08: Poisson and More

Jerry Cain
April 14, 2021
### Quick slide reference

<table>
<thead>
<tr>
<th>slide</th>
<th>topic</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Poisson</td>
<td>08a_poisson</td>
</tr>
<tr>
<td>11</td>
<td>Poisson, continued</td>
<td>08b_poisson_ii</td>
</tr>
<tr>
<td>17</td>
<td>Other Discrete RVs</td>
<td>08c_other_discrete</td>
</tr>
<tr>
<td>25</td>
<td>Exercises</td>
<td>LIVE</td>
</tr>
<tr>
<td>34</td>
<td>Poisson approximation</td>
<td>LIVE</td>
</tr>
<tr>
<td>52</td>
<td>Extra: Modeling exercise: Hurricanes</td>
<td>08e_extra_modeling</td>
</tr>
</tbody>
</table>
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:

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:

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

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

Expand

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

Def natural exponent

\[ = \lim_{n \to \infty} \frac{n!}{n^k(n-k)!} \frac{\lambda^k}{k!} \frac{e^{-\lambda}}{(1 - \frac{\lambda}{n})^k} \]

Expand

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

Limit analysis

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

Simplify

\[ = \frac{\lambda^k}{k!} e^{-\lambda} \]

\[ \lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda} \]
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, continued
Consider an experiment that lasts a fixed interval of time.

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.

Examples:
- # earthquakes per year
- # server hits per second
- # of emails per day
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, assuming the time that each success occurs is independent and the average # of requests over time is constant.

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

**PMF**

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

**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.
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."
Earthquakes

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

2. Solve
Are earthquakes really Poissonian?

Bulletin of the
Seismological Society of America

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

BY J. K. GARDNER and L. KNOPOFF

ABSTRACT

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

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
Geometric RV

Consider an experiment: independent trials of $\text{Ber}(p)$ 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**
\[ E[X] = \frac{1}{p} \]

**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 $\text{Ber}(p)$ random variables.

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

<table>
<thead>
<tr>
<th>$X \sim \text{NegBin}(r, p)$</th>
<th>PMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X = k) = \binom{k-1}{r-1} (1-p)^{k-r} p^r$</td>
<td></td>
</tr>
</tbody>
</table>

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

**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

<table>
<thead>
<tr>
<th>Number of successes</th>
<th>Time until success</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>One trial</strong></td>
<td></td>
</tr>
<tr>
<td>Ber((p))</td>
<td>Geo((p))</td>
</tr>
<tr>
<td>Bin((n, p))</td>
<td>NegBin((r, p))</td>
</tr>
<tr>
<td>(n = 1)</td>
<td>(r = 1)</td>
</tr>
<tr>
<td><strong>Several trials</strong></td>
<td></td>
</tr>
<tr>
<td>Poi((\lambda))</td>
<td></td>
</tr>
<tr>
<td><strong>Interval of time</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(tomorrow)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
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 5\textsuperscript{th} try?

1. Define events/RVs & state goal

   \( X \sim \text{some distribution} \)

   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 5\textsuperscript{th} try?

1. Define events/ RVs & state goal
   \[ X \sim \text{some distribution} \]
   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

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

Want: $P(X = 5)$
08: Poisson and More

Jerry Cain
April 14, 2021
Discrete RVs

The hardest part of problem-solving is determining what is a random variable.
CS109 Learning Goal: Use new RVs

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

You read about the M/D/1 queue:

“\( \lambda \) → Waiting Area → \( \mu \) → Service Node →

“The service time busy period is distributed as a Borel with parameter \( \mu = 0.2 \).”

Goal: You can recognize terminology and understand experiment setup.
### Big Q: Fixed parameter or random variable?

<table>
<thead>
<tr>
<th>Parameter</th>
<th>What is <strong>common</strong> among all outcomes of our experiment?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random variable</th>
<th>What <strong>differentiates</strong> our event from the rest of the sample space?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Examples so far:
- Prob. success
- # total trials
- # target successes
- Average rate of success

Examples so far:
- # of successes
- Time until success (for some definition of time)
Grid of random variables

- **Number of successes**
  - **in one trial**: $\text{Ber}(p)$
  - **in several trials**: $\text{Bin}(n, p)$
  - **in a fixed interval of time**: $\text{Poi}(\lambda)$
- **Time until success**
  - **until one success**: $\text{Geo}(p)$
  - **until several successes**: $\text{NegBin}(r, p)$
  - **(next time!)**: Interval of time until first success
Grid of random variables

Number of successes

- ...in one trial: $\text{Ber}(p)$
- ...in several trials: $\text{Bin}(n, p)$
- ...in a fixed interval of time: $\text{Poi}(\lambda)$

Time until success

- ...until one success: $\text{Geo}(p)$
- ...until several successes: $\text{NegBin}(r, p)$

Interval of time until first success

- (next time!): $n = 1$, $r = 1$
Check out the question on the next slide (Slide 32). Post any clarifications here!

https://edstem.org/us/courses/5090/discussion/357161

Breakout rooms: 4 minutes
Kickboxing with RVs

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. 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 in some decade with more than 6 Atlantic hurricanes

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

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. 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 in some decade with more than 6 Atlantic hurricanes

Note: These exercises are designed to build intuition; in a problem statement, you will generally have more clues.

Choose from:

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

C. $\text{Poi}(\lambda)$

D. $\text{Geo}(p)$ or E. $\text{NegBin}(1, p)$

A. $\text{Ber}(p)$ or B. $\text{Bin}(1, p)$

E. $\text{NegBin}(r = 5, p)$

B. $\text{Bin}(n = 10, p)$, where $p = P(\geq 6 \text{ hurricanes in a year})$ calculated from C. $\text{Poi}(\lambda)$
Poisson Approximation
Poisson Random Variable

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

Consider requests to a web server in 1 second.
- In the past, server load averages 2 hits/second, where hits arrive independently.
- Let $X = \#$ hits the server receives in a second.

What is $P(X < 5)$?

Define RVs Solve
Poisson Random Variable

\[ X \sim \text{Poi}(\lambda) \]

PMF

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

Support: \{0,1,2,...\}

Expectation

\[ E[X] = \lambda \]

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.
   \( \lambda = E[Y] = np \)
   Approximation works well even when trials not entirely independent.
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 \( P(\text{DNA storage is uncorrupted}) = P(X = 0) \)?

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

\( \approx 0.990049829 \)

2. Approach 2:

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

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

\( = e^{-0.01} \)

\( \approx 0.990049834 \) a good approximation!
Think

Slide 41 has a question to go over by yourself.

Post any clarifications here or in chat!

https://edstem.org/us/courses/5090/discussion/357161

Think by yourself: 1 min
When is a Poisson approximation appropriate?

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

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

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2021
Poisson approximation

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 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

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

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

- **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(\(\lambda\)) with the Poisson paradigm

Recall the Binomial:

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

<table>
<thead>
<tr>
<th></th>
<th>Expectation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Y)</td>
<td>(E[Y] = np)</td>
<td>(\text{Var}(Y) = np(1 - p))</td>
</tr>
</tbody>
</table>

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

\[
X \sim \text{Poi}(\lambda)
\]

<table>
<thead>
<tr>
<th></th>
<th>Expectation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X)</td>
<td>(E[X] = \lambda)</td>
<td>(\text{Var}(X) = \lambda)</td>
</tr>
</tbody>
</table>

Proof:

\[
E[X] = np = \lambda \\
\text{Var}(X) = np(1 - p) \to \lambda(1 - 0) = \lambda
\]
Poisson Approximation, approximately

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.
Think by yourself: 2 min

Slide 47 has a question to go over by yourself.

Post any clarifications here or in chat!

https://edstem.org/us/courses/5090/discussion/357161
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 $n \to \infty$, $p \to 0$
A Real License Plate Seen at Stanford

No, it’s not mine…
but I kind of wish it was.
Interlude for announcements
Announcements

Quiz #1

Time frame:       Wednesday 4/21 11:00am – Friday 4/23 10:00am PT
Covers:           Up to end of Week 2 (including Lecture 6)
Georgia’s Review session:  Monday 4/19 7pm PT
Info and practice:  http://web.stanford.edu/class/cs109/quizzes/
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.
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 Poissonian!, $\lambda = 8.5$-ish
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

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

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>X.pmf(k)</td>
<td>P(X = k)</td>
</tr>
<tr>
<td>X.cdf(k)</td>
<td>P(X ≤ k)</td>
</tr>
<tr>
<td>X.mean()</td>
<td>E[X]</td>
</tr>
<tr>
<td>X.var()</td>
<td>Var(X)</td>
</tr>
<tr>
<td>X.std()</td>
<td>SD(X)</td>
</tr>
</tbody>
</table>

2. Find a distribution

Until 1966, things look pretty Poissonian.

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?

\[
P(X > 30) = 1 - P(X \leq 30)
\]

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

\[
= 2.2 \times 10^{-9}
\]

\[X \sim \text{Poi}(\lambda = 8.5)\]
3. The distribution has changed.

1851–1966

Since 1966

Poi(16.7)? Really?
3. What changed?

![Graph showing CO2 levels over the last 10,000 years and annual anomaly relative to 1961-1990 (C).]
3. What changed?

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