Probability for Extreme Weather?
Review
Consider $n$ independent trials of an experiment with success probability $p$.

- $X$ is number of successes in $n$ trials
- $X$ is a **Binomial** Random Variable:

**Examples**
- # of heads in $n$ coin flips
- # of 1’s in randomly generated length $n$ bit string
- # of disk drives crashed in 1000 computer cluster
  - Assuming disks crash independently
Our random variable is distributed as a Binomial with these parameters:

\[ X \sim \text{Bin}(n, p) \]
If $X$ is a binomial with parameters $n$ and $p$,

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

**Probability Mass Function for a Binomial**

Probability that our variable takes on the value $k$.
Bernoulli is a type of RV that can take on two values, 1 (for success) with probability $p$ and 0 (for failure) with probability $(1 - p)$.

Binomial is the sum of $n$ Bernoullis,
**Fundamental Properties**

Random Variable

- Semantic Meaning
- \( P(X = k) \)
- \( E[X] \)
- \( Var(X) \)
- \( Std(X) \)
- \( E[X^2] \)
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
Average of 2.79 major earthquakes per year.
What is the probability of more than 1 major earthquake next year?
The expected probability mass function for the major earthquake given its mean rate = 2.79 events per year.
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 $n \rightarrow \infty$ 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

$N(t) = \#$ events in original time interval $\sim \text{Poi}(\lambda)$
Poisson is great when you have a rate!
Poisson is great when you have a rate and you care about # of occurrences!
IS THE SEQUENCE OF EARTHQUAKES IN SOUTHERN CALIFORNIA, WITH AFTERSHOCKS REMOVED, POISSONIAN?

By J. K. Gardner and L. Knopoff

Abstract

Yes.
Poisson can approximate a Binomial!
Storing Data on DNA

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.
Recall example of sending bit string over network

- In DNA (and real networks) send 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
Recall example of sending bit string over network

- In DNA (and real networks) send 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 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 $\lambda = np \ (n \to \infty \text{ and } p \to 0)$
  - $E[X] = np = \lambda$
  - $\text{Var}(X) = np(1 - p) = \lambda(1 - 0) = \lambda$
  - 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 is Chill

- 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
• 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 \( P(X = 5) \)?

• Model
  ▪ Assume server cannot acknowledge > 1 hit/msec.
  ▪ 1 sec = 1000 msec. (= large \( n \))
  ▪ \( P(\text{hit server in 1 msec}) = 2/1000 \) (= small \( p \))
  ▪ \( X \sim \text{Poi}(\lambda = 2) \)

\[
P(X = 5) = e^{-2} \frac{2^5}{5!} \approx 0.0361
\]
Probability for Extreme Weather?
Hurricanes per Year since 1851

Number of Hurricanes in Atlantic

<table>
<thead>
<tr>
<th>Year</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1851</td>
<td></td>
</tr>
<tr>
<td>1857</td>
<td></td>
</tr>
<tr>
<td>1863</td>
<td></td>
</tr>
<tr>
<td>1869</td>
<td></td>
</tr>
<tr>
<td>1875</td>
<td></td>
</tr>
<tr>
<td>1881</td>
<td></td>
</tr>
<tr>
<td>1887</td>
<td></td>
</tr>
<tr>
<td>1893</td>
<td></td>
</tr>
<tr>
<td>1899</td>
<td></td>
</tr>
<tr>
<td>1905</td>
<td></td>
</tr>
<tr>
<td>1911</td>
<td></td>
</tr>
<tr>
<td>1917</td>
<td></td>
</tr>
<tr>
<td>1923</td>
<td></td>
</tr>
<tr>
<td>1929</td>
<td></td>
</tr>
<tr>
<td>1935</td>
<td></td>
</tr>
<tr>
<td>1941</td>
<td></td>
</tr>
<tr>
<td>1947</td>
<td></td>
</tr>
<tr>
<td>1953</td>
<td></td>
</tr>
<tr>
<td>1959</td>
<td></td>
</tr>
<tr>
<td>1965</td>
<td></td>
</tr>
<tr>
<td>1971</td>
<td></td>
</tr>
<tr>
<td>1977</td>
<td></td>
</tr>
<tr>
<td>1983</td>
<td></td>
</tr>
<tr>
<td>1989</td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
</tr>
</tbody>
</table>
To the code!
Historically $\sim$ 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:

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

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

$$= 1 - 0.98$$

$$= 0.02$$

This is the pmf of a Poisson. Your favorite programming language has a function for it.
Twice since 1966 there have been 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:

\[
P(X > 30) = 1 - P(X \leq 30)
= 1 - \sum_{i=0}^{30} P(X = i)
= 1 - 0.999999997823
= 2.2 \times 10^{-9}
\]

This is the pdf of a Poisson. Your favorite programming language has a function for it.
Since 1966, looks like the distribution has changed

Lambda = 16.6?
What's Up?

Global annual average surface temperature

Annual anomaly relative to 1961-1990 (°C)

Year
What's Up?
## Python Scipy Poisson Methods

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pmf(k)</td>
<td>Probability mass function.</td>
</tr>
<tr>
<td>cdf(k)</td>
<td>Cumulative distribution function.</td>
</tr>
<tr>
<td>entropy()</td>
<td>(Differential) entropy of the RV.</td>
</tr>
<tr>
<td>mean()</td>
<td>Mean of the distribution.</td>
</tr>
<tr>
<td>var()</td>
<td>Variance of the distribution.</td>
</tr>
<tr>
<td>std()</td>
<td>Standard deviation of the distribution.</td>
</tr>
</tbody>
</table>
The Poisson Common Path
Next Time
Don’t have to memorize all of the following distributions. We want you to get a sense of how random variables work.
Geometric Random Variable

• \( X \) is **Geometric** Random Variable: \( X \sim \text{Geo}(p) \)
  - \( X \) is number of independent trials until first success
  - \( p \) is probability of success on each trial
  - \( X \) takes on values 1, 2, 3, …, with probability:
    \[
P(X = n) = (1 - p)^{n-1} p
    \]
  - \( E[X] = 1/p \) \( \quad \text{Var}(X) = (1 - p)/p^2 \)

• Examples:
  - Flipping a coin (\( P(\text{heads}) = p \)) until first heads appears
  - Urn with \( N \) black and \( M \) white balls. Draw balls (with replacement, \( p = N/(N + M) \)) until draw first black ball
  - Generate bits with \( P(\text{bit} = 1) = p \) until first 1 generated
Example of Geometric Distribution

Frequency

Time between Events
Negative Binomial Random Variable

- $X$ is **Negative Binomial** RV: $X \sim \text{NegBin}(r, p)$
  - $X$ is number of independent trials until $r$ successes
  - $p$ is probability of success on each trial
  - $X$ takes on values $r, r + 1, r + 2…$, with probability:
    \[ P(X = n) = \binom{n-1}{r-1} p^r (1 - p)^{n-r}, \text{ where } n = r, r + 1, \ldots \]
  - $E[X] = r/p \quad \text{Var}(X) = r(1 - p)/p^2$

- Note: Geo($p$) $\sim$ NegBin($1, p$)

- Examples:
  - # of coin flips until $r$-th “heads” appears
  - # of strings to hash into table until bucket 1 has $r$ entries
**Zipf Random Variable**

- $X$ is **Zipf** RV: $X \sim \text{Zipf}(s,N)$
  - $X$ is the rank index of a chosen word

\[
P(X = k) = \frac{1}{k^s \cdot H}
\]