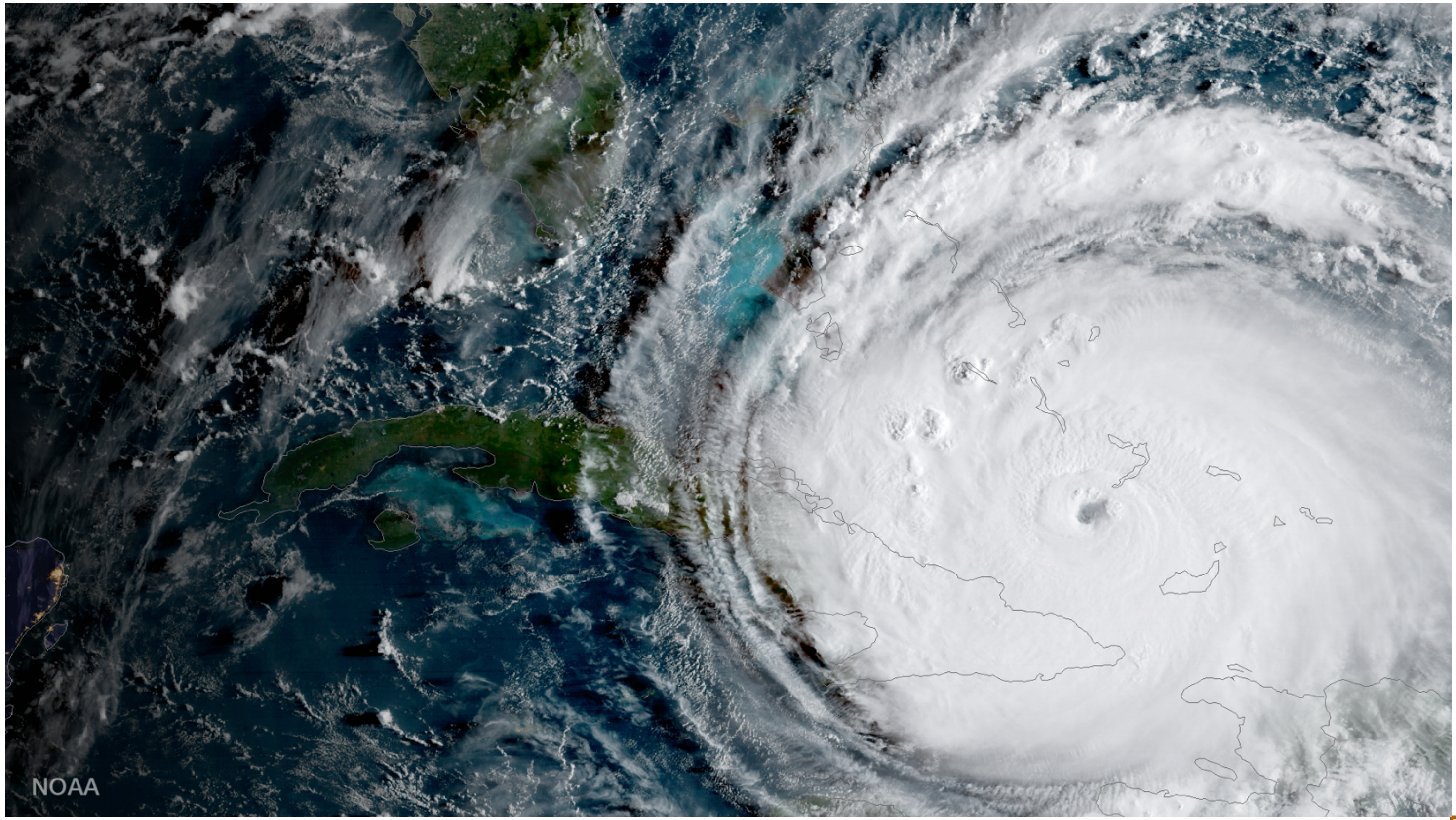


B.



Looks kinda Poissonian!

Hurricanes



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

2. Find a reasonable distribution and compute parameters.

2. Find a distribution: Python SciPy RV methods

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

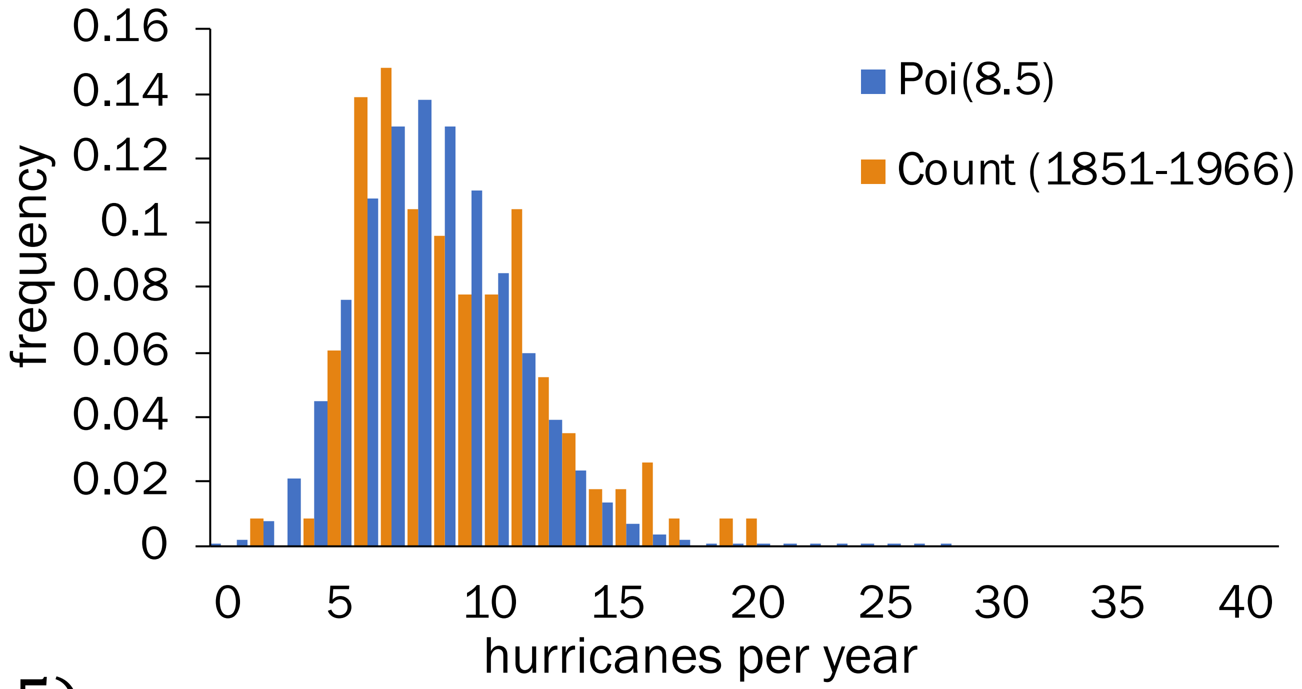
Function	Description
`X.pmf(k)`	$P(X = k)$
`X.cdf(k)`	$P(X \leq k)$
`X.mean()`	$E[X]$
`X.var()`	$\text{Var}(X)$
`X.std()`	$\text{SD}(X)$

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

2. Find a distribution

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?



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

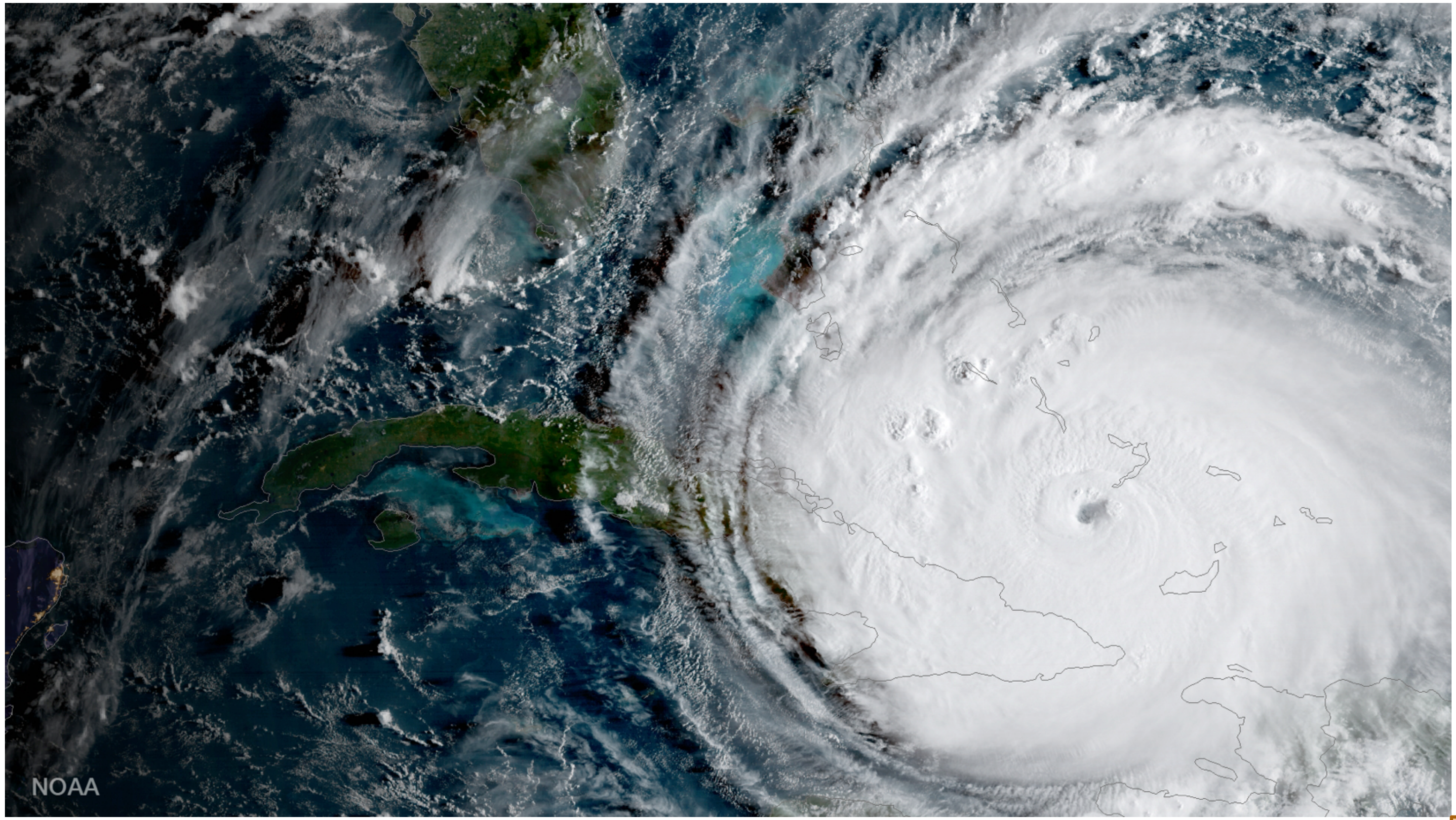
$$= 1 - \sum_{k=0}^{15} P(X = k)$$

$$= 1 - 0.986 = 0.014$$

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

You can calculate this PMF using your favorite programming language. Or Python3.

Hurricanes



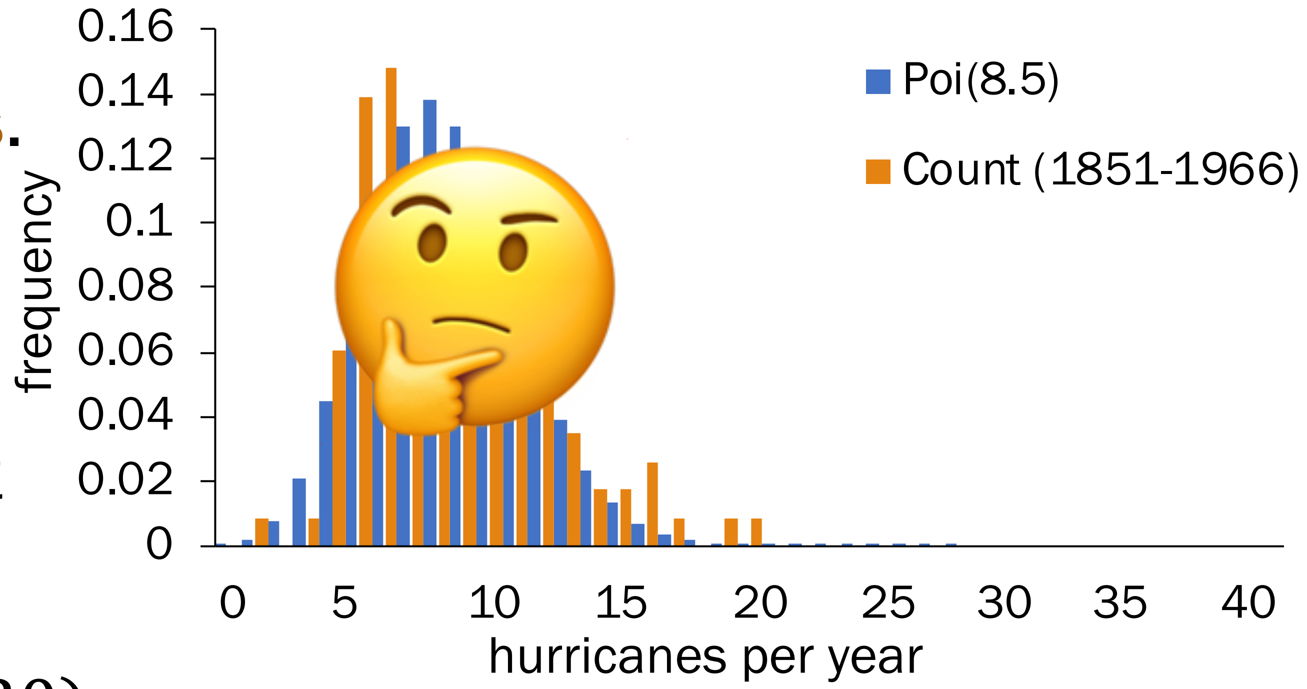
How do we model the number of hurricanes in a season (year)?

3. Identify and explain outliers.

3. Improbability

Since 1966, there have been **two years** with over 30 hurricanes.

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



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

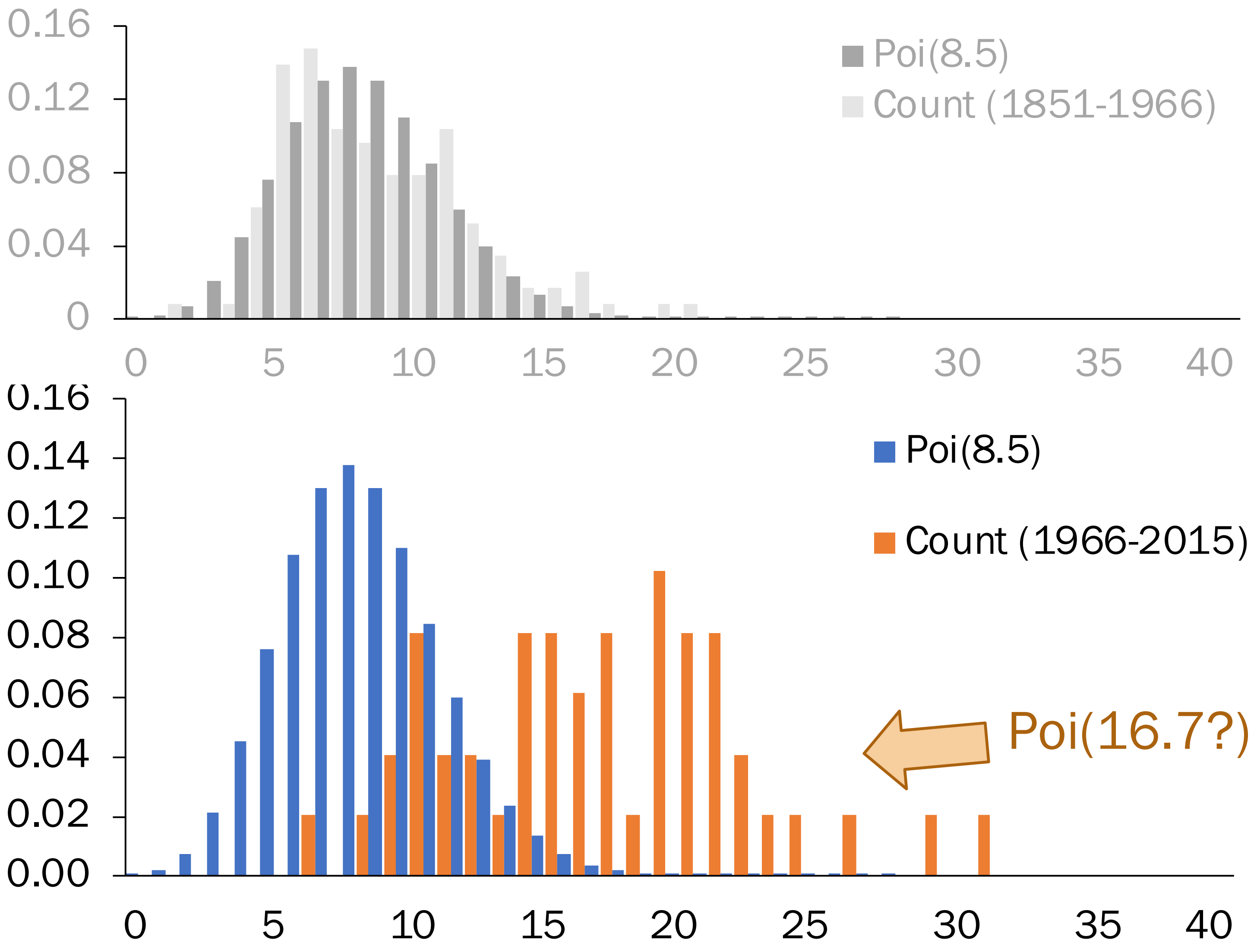
$$= 1 - \sum_{k=0}^{30} P(X = k) \qquad X \sim \text{Poi}(\lambda = 8.5)$$

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

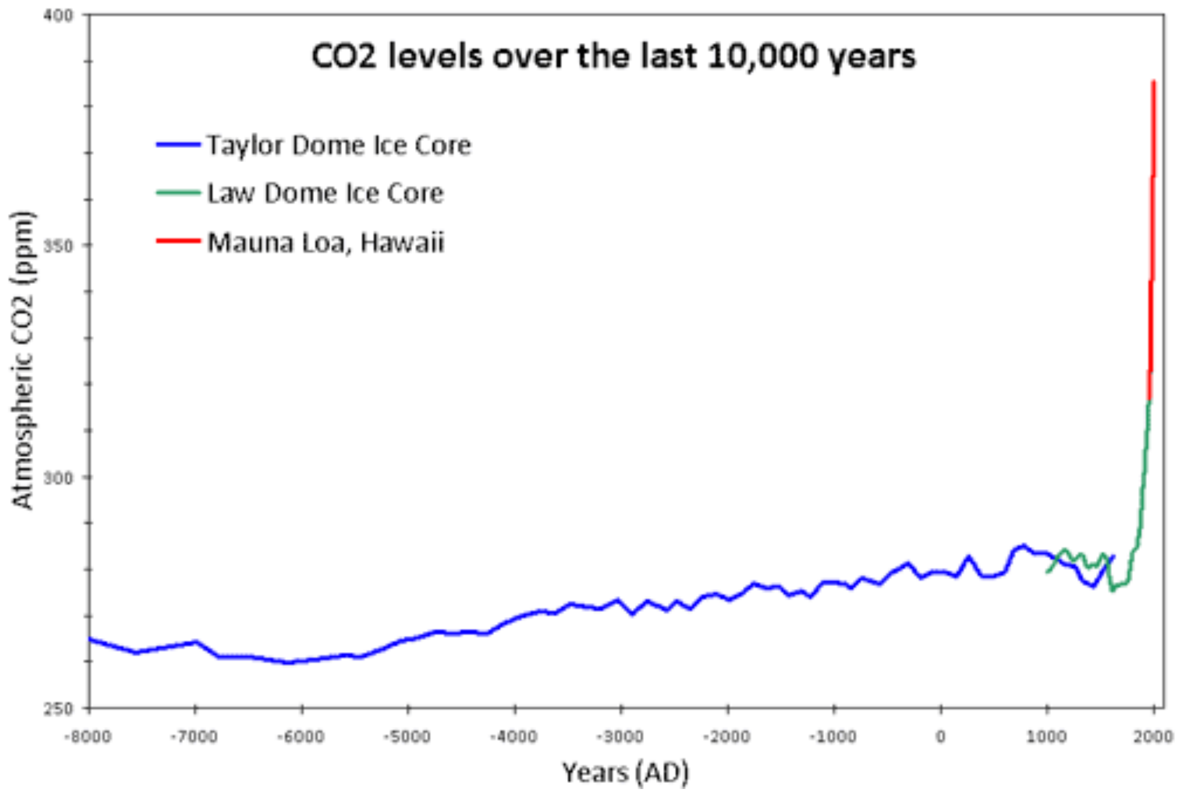
3. The distribution has changed.

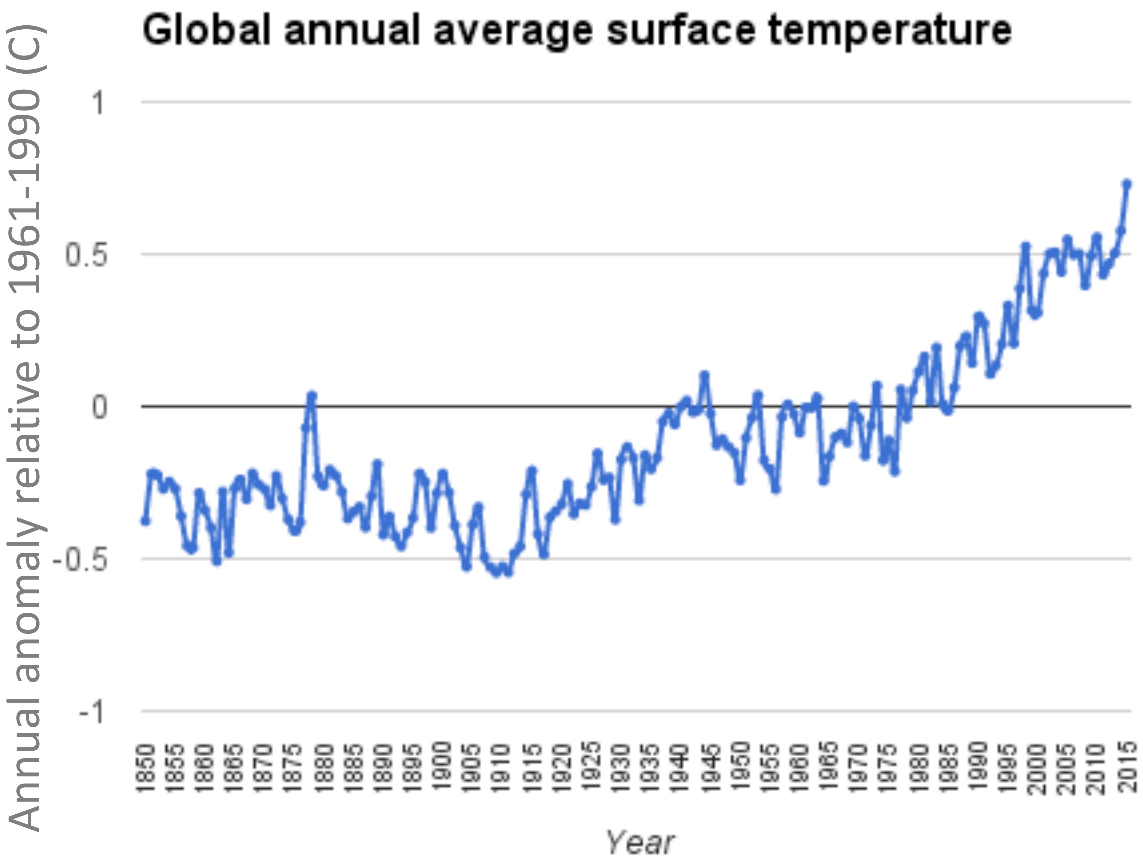
1851–
1966

Since
1966

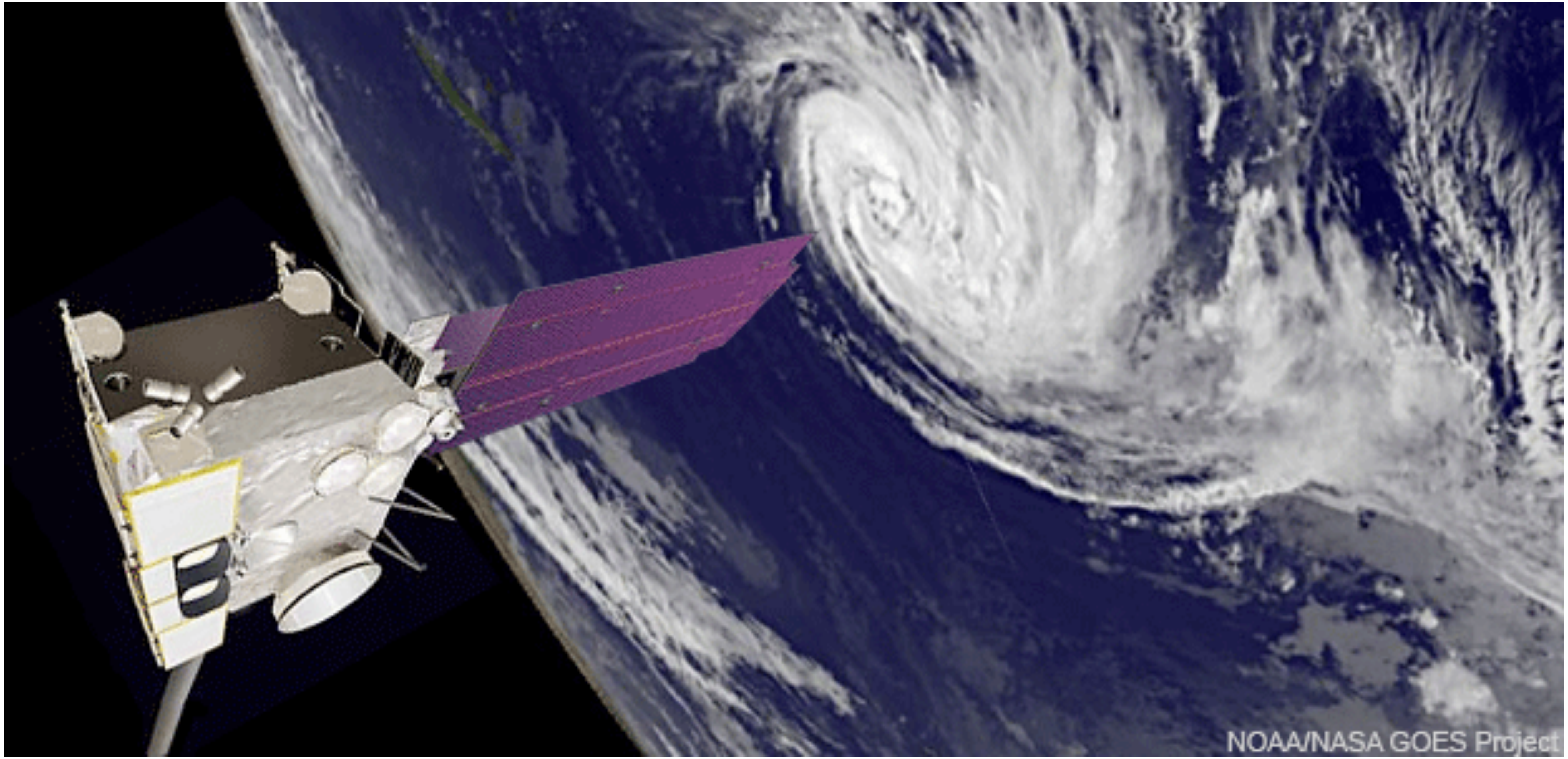


3. What changed?





3. What changed?



It’s not just climate change. We also have tools for better data collection.