Announcements

PS1:
- Grades out later today
- Solutions out after class today

PS2 due today

PS3 out today (due next Friday 7/20)
Midterm announcement

Tuesday, July 24, 7-9pm
Hewlett 201

- Closed book, closed computer, no calculators
- Two 8.5” x 11” pages of notes (front and back) allowed
- Covers up to and including Friday 7/20’s lecture
- Practice midterm (and PS4) to be released mid-next week

If you need alternate accommodations (OAE, academic reasons) and you have not yet heard from me, contact me ASAP
Goals for today

More discrete distributions
- Poisson approximation
- Geometric
- Negative Binomial
Summary of RVs

Expectation:

\[ E[X] = \sum_i x_i \, p(x_i) \]
\[ E[aX + b] = aE[X] + b \]  \hspace{1cm} \text{Linearity of Expectation}
\[ E[g(X)] = \sum_i g(x_i) \, p(x_i) \]

Variance:

\[ \text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2 \]
\[ \text{Var}(aX + b) = a^2 \text{Var}(X) \]
Discrete RV distributions, part 1

\[ X \sim \text{Ber}(p) \]
\[ X \in \{0,1\} \]

\[ Y \sim \text{Bin}(n,p) \]
\[ Y = \sum_{i=1}^{n} X_i \]
\[ \text{s.t. } X_i \sim \text{Ber}(p) \]

Binomial is sum of \( n \) independent Bernoullis

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

Poisson is Binomial as \( n \to \infty, p \to 0 \), where \( \lambda = np \)
Poisson is Binomial in Limit

When \( n \) is large, \( p \) is small, and \( \lambda = np \) is “moderate,” Poisson approximates Binomial.

“moderate”?  
- \( n > 20 \) and \( p < 0.05 \)  
- \( n > 100 \) and \( p < 0.1 \)  
- \( n \to \infty, p \to 0 \)
from matplotlib import pyplot as plt
import numpy as np
from scipy.stats import poisson, binom

def plot_pmf(lamb, n, p):
    fig = plt.figure()
    ax = plt.gca()
    # reasonable x range with nonzero probabilities
    x = np.arange(poisson.ppf(0.01, lamb),
                  poisson.ppf(0.9999, lamb))
    w = (x[1]-x[0])/3
    ax.bar(x, poisson.pmf(x, lamb), width=w, alpha=0.5,
           label='poi({})'.format(lamb))
    ax.bar(x+w, binom.pmf(x, n, p), width=w, alpha=0.5,
           label='bin({}, {})'.format(n, str(p)[1:])))
    ax.legend()
    plt.show()

# "moderate": n > 20, p < 0.05 or n > 100, p < 0.1
lamb = 500  # poisson parameter, lambda
n = 1000    # binomial parameters n
p = lamb/float(n)  # p
plot_pmf(lamb, n, p)
Poisson is like the pizza world record

Bake a giant Neapolitan pizza (2km long)

- Ingredients: Many basil leaves, lots of other stuff

Slice the pizza

- Some slices get more basil, some slices get less

Pick a slice

- Let $X$ be # of basil leaves in this slice.

$$X \sim \text{Poi}(\lambda), \text{where } \lambda = \frac{\text{total # basil leaves}}{\text{# pizza slices}}$$
Poisson is life?

Two ways to think about Poisson:

• # events that occur in an interval of time
  
  1 minute = 60000 milliseconds

• # basil leaves in a particular slice pizza slice, where there are many slices

\[ n \to \infty \]
CS = Show me your stepback

Hash tables
- Strings = basil leaves
- Buckets = pizza slices

Server crashes in data center
- Crashed servers = basil leaves
- Server rack = particular slice of pizza

Facebook login requests (i.e., web server requests)
- Requests = basil leaves
- Server receiving request = pizza slice
Defective Chips

Computer chips are produced.

- $p = 0.1$ that a chip is defective (chips are independent)
- Consider a sample of $n = 10$ chips.

$P(\text{sample contains } \leq 1\text{ defective chip})$?

Solution:

Define: $Y = \#$ defective chips in sample.

$P(Y \leq 1) = P(Y = 0) + P(Y = 1)$

Using $Y \sim \text{Bin}(10,0.1)$:

$P(Y \leq 1) = \binom{10}{0}(0.1)^0(0.9)^{10} + \binom{10}{1}(0.1)^1(0.9)^9 \approx 0.7361$

Using $Y \sim \text{Poi}(\lambda = (0.1)(10) = 1)$:

$P(Y \leq 1) = \frac{1^0}{0!} e^{-1} + \frac{1^1}{1!} e^{-1} = 2 e^{-1} \approx 0.7358$
The Poisson Paradigm

Poisson can still provide a good approximation, 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
Birthday Poisson Problem

What is the probability that of \( m \) people, none share the same birthday regardless of year?

To use Poisson approximation:

- Define trial (with probability of success, \( p \))
  
  Pair of people \((x, y), x \neq y\)
  
  \( E_{x,y} = x \text{ and } y \text{ have same birthday (trial success)} \)
  
  \( p = P(E_{x,y}) = 1/365 \)  
  - not all independent

- Compute # of said trials, \( n \)
  
  \( n = \binom{m}{2} = m(m-1)/2 \)

- Compute \( \lambda \) and answer problem.

\( X \sim \text{Poi}(\lambda) \) where \( \lambda = np = m(m-1)/730 \)

\[
P(X = 0) = \frac{(m(m-1)/730)^0}{0!} e^{-m(m-1)/730}
\]
Birthday Poisson Problem

What is the probability that of $m$ people, none share the same birthday regardless of year?

To use Poisson approximation:

- Define trial (with probability of success, $p$)
- Compute # of said trials, $n$ $\quad n = \binom{m}{2} = \frac{m(m-1)}{2}$
- Compute $\lambda$ and answer problem.
  
  $X \sim \text{Poi}(\lambda)$ where $\lambda = np = \frac{m(m-1)}{730}$

\[
P(X = 0) = \frac{(\frac{m(m-1)}{730})^0}{0!} e^{-\frac{m(m-1)}{730}} = e^{-\frac{m(m-1)}{730}}
\]

Solve for smallest integer $m$: $e^{-\frac{m(m-1)}{730}} \leq 0.5$

\[
\ln(e^{-\frac{m(m-1)}{730}}) \leq \ln(0.5) \Rightarrow m \geq 23 \quad \text{same as before!}
\]
Break

Attendance: tinyurl.com/cs109summer2018
Geometric RV

Consider independent trials of Ber(p) random variables.

• X is # of trials until first success.

Geometric RV, X:

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

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

\[ E[X] = 1/p \]

\[ \text{Var}(X) = (1 - p)/p^2 \]

\[ X \in \{1, 2, \ldots\} = \mathbb{Z}^+ \]

Examples:

• Flipping a coin (P(heads) = p) until first heads appears

• Urn with N red and M white balls. Draw balls (with replacement), \( p = N/(N+M) \) until draw first red ball.

• Generate bits with \( P(\text{bit} = 1) = p \) until first 1 generated
Let $X \sim \text{Geo}(p)$. (# of trials until first success, $P(\text{trial success}) = p$)

Recall CDF = cumulative distribution function, $F_X(k) = P(X \leq k)$.

What is the CDF of $X$?

**Solution 1:**

\[ P(X \leq k) = \sum_{m=1}^{k} P(X=m) = \sum_{m=1}^{k} (1 - p)^{m-1}p \]

\[ = p \sum_{m=0}^{k-1} (1 - p)^{m} \]

\[ = p \frac{1-(1-p)^{k-1+1}}{1-(1-p)} \]

\[ = 1 - (1 - p)^{k} \]

**Calculation Reference:**

\[ \sum_{i=0}^{n} x^i = \frac{1 - x^{n+1}}{1 - x} \]
CDF of a Geometric RV

Let $X \sim \text{Geo}(p)$. (# of trials until first success, $P(\text{trial success}) = p$)

Recall CDF = cumulative distribution function, $F_X(k) = P(X \leq k)$.

What is the CDF of $X$?

Solution 2:

Define: $C_i = \text{success on } i\text{th trial}$

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

$= 1 - P(C_1^c C_2^c \ldots C_k^c)$

$= 1 - P(C_1^c)P(C_2^c)\ldots P(C_k^c)$

$= 1 - (1 - p)^k$
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.

How many do you need to bring with you so that the probability of successful capture is at least 99%?

Solution:

Define: $X = \#$ of Pokeballs needed until capture

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

WTF: $F_X(k) = P(X \leq k) \geq 0.99$

$P(X \leq k) = 1 - (1-p)^k \geq 0.99$

$0.01 \geq (1-p)^k$

$\ln(0.01) \geq k \ln(1-p) \rightarrow k \leq \frac{\ln(0.01)}{\ln(0.9)} \approx 43.7$
Consider independent trials of Ber(p) random variables.

- $X$ is the number of trials until $r^{th}$ success.

**Negative Binomial RV, $X$:**

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

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

- $E[X] = r/p$
- $\text{Var}(X) = r (1 - p)/p^2$

**Examples:**

- # of coin flips until $r$-th heads appears
- # of strings to hash into table until bucket 1 has $r$ entries

Note: Geo(p) = NegBin(1,p)
Getting that degree

A conference accepts papers randomly (???).

- Each paper is accepted independently with probability $p = 0.25$.
- A hypothetical grad student needs 3 accepted papers to graduate.

What is $P(\text{grad student needs exactly 10 submissions})$?

Solution:

Define: $X = \# \text{ of tries to get 3 accepts}$

$X \sim \text{NegBin}(3, 0.25)$

WTF:

$P(X = 10) = \binom{10 - 1}{3 - 1} 0.25^3 (1 - 0.25)^{10-3}$

$= \binom{9}{2} 0.25^3 (0.75)^7 \approx 0.075$
Hypergeometric RV

Consider $m$ red balls and $N - m$ black balls in a hat.

- Draw $n$ balls without replacement.
- $X$ is # of red balls drawn.

Hyper Geometric RV, $X$:

$$X \sim \text{HypG}(n, N, m)$$

$$P(X=k) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}$$

$$E[X] = n \left( \frac{m}{N} \right)$$

$$\text{Var}(X) = \frac{nm(N-n)(N-m)}{N^2(N-1)}$$

$$X \in \{0, 1, 2, \ldots, m\}$$

Note: If $p = \frac{m}{N}$ (probability of drawing red on 1st draw),

$\text{HypG}(n, N, m) \Rightarrow \text{Bin}(n, \frac{m}{N})$,

as $N \rightarrow \infty$ $m/N$ remains constant
Defining Random Variables and PMFs.

1. Read the problem. Define your experiment.

2. Determine the constant factors in the experiment and the variable desired outcome we want.

3. Define your random variable.

4. Pattern match to an existing probability mass function, or define a new one via counting.
Check your experiment goals

Want To Find (WTF):

# successes in an experiment
1. Start and finish experiment.
2. Review and count occurrences.

Bernoulli:
1. Flip coin once.
2. Check if heads.

Binomial:
1. Flip coin \( n \) times.
2. Check if \( k \) heads.

Poisson:
1. Wait a unit of time.
2. Check if \( k \) occurrences.

# trials until success
1. Start experiment.
2. Once success is observed, finish the experiment.

Geometric:
1. Flip coin repeatedly.
2. If Heads, exit.

Negative-Binomial:
1. Flip coin repeatedly
2. If \( k \) heads, exit.
More RV practice

Are the below Ber(p), Bin(n,p), Poi(\(\lambda\)), Geo(p), NegBin(r,p)?

1. # of snapchats you receive in a day  
   \(\text{Poi}(\lambda)\)

2. # of children until the first one with brown eyes  
   \(\text{Geo}(p)\)

3. Whether stock went up or down  
   \(\text{Ber}(p)\) aka Bin(1,p)

4. # of probability problems you try until you get 5 correct (if you are randomly correct)  
   \(\text{NegBin}(r,p)\)

5. # of years in some decade with more than 6 Atlantic hurricanes  
   \(\text{Bin}(n,p)\), where \(p = P(\geq 6 \text{ hurricanes in a year})\)
Midterm question: Bit Coin Mining

You “mine a bitcoin” if, for a given data $D$, you find a number $N$ such that $\text{Hash}(D,N)$ produces an $M$-length bitstring that starts with $g$ zeros.

1. What is the probability that the first number you try will produce a (random) bit string that starts with $g$ zeros (i.e., you mine a bitcoin)?
   - WTF: $P(\text{M bits starts with } g \text{ zeros})$
   - $P(\text{M-length starts with } g \text{ zeros}) = (0.5)^g$

2. How many different numbers do you expect to have to try before you mine 5 bitcoins?
   - WTF: Expectation of # of tries of different $N$ until 5 coins
   - Observation: $P(\text{a particular } N \text{ mines a bitcoin}) = (0.5)^g$
   - Define: $X = \# \text{ trials until 5 successes}$
   - $X \sim \text{NegBin}(r,p)$, $r = 5$, $p = (0.5)^g$
   - $E[X] = 5/(0.5)^g$