Poisson Random Variable

Expected # of autos in an hour = $\underline{10}$

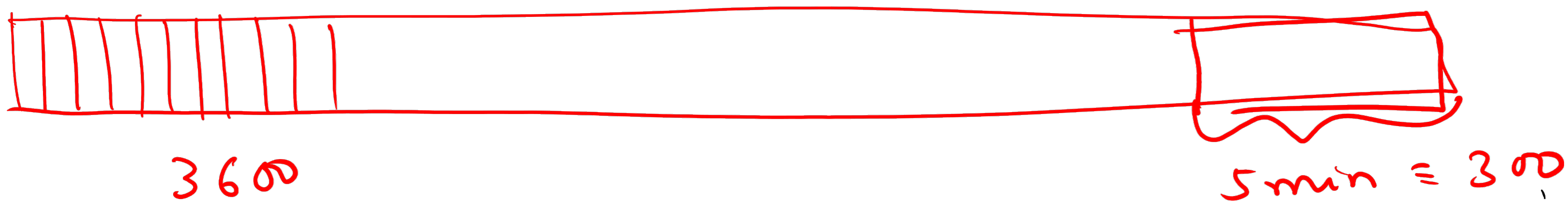
Time interval = 5 minutes.

Probability that one auto will arrive in the next 5 minutes.

improve approximation

we are calculating the probability that an auto will arrive in next 55 minutes only and that no auto arrives in first 5 minute then subtract it from total prob gives us probability that at-least one auto will arrive in next 5 minutes.

$$\approx 1 - \frac{\binom{55}{10}}{\binom{60}{10}} \quad \Bigg| \quad 1 - \frac{\binom{3600 - 5\times 60}{10}}{\binom{3600}{10}}$$



$p = \dfrac{10}{3600}$ $\qquad X$ # of autos in

$n = 300$

$P(X \geq 1) = 1 - P(X = 0)$

at-least an auto arrives. $\qquad$ No auto arrives

$$= 1 - \binom{300}{k}(1-p)^{300-k}\, p^{k}, \quad k = 0$$

$$= 1 - (1-p)^{300} = 1 - \left(1 - \frac{10}{3600}\right)^{300}.$$

$$\approx 0.57$$

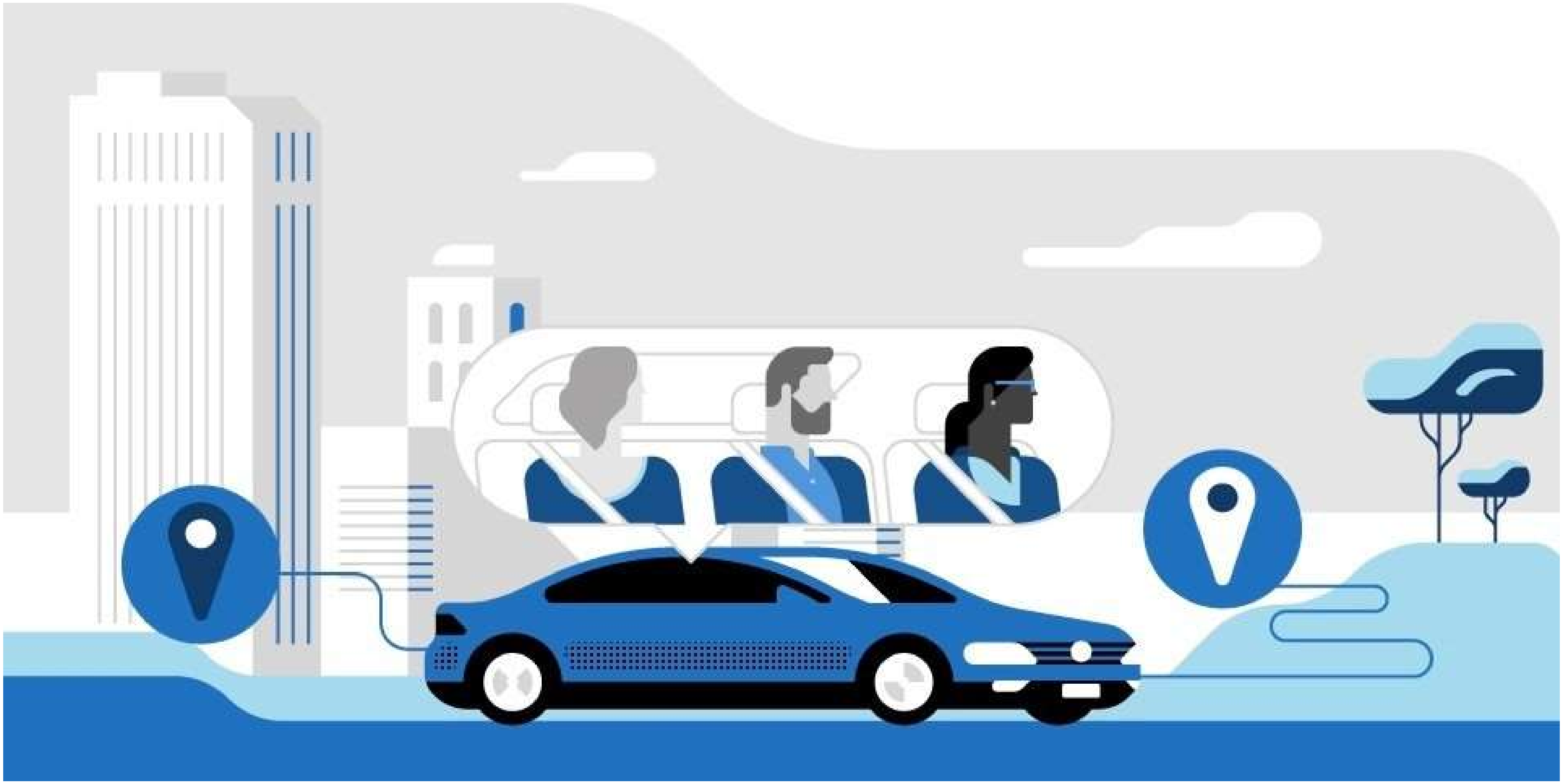
Situations from Poisson R.V is useful

- In all four discrete R.V.s so far (Bernoulli, Binomial, Geometric, Negative binomial), we were counting some outcome from a set of possible discrete options.
  - Multiple dice rolls
  - Servers in operation
  - View of ads on YouTube.

- In many real-life applications, the substrate is continuous, example time.
  - We are counting outcomes of interest in this continuous space.

We want outcome of interest in this continuous space.

Case Study: Ride Sharing Apps



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

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

On average, λ = 5 requests per minute

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

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

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

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 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

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

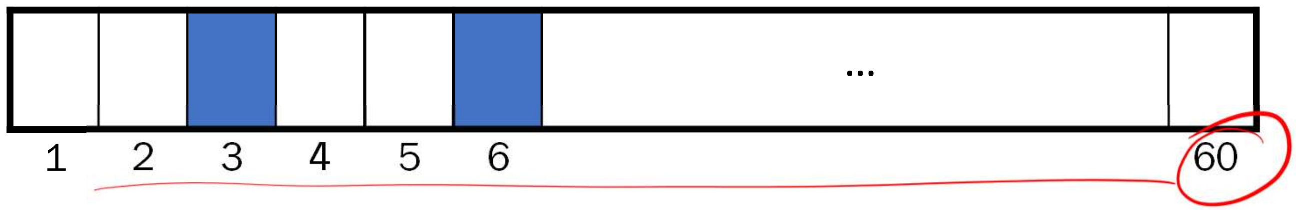
$$p = \frac{\lambda}{n}$$

Probability depending on number of bins.

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

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, $\lambda = 5$ requests per minute

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

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

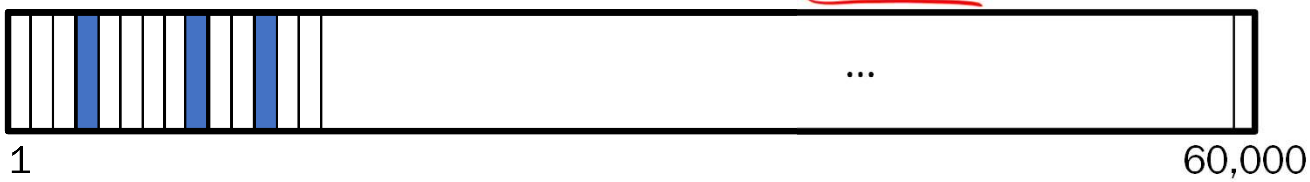
$$P(X = 3) = \binom{60}{3} (5/60)^3 (1 - 5/60)^{57}$$

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

we increase the bins, and increase it till infinity.

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

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

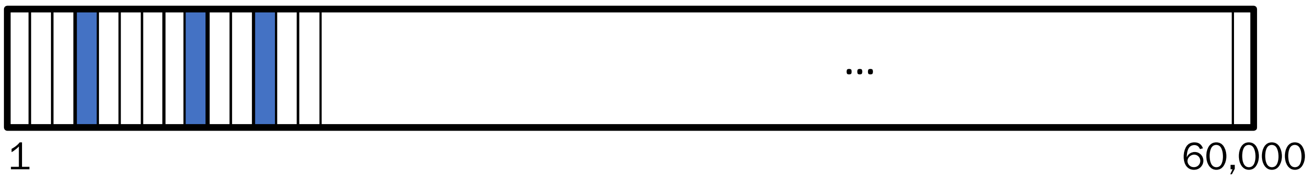


At each ms, you either get a request or don't.
Let X be the number of requests in a minute.

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

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

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

$$P(X = k) = \binom{60000}{k}\left(\frac{5}{60000}\right)^k\left(1 - \frac{5}{60000}\right)^{n-k}$$

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

In each bucket, 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

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

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

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

Probability of *k* Requests From This Area Each Minute

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

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

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

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

The Poisson Random Variable

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

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

PMF:

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

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."*

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.

You can say independence is the very desired thing everyone wants!

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.

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

It should come at the rate of lambda.

Is Lambda All You Need? Yes

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

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

Calculate $E[X]$, $\text{Var}(X)$

First calculate moment generating function of X.

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: $E[X] = \lambda$

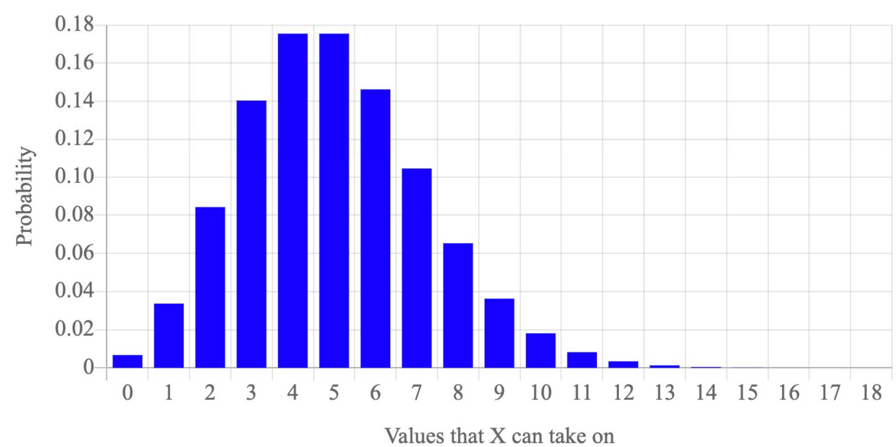
Variance: $\text{Var}(X) = \lambda$

np

np(1−p)

PMF graph:

Parameter λ: 5



t

Moment Generating Function

Expected value of a special function that will make it easy to calculate mean and variance of several random variables.

Let X be a random variable, and P(X) be its pmf or density function.

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

Let $g(X) = e^{tX}$, $\phi(t) = E[e^{tX}]$ $= \sum_{x \in X} e^{tx} P(x)$

For many special random variables, $\phi(t)$ can be calculated in closed form.

Moment Generating Function

$\underline{\phi(t)}$

$$E_p(X) = \sum_x x\, p(x)$$

$$\phi(t) = \sum_x e^{tx} p(x)$$

$$\frac{\partial \phi(t)}{\partial t} = \frac{\partial}{\partial t} \sum_x e^{tx} p(x) = \sum_x x\, e^{tx} p(x) \Big|_{t=0} = \sum_x x\, p(x)$$

$$= E(X)$$

$$\phi'(t)\Big|_{t=0} = E(X)$$

$$\boxed{\phi''(t)\Big|_{t=0} = E(X^2)}$$

MGF of Poisson distribution

$$\phi(t) = \sum_x e^{tx} \frac{\lambda^x}{x!} e^{-\lambda}$$

$$= e^{-\lambda} \boxed{\sum_{x=0}^{\infty} \frac{(e^t \lambda)^x}{x!}} = e^{-\lambda} e^{e^t \lambda}$$

$$\phi(t) = e^{-\lambda(1-e^t)}$$

$$\phi'(t) = e^{-\lambda(1-e^t)} \cdot \lambda e^t \Big|_{t=0}$$

$$= \lambda$$

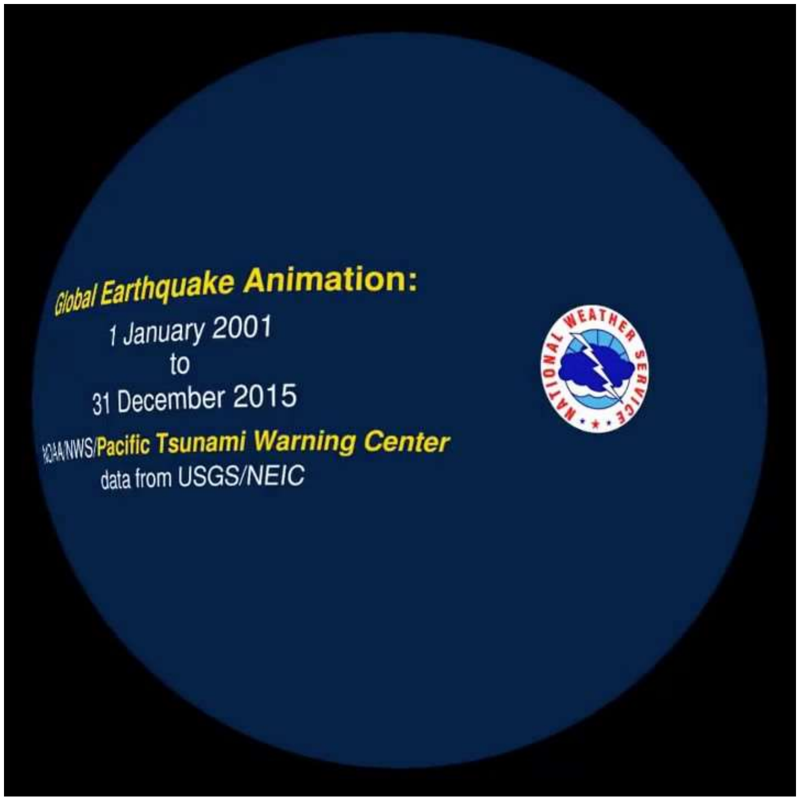
$$\phi''(t) = \frac{\partial}{\partial t}\left(\lambda e^{-\lambda+\lambda e^t + t}\right) = \lambda e^{-\lambda+\lambda e^t + t} (\lambda e^t + 1)\Big|_{t=0}$$

$$= \lambda + \lambda^2 = E(X^2)$$

$$Var(X) = E(X^2) - [E(X)]^2$$

$$= \lambda$$

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.

Earthquakes

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

Here's the PMF for X:

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

X is a Poisson!

What is E[X] (11)?

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

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

This one here is the discrete one!

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 $P(X < 5)$?

$\lambda = \frac{120}{60} = 2$ average # of requests per second

$$P(x < 5) = \sum_{x=0}^{4} \frac{e^{-\lambda} \lambda^{x}}{x!}$$

Practice: Web Server Load

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

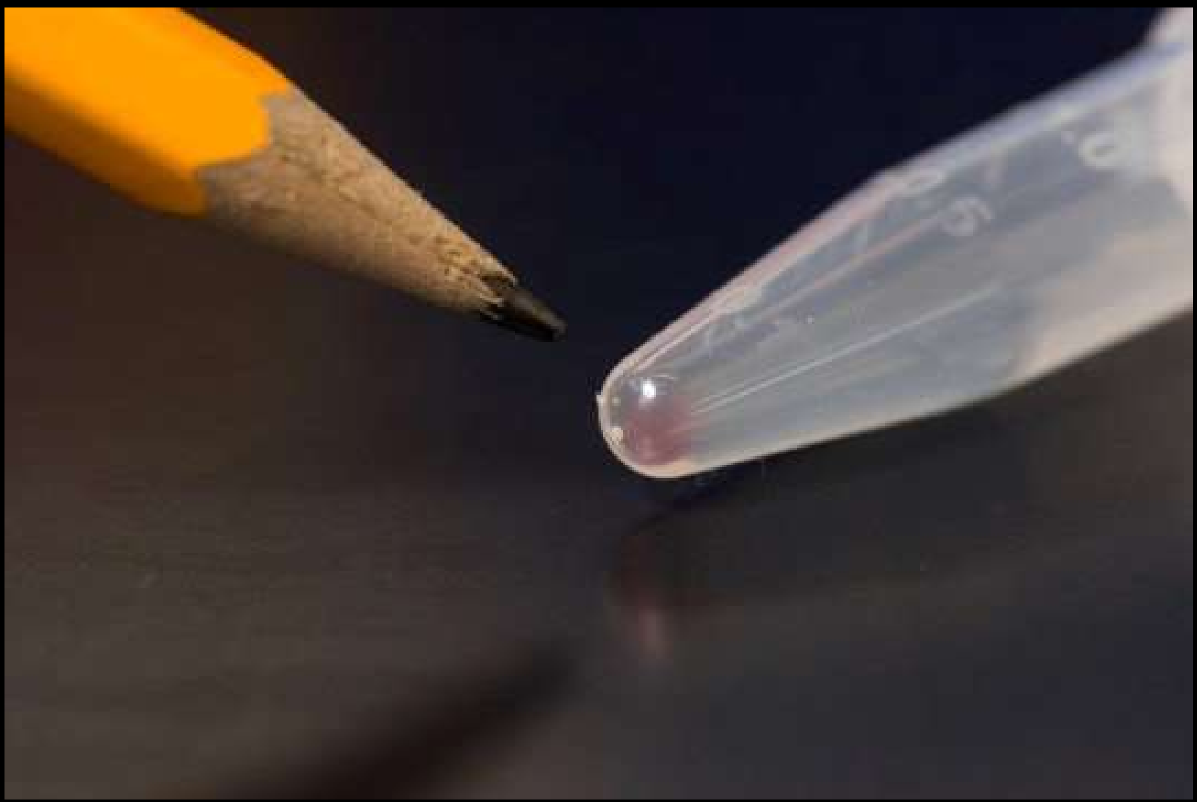
Let X be the number of hits this server receives in a **second**. What is $P(X < 5)$?

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

The Poisson approximates the Binomial when n is large

Too much data can be stored into small chunk if we could develop such a technology!

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.

https://en.wikipedia.org/wiki/DNA_digital_data_storage#:~:text=DNA%20digital%20data%20storage%20is,slow%20read%20and%20write%20times.

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.

$$\sum_{x=0}^{10^6-1} \binom{n}{x} p^{x} (1\ p)^{n-x} \quad = \quad X \sim \text{Bin}(\underbrace{10^8}_{n}, 10^{-6})$$

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

lambda = 100, poison distribution with bla bla...

Approximating with Poisson

When the number of possible events is very large probability is very less, then we can assume continuity over discrete for such a large data.

Let X be the number of corrupted positions.

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

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

lambda = n*p

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 $n \rightarrow \infty$ and $p \rightarrow 0$, where $np = \lambda$

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