Announcements

- Happy Friday
- PSet 1 Solutions available, grades to be released soon.
In Person Next Monday

Details will be sent on Sunday evening. I am not sure if the first lecture will be “live” online.
Grades for PSet 1 Going Out

Number of Students

Grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.07]</td>
<td>0</td>
</tr>
<tr>
<td>(0.07, 0.14]</td>
<td>5</td>
</tr>
<tr>
<td>(0.14, 0.21]</td>
<td>10</td>
</tr>
<tr>
<td>(0.21, 0.28]</td>
<td>50</td>
</tr>
<tr>
<td>(0.28, 0.35]</td>
<td>100</td>
</tr>
<tr>
<td>(0.35, 0.42]</td>
<td>150</td>
</tr>
<tr>
<td>(0.42, 0.49]</td>
<td>200</td>
</tr>
<tr>
<td>(0.49, 0.56]</td>
<td>250</td>
</tr>
<tr>
<td>(0.56, 0.63]</td>
<td>300</td>
</tr>
<tr>
<td>(0.63, 0.7]</td>
<td>350</td>
</tr>
<tr>
<td>(0.7, 0.77]</td>
<td>400</td>
</tr>
<tr>
<td>(0.77, 0.84]</td>
<td>450</td>
</tr>
<tr>
<td>(0.84, 0.91]</td>
<td>500</td>
</tr>
<tr>
<td>(0.91, 0.98]</td>
<td>550</td>
</tr>
<tr>
<td>(0.98, 1.05]</td>
<td>600</td>
</tr>
</tbody>
</table>
Happy Friday
Probability for Extreme Weather?
Review
Fundamental Properties of Random Variables

Random Variable

Semantic Meaning

\[ P(X = x) \]

E\[X\]

Summary stats

Var\( (X) \)

Std\( (X) \)

Measure of spread
Exactly $k$ heads in $n$ coin flips

Probability of exactly $k$ heads, in $n$ coin flips, where each flip is heads with probability $p$:

$$\binom{n}{k} p^k (1 - p)^{n-k}$$
Binomial Random Variable

The number of successes, in $n$ independent trials, where each trial is a success with probability $p$:

$$\binom{n}{k} p^k (1 - p)^{n-k}$$
Declare a Random Variable to be Binomial

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

- Our random variable
- Is distributed as a Binomial
- With these parameters
- Num trials
- Probability of success on each trial
Automatically Know the PMF

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

* This is also called the binomial term

Probability that our variable takes on the value \( k \)
The PMF as a Graph: $X \sim \text{Bin}(n = 20, p = 0.6)$

Parameter $n$: 20  
Parameter $p$: 0.60

Values that $X$ can take on

Probability

$P(x): 0.03499$
### Binomial Random Variable

**Notation:** \( X \sim \text{Bin}(n, p) \)

**Description:** Number of "successes" in \( n \) identical, independent experiments each with probability of success \( p \).

**Parameters:**
- \( n \in \{0, 1, \ldots\} \), the number of experiments.
- \( p \in [0, 1] \), the probability that a single experiment gives a "success".

**Support:** \( x \in \{0, 1, \ldots, n\} \)

**PMF equation:**
\[
\text{Pr}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}
\]

**Expectation:**
\[
\mathbb{E}[X] = np
\]

**Variance:**
\[
\text{Var}(X) = np(1-p)
\]

**PMF graph:**

<table>
<thead>
<tr>
<th>Parameter ( n ):</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter ( p ):</td>
<td>0.60</td>
</tr>
</tbody>
</table>

### Bernoulli Random Variable

**Notation:** \( X \sim \text{Bern}(p) \)

**Description:** A boolean variable that is 1 with probability \( p \).

**Parameters:** \( p \), the probability that \( X = 1 \).

**Support:** \( x \) is either 0 or 1

**PMF equation:**
\[
\text{Pr}(X = x) = \begin{cases} 
  p & \text{if } x = 1 \\
  1 - p & \text{if } x = 0 
\end{cases}
\]

**Expectation:**
\[
\mathbb{E}[X] = p
\]

**Variance:**
\[
\text{Var}(X) = p(1-p)
\]

**PMF graph:**

| Parameter \( p \): | 0.80 |

Values that \( X \) can take on

<table>
<thead>
<tr>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
</tr>
<tr>
<td>0.80</td>
</tr>
<tr>
<td>0.60</td>
</tr>
<tr>
<td>0.40</td>
</tr>
<tr>
<td>0.20</td>
</tr>
<tr>
<td>0.00</td>
</tr>
</tbody>
</table>

Values that \( X \) can take on

<table>
<thead>
<tr>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
</tr>
<tr>
<td>0.00</td>
</tr>
</tbody>
</table>
Natural Exponent Definition

Natural Exponent def:

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

Jacob Bernoulli

https://en.wikipedia.org/wiki/E_(mathematical_constant)
End Review
Algorithmic Ride Sharing
Probability of $k$ requests from this area in the next 1 min
Probability of $k$ requests from this area in the next 1 min
Probability of $k$ requests from this area in the next 1 min

On average $\lambda = 5$ requests per minute
Probability of $k$ requests from this area in the next 1 min

On average $\lambda = 5$ requests per minute

We can break the next minute down into seconds

1 2 3 4 5 6 ... 60
Probability of \( k \) requests from this area in the next 1 min

On average \( \lambda = 5 \) requests per minute

We can break the next minute down into seconds

At each second either get a request or you don’t.
Probability of \( k \) requests from this area in the next 1 min

On average \( \lambda = 5 \) requests per minute

We can break the next minute down into seconds

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

Let \( X = \) Number of requests in the minute

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

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

\[
P(X = 3) = \binom{60}{3} \left(\frac{5}{60}\right)^3 \left(1 - \frac{5}{60}\right)^{57}
\]
On average $\lambda = 5$ requests per minute

At each second either get a request or you don’t.
Let $X = \text{Number of requests in the minute}$

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

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

But what if there are two requests in the same second?
Probability of \( k \) requests from this area in the next 1 min

On average \( \lambda = 5 \) requests per minute

We can break that next minute down into milli-seconds

At each milli-second either get a request or you don’t.
Let \( X = \) Number of requests in the minute

But what if there are two requests in the same second?
Probability of \( k \) requests from this area in the next 1 min

On average \( \lambda = 5 \) requests per minute

We can break that next minute down into \textit{milli}-seconds

At each \textit{milli}-second either get a request or you don’t.

Let \( X = \) Number of requests in the minute

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

\[
P(X = k) = \binom{n}{k} (\lambda/n)^k (1 - \lambda/n)^{n-k}
\]

Can we do any better than \textit{milli}-seconds?
Probability of \( k \) requests from this area in the next 1 min

On average \( \lambda = 5 \) requests per minute

We can break that minute down into infinitely small buckets

OMG so small

Let \( X = \) Number of requests in the minute

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

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

Who wants to see some cool math?
Probability of \( k \) requests from this area in the next 1 min

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

\[
= \lim_{n \to \infty} \frac{n!}{(n-k)!k!} \cdot \frac{\lambda^k}{n^k} \cdot \frac{(1 - \frac{\lambda}{n})^n}{(1 - \frac{\lambda}{n})^k}
\]
By expanding each term

\[
= \lim_{n \to \infty} \frac{n!}{(n-k)!k!} \cdot \frac{\lambda^k}{n^k} \cdot \frac{e^{-\lambda}}{1}
\]
By definition of natural exp

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

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

\[
= \frac{\lambda^k e^{-\lambda}}{k!}
\]
Simplifying
Probability of $k$ requests from this area in the next 1 min
Simeon-Denis Poisson (1781-1840) was a prolific French mathematician.

Published his first paper at 18, became 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.”

I’m going with French Martin Freeman.
X is a **Poisson** Random Variable: the number of occurrences in a fixed interval of time.

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

- \(\lambda\) is the “rate”
- \(X\) takes on values 0, 1, 2…
- has distribution (PMF):

\[
P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}
\]
Consider events that occur over time
- Earthquakes, radioactive decay, hits to web server, etc.
- Have time interval for events (1 year, 1 sec, whatever...)
- Events arrive at rate: $\lambda$ events per interval of time

Split time interval into $n \to \infty$ sub-intervals
- 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

$N(t) = \# \text{ events in original time interval} \sim \text{Poi}(\lambda)$
To the reader!

**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:** $\mathbb{E}[X] = \lambda$

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

**PMF graph:**

Parameter $\lambda$: 5

- Values that $X$ can take on
- Probability
Poisson is great when you have a rate!
Poisson is great when you have a rate and you care about # of occurrences!
Make sure that the time unit for “rate” and match the probability question.
Earthquakes

Average of 2.79 major earthquakes per year.
What is the probability of 3 major earthquakes next year?
Let $X = \text{number of earthquakes next year}$

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

The expected probability mass function for the major earthquake given its mean rate = 2.79 events per year

$P(X = 3) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{2.79^3 e^{-2.79}}{3!} \approx 0.23$
IS THE SEQUENCE OF EARTHQUAKES IN SOUTHERN CALIFORNIA, WITH AFTERSHOCKS REMOVED, POISSONIAN?

By J. K. Gardner and L. Knopoff

Abstract

Yes.
Average of 0.23 major earthquakes per month. What is the probability of 3 major earthquakes next year?
Poisson can approximate a Binomial!
All the movies, images, emails and other digital data from more than 600 smartphones (10,000 gigabytes) can be stored in the faint pink smear of DNA at the end of this test tube.
Storing Data in DNA

Will the DNA storage become corrupt?
- In DNA (and real networks) store large strings
- Length $n \approx 10^4$
- Probability of corruption of each base pair is very small $p \approx 10^{-6}$
- $X \sim \text{Bin}(10^4, 10^{-6})$ is unwieldy to compute

Extreme $n$ and $p$ values arise in many cases
- # bit errors in steam sent over a network
- # of servers crashes in a day in giant data center
Will the DNA storage become corrupt?

- In DNA (and real networks) store large strings
- Length $n \approx 10^4$
- Probability of corruption of each base pair is very small $p \approx 10^{-6}$
- $X \sim \text{Poi}(\lambda = 10^4 \times 10^{-6} = 0.01)$

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

$$P(X = 0) = e^{-\lambda} \frac{1}{0!}$$

$$= e^{-0.01} \approx 0.99$$
Poisson is a Binomial in the Limit

Poisson approximates Binomial where $n$ is large, $p$ is small, and $\lambda = np$ is “moderate”

Different interpretations of "moderate"
- $n > 20$ and $p < 0.05$
- $n > 100$ and $p < 0.1$

Really, Poisson is Binomial as

$$n \to \infty \text{ and } p \to 0, \text{ where } np = \lambda$$
Bin(10, 0.3) vs Bin(100, 0.03) vs Poi(3)
Poisson can be used to approximate a Binomial where $n$ is large and $p$ is small.
Tender (Central) Moments with Poisson

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

$X \sim \text{Poi}(\lambda)$ where $\lambda = np$ ($n \to \infty$ and $p \to 0$)
- $E[X] = np = \lambda$
- $\text{Var}(X) = np(1 - p) = \lambda(1 - 0) = \lambda$
- Yes, expectation and variance of Poisson are same
- It brings a tear to my eye…
No, it’s not mine…
but I kind of wish it was.
Poisson can still provide a good approximation even when assumptions are “mildly” violated

“Poisson Paradigm”

Can apply Poisson approximation when...

- “Successes” in trials are not entirely independent
  - Example: # entries in each bucket in large hash table

- Probability of “Success” in each trial varies (slightly)
  - Small relative change in a very small $p$
  - Example: average # requests to web server/sec. may fluctuate slightly due to load on network
Web Server Load

Consider requests to a web server in 1 second
- In past, server load averages 2 hits/second
- \( X = \# \) hits server receives in a second
- What is \( P(X < 5) \)?

Solution

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

\[ P(X < 5) = \sum_{i=0}^{4} P(X = i) \]

\[ = \sum_{i=0}^{4} e^{-\lambda} \frac{\lambda^i}{i!} \]

Since \( X \) is Poisson

\[ = \sum_{i=0}^{4} e^{-2} \frac{2^i}{i!} \approx 0.95 \]

Since \( \lambda = 2 \)
Probability for Extreme Weather?
To the code!
Historically ~ Poisson(8.5)

Until 1966, things look pretty Poisson

Lambda = 8.5
What is the probability of over 15 hurricanes in a season given that the distribution doesn’t change?

• Let $X =$ # hurricanes in a year. $X \sim \text{Poi}(8.5)$

Solution:

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

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

$$= 0.0135$$

This is the pmf of a Poisson. Your favorite programming language has a function for it.
Twice since 1966 there have been two years with over 30 hurricanes
What is the probability of over 30 hurricanes in a season given that the distribution doesn’t change?

- Let $X = \#$ hurricanes in a year. $X \sim \text{Poi}(8.5)$

Solution:

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

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

$$= 1 - 0.999999997823$$

$$= 2.2e - 09$$

* Challenge: Calculate the probability of two years with over 30 hurricanes
The Distribution has Changed

Since 1966, looks like the distribution has changed

Lambda = 16.6?
What’s Up?

CO2 levels over the last 10,000 years

- Taylor Dome Ice Core
- Law Dome Ice Core
- Mauna Loa, Hawaii

Years (AD)

Atmospheric CO2 (ppm)
What’s Up?

Global annual average surface temperature

Annual anomaly relative to 1961-1990 (°C)
What’s Up?
from scipy import stats  # great package
X = stats.poisson(2.5)  # X ~ Pois(λ = 2.5)
print(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.entropy()</td>
<td>(Differential) entropy of X</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>Std(X)</td>
</tr>
</tbody>
</table>
Next Time
Bernoulli:
• indicator of coin flip $X \sim \text{Ber}(p)$

Binomial:
• # successes in $n$ coin flips $X \sim \text{Bin}(n, p)$

Poisson:
• # successes in $n$ coin flips $X \sim \text{Poi}(\lambda)$

Geometric:
• # coin flips until success $X \sim \text{Geo}(p)$

Negative Binomial:
• # trials until $r$ successes $X \sim \text{NegBin}(r, p)$

Zipf:
• The popularity rank of a random word, from a natural language
• $X \sim \text{Zipf}(s)$
The Poisson Common Path

Solution

"Backbone"