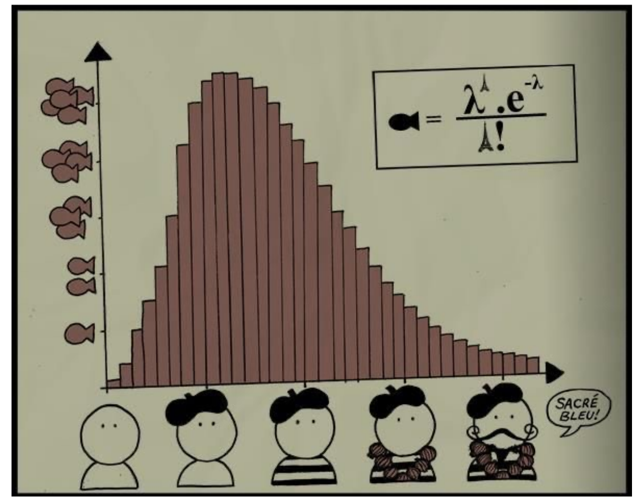
08: Poisson and More

Jerry Cain

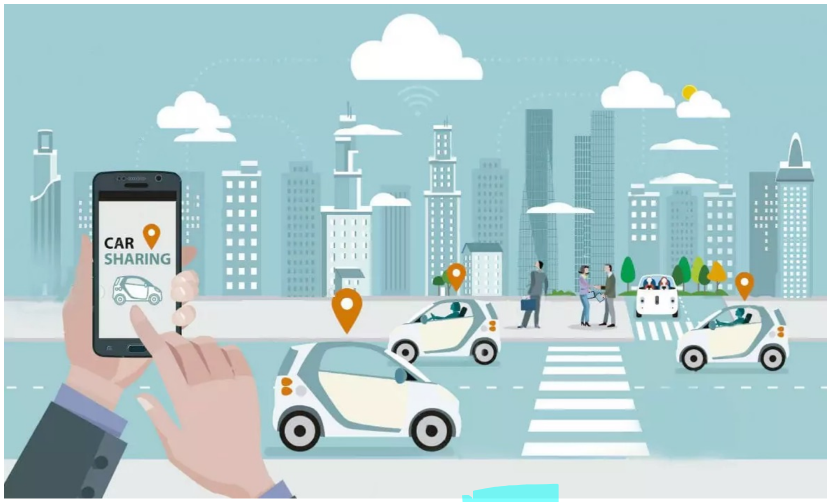
April 17th, 2024

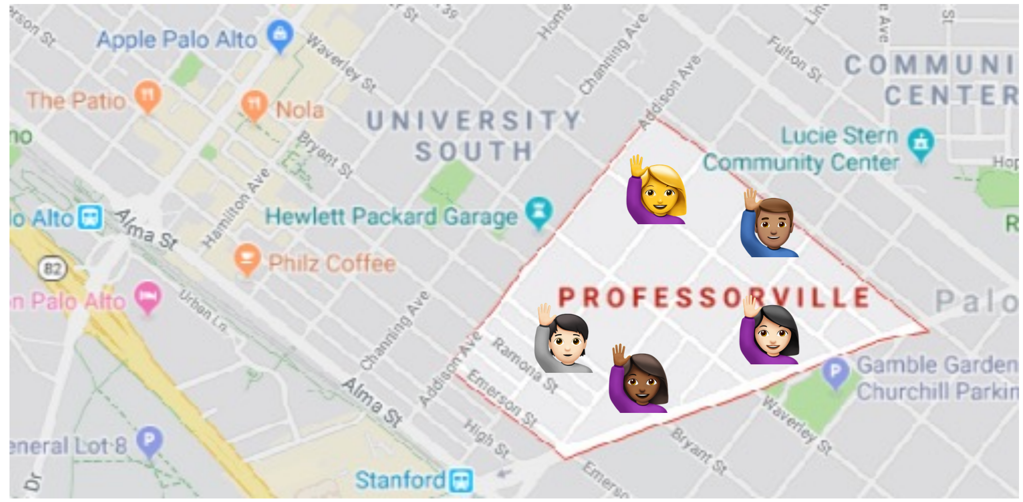
Lecture Discussion on Ed

Poisson



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

k can, in theory, be any nonnegative integer.

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

Let X = # of requests in minute.

$E[X] = \lambda = 5$

recall: if $X \sim \text{Bin}(n, p)$, then $E[X] = np$

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

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

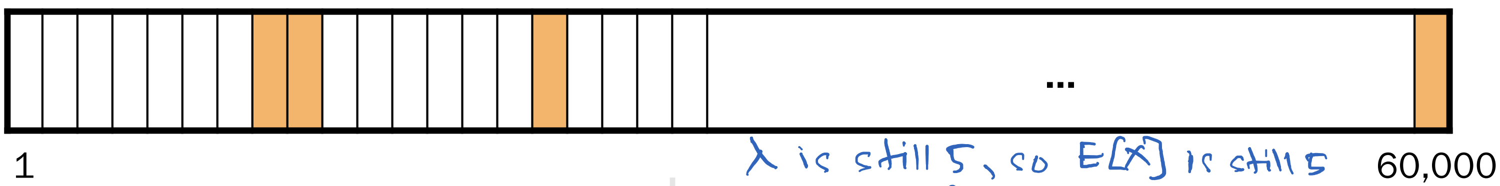
But what if there are *two* requests in the same second?

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

Let X = # of requests in minute.

$E[X] = \lambda = 5$

λ is still 5, so E[X] is still 5 as well. If n = 60,000 now, then p must be 5/60000

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

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

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

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

Let $X = \#$ of requests in minute.

$E[X] = \lambda = 5$

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

still modelling as a Binomial, but n is now arbitrarily large.

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

Gnarly math incoming!

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

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

these cancel

this is p, which approaches 0 as n approaches ∞.

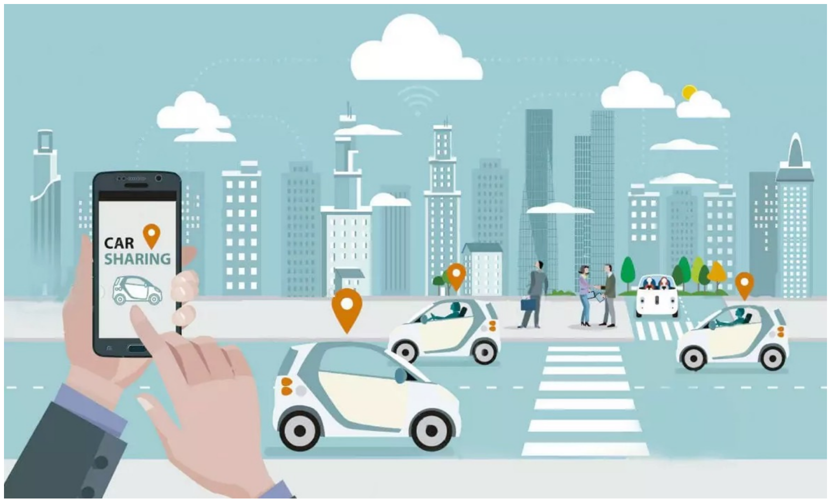
Limit analysis + cancel

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

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

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

Poisson distribution

Poisson Random Variable

key constraint.

when a problem statement is framed in terms of a fixed amount of time, you should think of Poisson or something similar.

Consider an experiment that lasts a fixed interval of time.

<u>def</u> A Poisson random variable X is the number of successes over the experiment duration, assuming **the time that each success occurs is independent** and the average # of requests over time is constant.

$X \sim \text{Poi}(\lambda)$	PMF	$P(X = k) = e^{-\lambda} \dfrac{\lambda^k}{k!}$
	Expectation	$E[X] = \lambda$
Support: $\{0, 1, 2, \dots\}$	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, and major earthquakes occur independently.

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

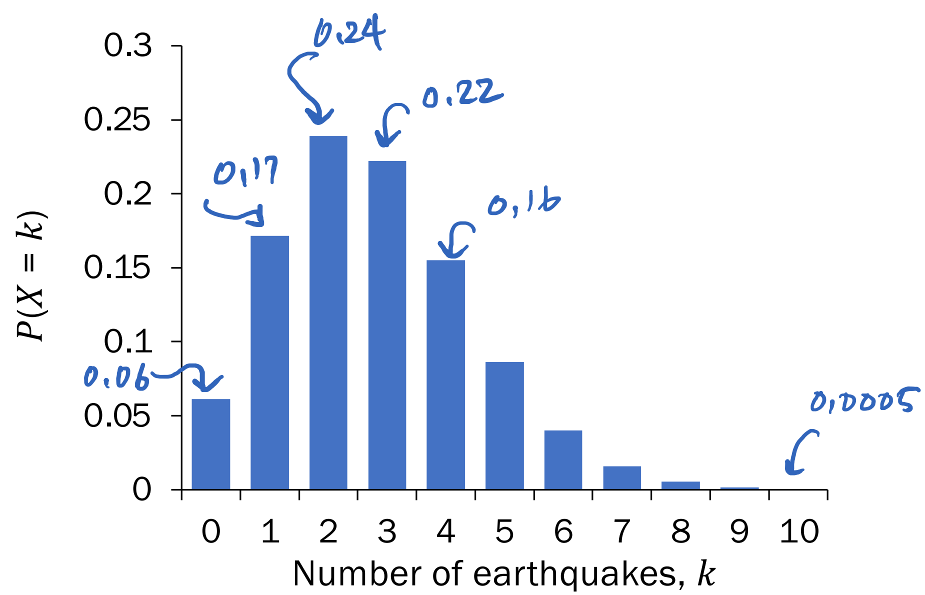
1. Define RVs

$X \sim Poi(\lambda = 2{,}79)$

units of λ is events per time period

2. Solve

$P(X=3) = e^{-2{,}79} \frac{2{,}79^3}{3!} \approx \boxed{0{,}22}$



Other Discrete RVs

Grid of random variables

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

Geometric RV

consider a fair coin: H: $P(X=1) = 0.5$
TH: $P(X=2) = 0.25$
TTH: $P(X=3) = 0.125$
TTTH: $P(X=4) = 0.0675$ etc.

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

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

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

need at least r trials if we need r successes

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

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

Grid of random variables

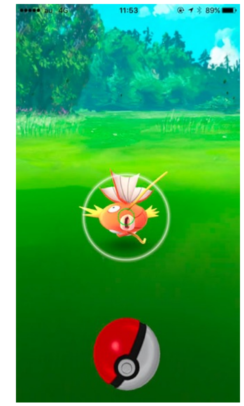
	Number of successes	Time until success	
One trial	Ber(p)	Geo(p)	One success
	$n = 1$	$r = 1$	
Several trials	Bin(n, p)	NegBin(r, p)	Several successes
Interval of time	Poi(λ)	(this Friday)	Interval of time to first success

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

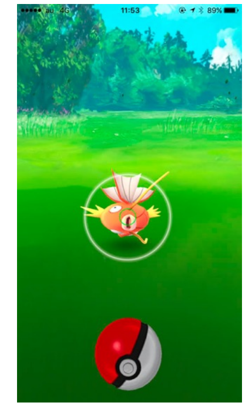
A. $X \sim \text{Bin}(5, 0.1)$
B. $X \sim \text{Poi}(0.5)$
C. $X \sim \text{NegBin}(5, 0.1)$
D. $X \sim \text{NegBin}(1, 0.1)$
E. $X \sim \text{Geo}(0.1)$

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

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

these are the same distribution

first four attempts fail

success on fifth attempt

$P(X=5) = 0.9^4 \cdot 0.1$

$= 0.06561$

Exercises

The hardest part of is almost always deciding what you’re modeling and what random variable to use.

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 might you model the following?

1. # of snapchats you receive in a day
2. # of children born to the same parents until the first one with green eyes
3. If stock went up (1) or down (0) in a day
4. # of probability problems you try until you get 5 correct (if you are randomly correct)
5. # of years between now and 2050 with more than 6 Atlantic hurricanes

Kickboxing with RVs

Choose from:
A. Ber(p)
B. Bin(n, p)
C. Poi(λ)
D. Geo(p)
E. NegBin(r, p)

How might you model the following?

1. # of snapchats you receive in a day
   fixed time interval

   C. Poi(λ)

2. # of children born to the same parents until the first one with green eyes
   yes, each baby is an independent trial :)

   D. Geo(p) or E. NegBin($1, p$)

3. If stock went up (1) or down (0) in a day

   A. Ber(p) or B. Bin($1, p$)

4. # of probability problems you try until you get 5 correct (if you are randomly correct)

   E. NegBin($r = 5, p$)

5. # of years between now and 2050 with more than 6 Atlantic hurricanes

   2050 − 2024 + 1

   B. Bin($n = 27, p$), where $p = P(\geq$ 6 hurricanes in a year) calculated from C. Poi(λ)

   tricky :)

Note: These exercises are designed to build intuition; in a problem statement, you'll generally be given more detail.

Poisson Random Variable

Review

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

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

1. \# of successes in a fixed interval of time, where successes are independent
   $\lambda = E[X]$, average success/interval

1. 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, where requests arrive independently.
- Let X = # requests the server receives in a second.

What is $P(X < 5)$?

Define RVs

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

unit of time is the second

Solve

$$P(X<5) = P(X \le 4)$$

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

$$\approx 0.9473$$

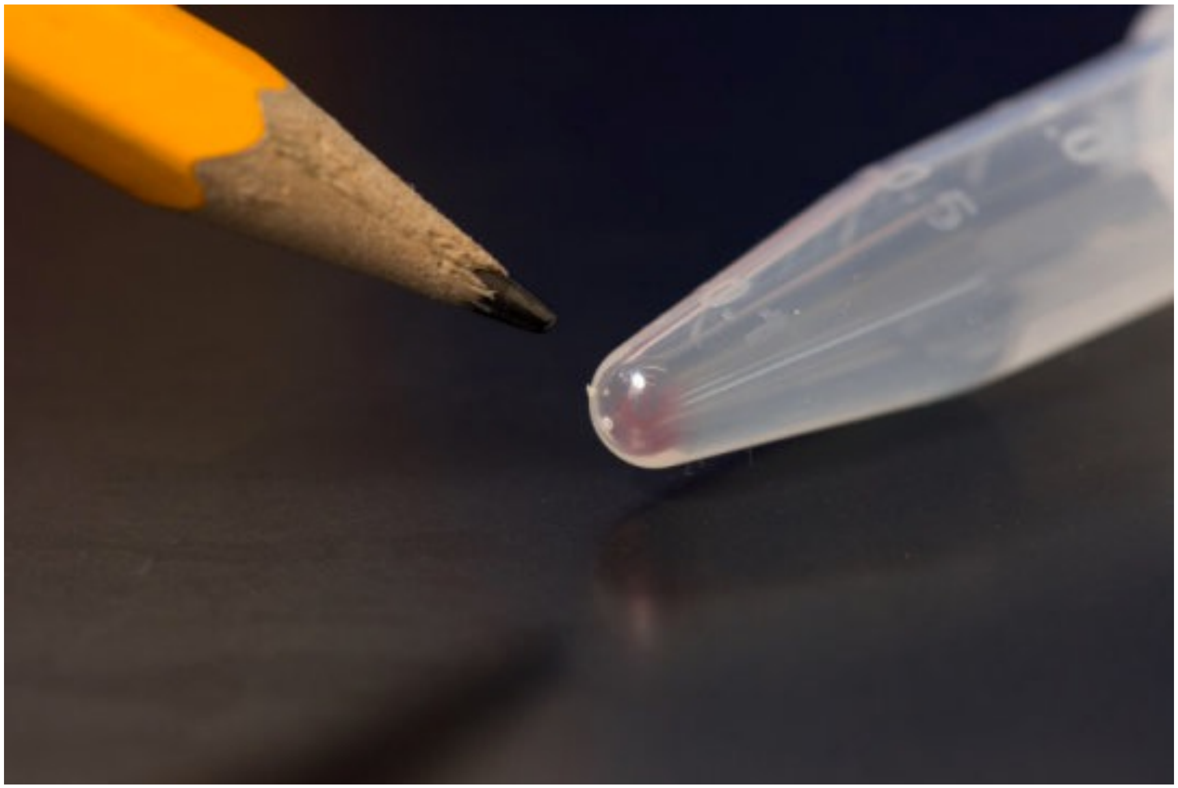
Poisson Random Variable

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

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

1. # of successes in a fixed time interval, where successes are independent
   $\lambda = E[X]$, average success/interval
2. Approximation of $Y \sim \text{Bin}(n, p)$ where n is large and p is small.
   $\lambda = E[Y] = np$ remember that the Poisson was born in the limit of a Binomial with a fixed average value and $n \to \infty$.
   Approximation works well even when trials not entirely independent.

2. 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?

2. 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) = $P(X = 0)$?

here, we choose λ to match the expected value of the original Binomial.

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

≈ 0.990049829

2. Approach 2:

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

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

$$= e^{-0.01}$$

≈ 0.990049834 a good ✅ approximation!

When is a Poisson approximation appropriate?

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

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

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

this is true because n is huge, approaching ∞

Limit analysis

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

this is valid because $p = \frac{\lambda}{n}$ is tiny and approaching 0

Simplify

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

Under which conditions will $X \sim \text{Bin}(n,p)$ behave like $\text{Poi}(\lambda)$, where $\lambda = np$?

A. Large n, large p
B. Small n, small p
C. Large n, small p
D. Small n, large p
E. Other

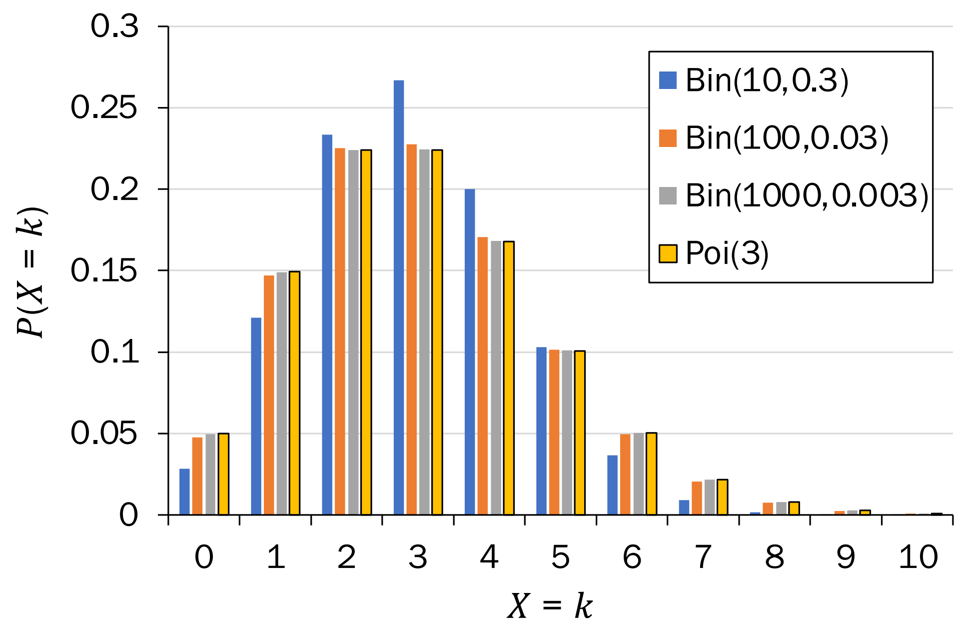
Poisson approximation

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

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

Different interpretations of "moderate":

- $n > 20$ and $p < 0.05$ } CS109 usually relies on this one.
- $n > 100$ and $p < 0.1$



Poisson is Binomial in the limit:

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

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	$E[X] = \lambda$
Support: $\{0, 1, 2, \ldots\}$	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:

$$E[X] = np = \lambda$$

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

very small

Poisson Approximation, approximately

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

You still can apply the Poisson approximation when:

- "Successes" in trials are almost, but 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 or time of day

We won't explore this too much, but we want you to know about it anyway.

Can these Binomial RVs be approximated?

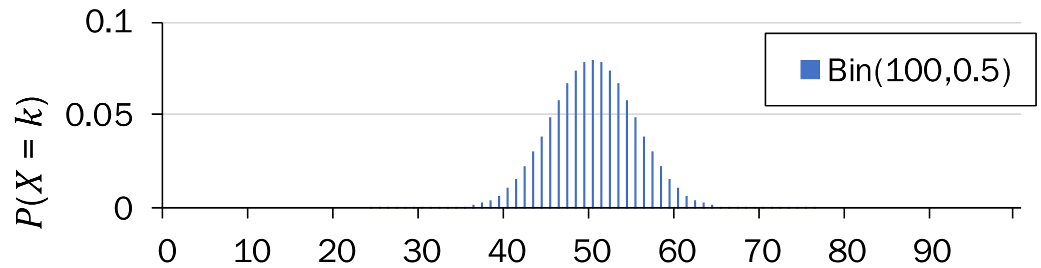
Poisson approximates Binomial when n is large, p is small, and $\lambda = np$ is "moderate".

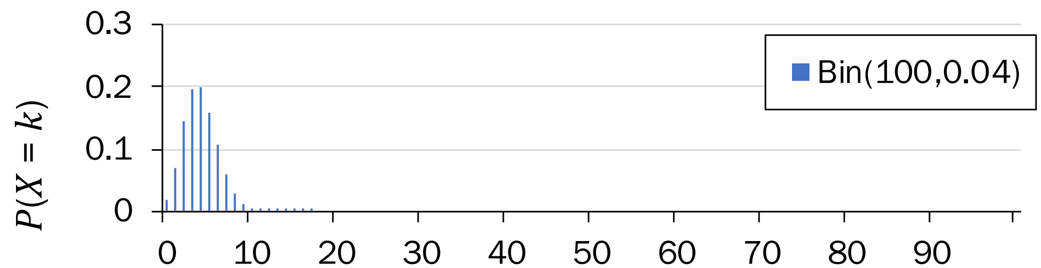
Different interpretations of "moderate":

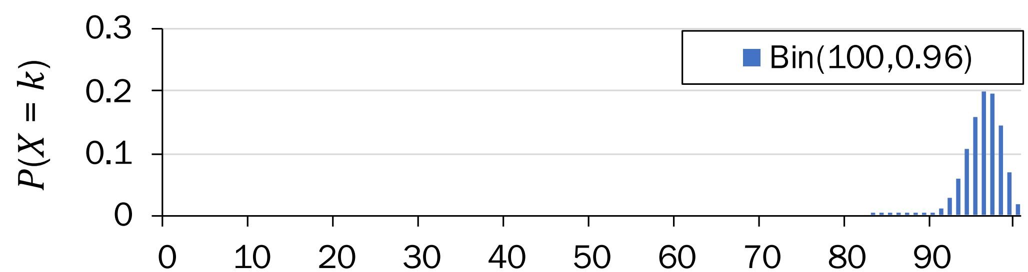
- $n > 20$ and $p < 0.05$
- $n > 100$ and $p < 0.1$

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