Discrete random variables

Definition

Properties

Experiment outcomes

Discrete Random Variable, $X$

$P(X = x) = p(x)$

$E[X]$

$E[X^2]$

$\text{Var}(X)$

$SD(X)$

Note: Random Variables also called distributions
Variance

The variance of a random variable $X$ with mean $E[X] = \mu$ is

$$\text{Var}(X) = E[(X - \mu)^2]$$

An easier way to compute variance: $\text{Var}(X) = E[X^2] - (E[X])^2$
Binomial random variable

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

PMF

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

Expectation

\[ E[X] = np \]

Variance

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

1. The random variable
2. is distributed as a
3. Binomial
4. with parameters

Range: \{0, 1, \ldots, n\} (aka support)

P(success) on each trial

\# independent trials
Today’s plan

Poisson

Poisson Paradigm

Some more Discrete RVs (if time)
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)
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:

A smearing of infinitesimal buckets

For each time bucket:
- Independent trial
- You get a request (1) or you don’t (0).

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

$E[X] = \lambda = 5$

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

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

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

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

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

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

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

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

Rearrange

Expand

Def natural exponent

Limit analysis

+ cancel

Simplify
Probability of $k$ requests from this area in the next 1 minute?

On average, $\lambda = 5$ requests per minute

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

\[
\begin{align*}
X & \sim \text{Poi}(\lambda) \\
\text{PMF} & \quad P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \\
\text{Range: } & \quad \{0, 1, 2, \ldots\} \\
\text{Expectation} & \quad E[X] = \lambda \\
\text{Variance} & \quad \text{Var}(X) = \lambda
\end{align*}
\]

Examples:
- # earthquakes per year
- # server hits per second
- # of emails per day

Yes, expectation and variance of Poisson are the same (shown later)
Earthquakes

There are an average of 2.79 major earthquakes in the world each year. What is the probability of 3 major earthquakes happening next year?

1. Define RVs

2. Solve
Earthquakes

There are an average of 2.79 major earthquakes in the world each year. What is the probability of 3 major earthquakes happening next year?

1. Define RVs

\[ X \sim \text{Poi}(2.79) \]

2. Solve

\[ P(X = 3) = e^{-\lambda} \frac{\lambda^k}{k!} \]

where \( k = 3 \), \( \lambda = 2.79 \)

\[ = e^{-2.79} \left( \frac{2.79}{3!} \right)^3 \approx 0.23 \]
Are earthquakes really Poissonian?

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.
Web server load

Consider requests to a web server in 1 second.

- In the past, server load averages 2 hits/second.
- Let $X = \# \text{ hits the server receives in a second.}$

What is $P(X < 5)$?

1. Define RVs  
2. Solve

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

$$E[X] = \lambda$$

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$
Today’s plan

Poisson

Poisson Paradigm

Some more Discrete RVs
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?
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:

   $X \sim \text{Bin}(n = 10^4, p = 10^{-6})$
   
   $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$
   
   $= \binom{10^4}{0} 10^{-6 \cdot 0} (1 - 10^{-6})^{10^6 - 0}$
   
   $\approx 0.99049829$

2. Approach 2:

   $X \sim \text{Poi}(\lambda = 10^4 \cdot 10^{-6} = 0.01)$
   
   $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} = e^{-0.01} \frac{0.01^0}{0!}$
   
   $= e^{-0.01}$
   
   $\approx 0.99049834$  
   
   a good approximation!
The Poisson Paradigm

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 $n \rightarrow \infty$, $p \rightarrow 0$

Poisson can approximate Binomial!
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 $n \to \infty, p \to 0$
Announcements

Problem Set 1 Grades
Grades/feedback on Gradescope: (hopefully) tomorrow
Regrades: submit through Gradescope
Coding problem is out of 10 points, not 20
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.

### Poisson Random Variable

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

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

- **Expectation**
  
  $$E[X] = \lambda$$

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

**Range:** $\{0, 1, 2, \ldots\}$

**Examples:**

- # earthquakes per year
- # server hits per second
- # of emails per day

Yes, expectation and variance of Poisson are the same (intuition now)
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:

\[
E[X] = np = \lambda \\
\text{Var}(X) = np(1 - p) \to \lambda(1 - 0) = \lambda
\]
Today’s plan

Poisson

Poisson Paradigm

Some more Discrete RVs
More discrete RVs

Part of CS109 learning goals:
• Translate a problem statement into a random variable
• Understand new random variables

We focus primarily on Binomial, Bernoulli, and Poisson.

Here are a few more to get a sense of how random variables work.

Focus on understanding how and when to use RVs, not on memorizing PMFs.
Consider an experiment: independent trials of Ber\((p)\) random variables. A **Geometric** random variable \(X\) is the # of trials until the first success.

**Geometric RV**

\[
X \sim \text{Geo}(p)
\]

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
Consider an experiment: independent trials of Ber(p) random variables.

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

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

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

Range: \{r, r + 1, ... \}

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>One trial</td>
<td>One success</td>
</tr>
<tr>
<td>Bin((n, p)) (n = 1)</td>
<td>Geo((p)) (r = 1)</td>
</tr>
<tr>
<td>Several trials</td>
<td>Several successes</td>
</tr>
<tr>
<td>Poi((\lambda))</td>
<td>Interval of time to first success</td>
</tr>
<tr>
<td>Interval of time</td>
<td>(next time)</td>
</tr>
</tbody>
</table>
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
   - $X \sim \text{Bin}(5, 0.1)$
   - $X \sim \text{Poi}(0.5)$
   - $X \sim \text{NegBin}(5, 0.1)$
   - $X \sim \text{NegBin}(1, 0.1)$
   - $X \sim \text{Geo}(0.1)$
   - None/other

2. Solve
   - $P(X = 5)$
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: $P(X = 5)$

2. Solve

   $P(X = 5) = (1 - p)^{k-1}p$, where $k = 5, p = 0.1$

   $= (0.9)^4(0.1)$

   $\approx 0.066$
Hurricanes

What is the probability of an extreme weather event?

How do we model the number of hurricanes in a season (year)?

Step 1. graph your distribution.
Hurricanes per year since 1851

Which graph is a histogram (i.e., distribution) of frequency (# of hurricanes per year)?

A.

B.
Hurricanes

How do we model the number of hurricanes in a season (year)?

Step 2. Find a reasonable distribution (Poisson) and compute parameters.
Improbability

Until 1966, things look pretty Poisson.

What is the probability of over 15 hurricanes in a season (year) given that the distribution doesn’t change?

\[ P(X > 15) = 1 - P(X \leq 15) \]

\[ = 1 - \sum_{k=0}^{15} P(X = k) \]

\[ = 1 - 0.986 = 0.014 \]

This is the PMF of a Poisson. Your favorite programming language has a function for it.

In Python 3: `from scipy import stats`  
\[ X = \text{stats.poisson}(8.5) \]
\[ X.pmf(k) \]
Hurricanes

How do we model the number of hurricanes in a season (year)?

Step 3. See if there are outliers
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?

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

This is the PMF of a Poisson. Your favorite programming language has a function for it.

In Python 3: 
from scipy import stats  
X = stats.poisson(8.5)  
X.pmf(k)
The distribution has changed.

1851–1966

Since 1966

Poi(8.5)

Count (1851-1966)

Poi(16.7?)

Count (1966-2015)
What changed?

CO₂ levels over the last 10,000 years

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

Global annual average surface temperature

Annual anomaly relative to 1961-1990 (°C)
What changed?

It’s not just climate change. We also have better data collection now.
### Python SciPy RV methods

```python
from scipy import stats  # great package
X = stats.poisson(8.5)  # X ~ Poi(\lambda = 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 \leq k))</td>
</tr>
<tr>
<td>X.mean()</td>
<td>(E[X])</td>
</tr>
<tr>
<td>X.var()</td>
<td>(\text{Var}(X))</td>
</tr>
<tr>
<td>X.std()</td>
<td>(\text{SD}(X))</td>
</tr>
</tbody>
</table>