CSCI 3022: Intro to Data Science

Lecture 8: More Discrete Random Variables and their Distributions

Rachel Cox

Department of Computer Science

"I don't know if you're a mathematician but my wife's not happy with her Poisson distribution."

Announcements & Reminders

HW2 - Due Friday at 11:59 PM

Midterm - Tuesday Feb. 25

6:30 - 8pm

MATH 100

makeup day Wednesday at 6:30pm

Practicum 1 - Due Friday the 28th

What will we learn today?

- ❑ Geometric Distribution
- ❑ Negative Binomial Distribution
- ❑ Poisson Distribution

- ❑ *A Modern Introduction to Probability and Statistics, Chapter 4 & 5*

Review from Last Time

A **discrete random variable** (r.v.) X is a function that maps the elements of the sample space Ω to a finite number of values $a_1, a_2, \ldots, a_n$ or an infinite number of values $a_1, a_2, \ldots$

A discrete random variable X has a **Bernoulli distribution** (X~Ber(p)), where $0 \leq p \leq 1$, if its probability mass function is given by

$$f(1) = p_X(1) = P(X = 1) = p \quad \text{and} \quad p_X(0) = P(X = 0) = 1 - p$$

A discrete random variable X has a **Binomial Distribution** (X~Bin(n, p)) with $n = 1, 2, \ldots$ and $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n$$

Binomial–like Distributions

Example: You are doing an exit poll outside of a voting station on Election Day. As people exit, you ask them questions about their political affiliation, who they voted for, etc.

In particular, you are interested in how registered Independents voted. You know that about 20% of registered voters are registered as Independents.

Suppose you interview 100 people. ***Let X be a random variable describing the number of actual Independents you encounter.***

What distribution does X have?

Binomial Distribution $\quad \mathrm{Bin}(100, .2)$

Binomial–like Distributions

Example: You are doing an exit poll outside of a voting station on Election Day. As people exit, you ask them questions about their political affiliation, who they voted for, etc.

In particular, you are interested in how registered Independents voted. You know that about 20% of registered voters are registered as Independents.

Suppose you talk to a lot of registered Republicans and Democrats, but haven't found an Independent yet. ***Let X be a random variable describing the number of people you have interviewed up to and including your first registered Independent voter.***

What distribution does X have?

Geometric Distribution

Binomial–like Distributions

<u>Example</u>: You are doing an exit poll outside of a voting station on Election Day. As people exit, you ask them questions about their political affiliation, who they voted for, etc.

In particular, you are interested in how registered Independents voted. You know that about 20% of registered voters are registered as Independents.

Suppose you are really interested in talking to a lot of Independents. ***Let X be a random variable describing the number of people you have to talk to in order to interview exactly 100 registered Independents.***

What distribution does X have?

Negative Binomial Distribution

Binomial–like Distributions

<u>Example</u>: You are doing an exit poll outside of a voting station on Election Day. As people exit, you ask them questions about their political affiliation, who they voted for, etc.

In particular, you are interested in how registered Independents voted. You know that about 20% of registered voters are registered as Independents.

You are concerned about being overwhelmed during peak voting times, so you track the number of people arriving in line at the voting station. ***Let X be a random variable describing the number of voters that arrive at the station over a 15-minute period.***

What distribution does X have?

Poisson distribution

Geometric Distribution

Example: Suppose that you flip the same biased coin repeatedly. How many times do you flip the coin before you see your first Heads?

$$P(H) = p$$

$$P(TH) = (1-p) \cdot p$$

$$P(TTH) = (1-p)^2 \cdot p$$

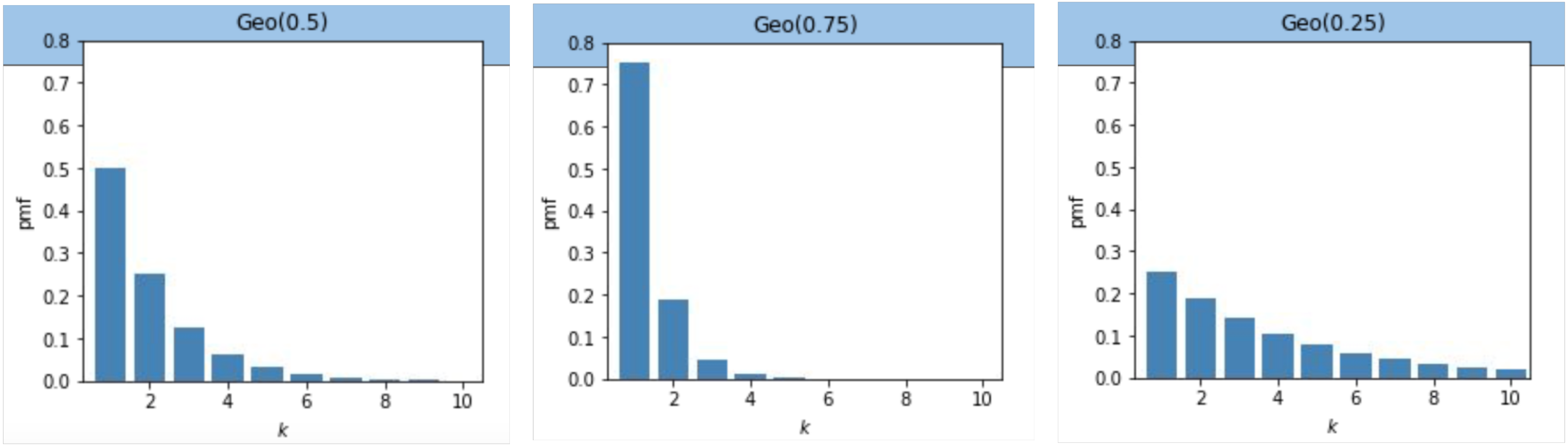
$$\vdots$$

$$P(TT \cdots \underset{\uparrow\, k^{th} \text{ flip}}{T}\, H) = (1-p)^{k-1} \cdot p$$

Geometric Distribution

A discrete random variable X has a **Geometric distribution** (X~Geo(p)) with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(k) = P(X = k) = (1-p)^{k-1} \cdot p \quad \text{for } k = 1, 2, 3, \ldots$$



Geometric Distribution

A discrete random variable X has a **Geometric distribution** (X~Geo(p)) with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(k) = P(X = k) = (1 - p)^{k-1} \cdot p \quad \text{for } k = 1, 2, 3, \ldots$$

Question: What assumptions did we implicitly make in deriving the Geometric distribution?

- Each trial is independent.
- Each trial is a Bernoulli random variable with probability of success p.

Negative Binomial Distribution

success on k^{th} flip

Example: Suppose that you flip the same biased coin repeatedly. How many times do you flip the coin before you see 3 Heads total?

Suppose we observe the 3^{rd} head (success) on flip # K

$\Rightarrow$ In the first $K-1$ flips we had 2 successes (H)

use the Binomial distribution to find the probability of 2 successes in $k-1$ flips.

$$p_X(K) = \left[\text{Binomial Random variable for } 2 \text{ successes in } k-1 \text{ trials}\right] \cdot p$$

$$= \binom{k-1}{2} p^2 (1-p)^{(k-1)-2} \cdot p = \boxed{\binom{k-1}{2} p^3 (1-p)^{k-3}}$$

Negative Binomial Distribution

A discrete random variable X has a **Negative Binomial distribution** (X~NB(r, p)) with parameters r and p, where $r > 1$ and $0 \leq p \leq 1$, if its probability mass function is given by

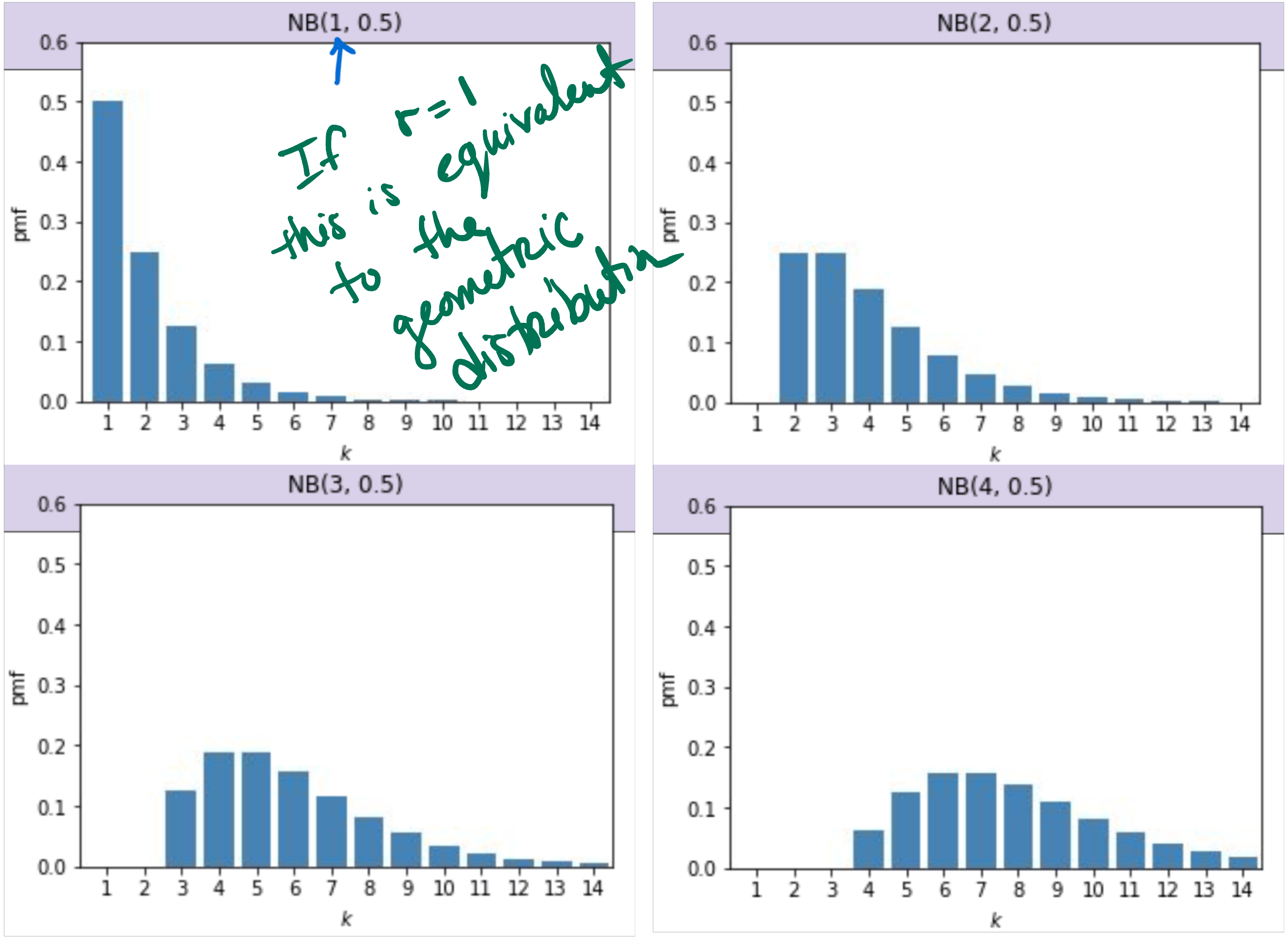
$$p_X(k) = P(X = k) = \binom{k-1}{r-1} \mathrm{p}^{\mathrm{r}}(1-\mathrm{p})^{k-r}$$

p = probability of success for each trial

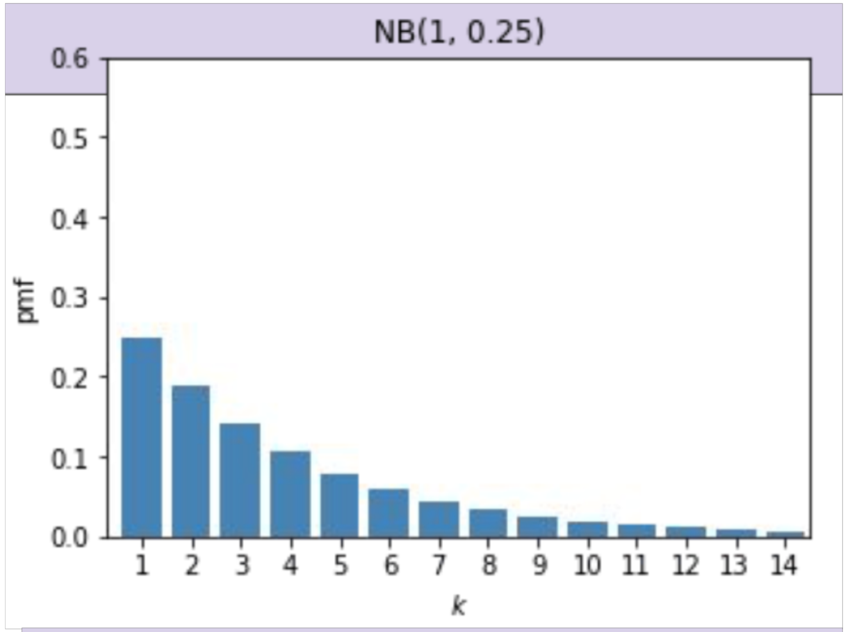
r = number of successes we want to observe

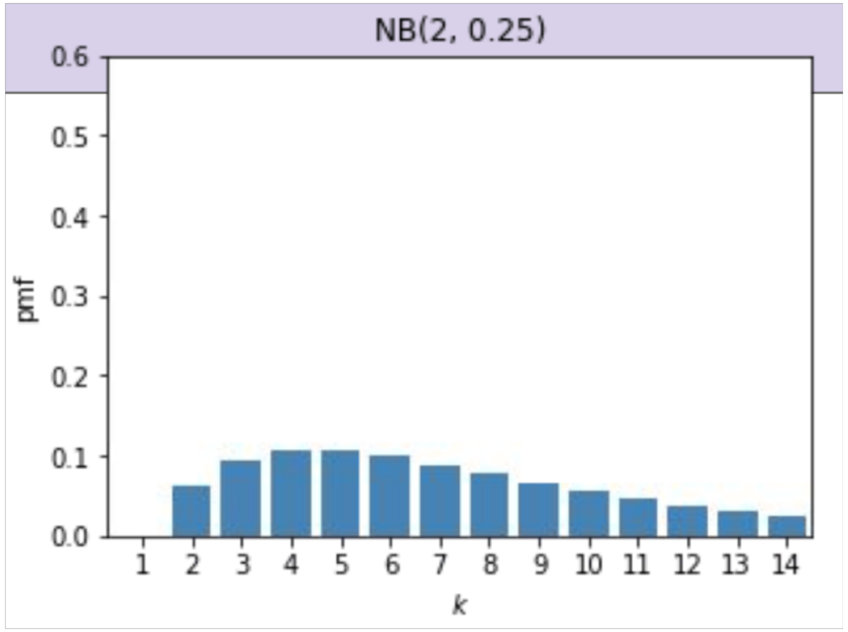
X = number of trials needed before we observe r successes

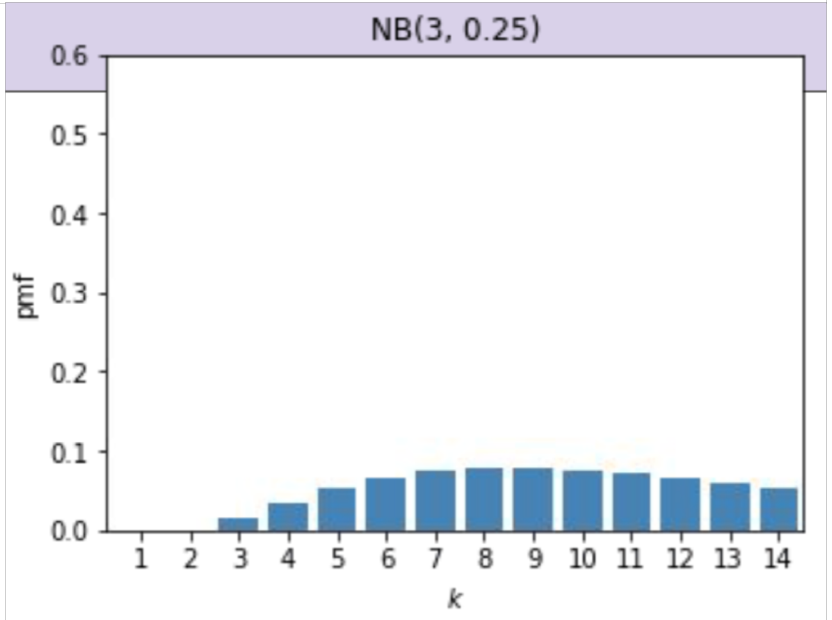
Negative Binomial Distribution



Negative Binomial Distribution







Question: What assumptions did we implicitly make in deriving the Negative Binomial distribution?

- Each trial is a Bernoulli r.v. with probability of success *p*
- Each trial is independent

Poisson Distribution

Example: A grocery store check-out line moves people through at an average rate of 1 customer per 3 minutes. What is the probability that they check-out:
(i) 1 customer in 5 minutes? (ii) 3 customers in 5 minutes? (iii) 10 customers in 5 minutes?

$$\mu = \frac{1 \text{ customer}}{3 \text{ minutes}} = \frac{1/3 \text{ customer}}{1 \text{ minute}}$$

Think of this process as the limit of a Binomial random variable, as we pack more and more trials into a fixed slice of time.

$\mu = np$ $n =$ time slices, $p =$ probability of a customer in that time slice

$\mu/n = p$

What is the probability of seeing k successes in that slice of time?

$$\lim_{n\to\infty} \underbrace{\binom{n}{k} p^k (1-p)^{n-k}}_{\text{Binomial Distribution}} = \underbrace{\lim_{n\to\infty} \binom{n}{k} \left(\frac{\mu}{n}\right)^k \left(1-\frac{\mu}{n}\right)^{n-k}}_{\text{Poisson distribution}}$$

Poisson Distribution

Recall: $\lim\limits_{n\to\infty}\left(1+\frac{a}{n}\right)^{bn} = e^{ab}$

Example: A grocery store check-out line moves people through at an average rate of 1 customer per 3 minutes. What is the probability that they check-out:

(i) 1 customer in 5 minutes? (ii) 3 customers in 5 minutes? (iii) 10 customers in 5 minutes?

Derivation:

Poisson

$$P_X(k) = \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \lim_{n\to\infty} \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

$$= \lim_{n\to\infty} \frac{n!}{k!\,(n-k)!} \left(\frac{\mu}{n}\right)^k \left(1-\frac{\mu}{n}\right)^{n-k}$$

$$= \lim_{n\to\infty} \frac{n(n-1)(n-2)\cdots(n-k+1)\cancel{(n-k)!}}{k!\,\cancel{(n-k)!}} \cdot \frac{\mu^k}{n^k} \cdot \frac{\overbrace{\left(1-\mu/n\right)^n}^{e^{-\mu}}}{\underbrace{\left(1-\mu/n\right)^k}_{\to 1}}$$

$$= \frac{\mu^k e^{-\mu}}{k!} \lim_{n\to\infty} \frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k}$$

$$= \frac{\mu^k e^{-\mu}}{k!} \lim_{n\to\infty} \boxed{\frac{n}{n}}\boxed{\frac{(n-1)}{n}}\boxed{\frac{(n-2)}{n}}\cdots\boxed{\frac{(n-k+1)}{n}}$$

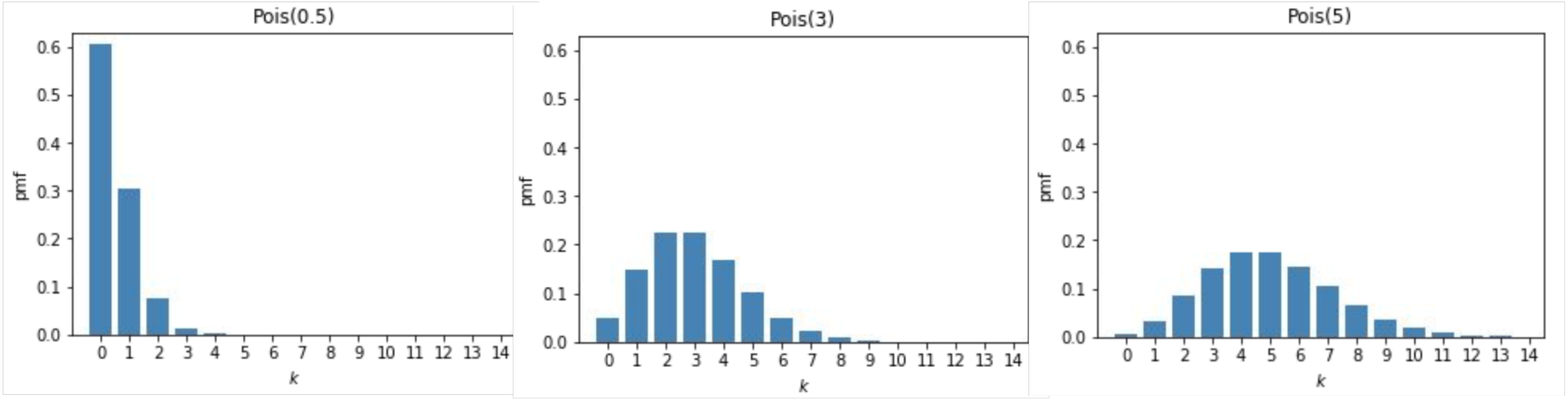
1

$$= \frac{\mu^k e^{-\mu}}{k!}$$

Poisson Distribution

A discrete random variable X has a **Poisson distribution** (X~Pois(μ)) with parameter μ, where $\mu > 0$, if its probability mass function is given by

$$p_X(k) = P(X = k) = \frac{\mu^k e^{-\mu}}{k!} \quad \text{for } k = 0, 1, 2, \ldots$$



Poisson Distribution

$\mu = \dfrac{1 \text{ customer}}{3 \text{ minutes}}$

$\underset{n}{\sim}$ $P = \left(\frac{1}{3}\text{customer}\right) \cdot \frac{1}{\text{minute}}$ $\Big\}$ $P = \frac{1}{3}$

1 minute - 1 time chunk

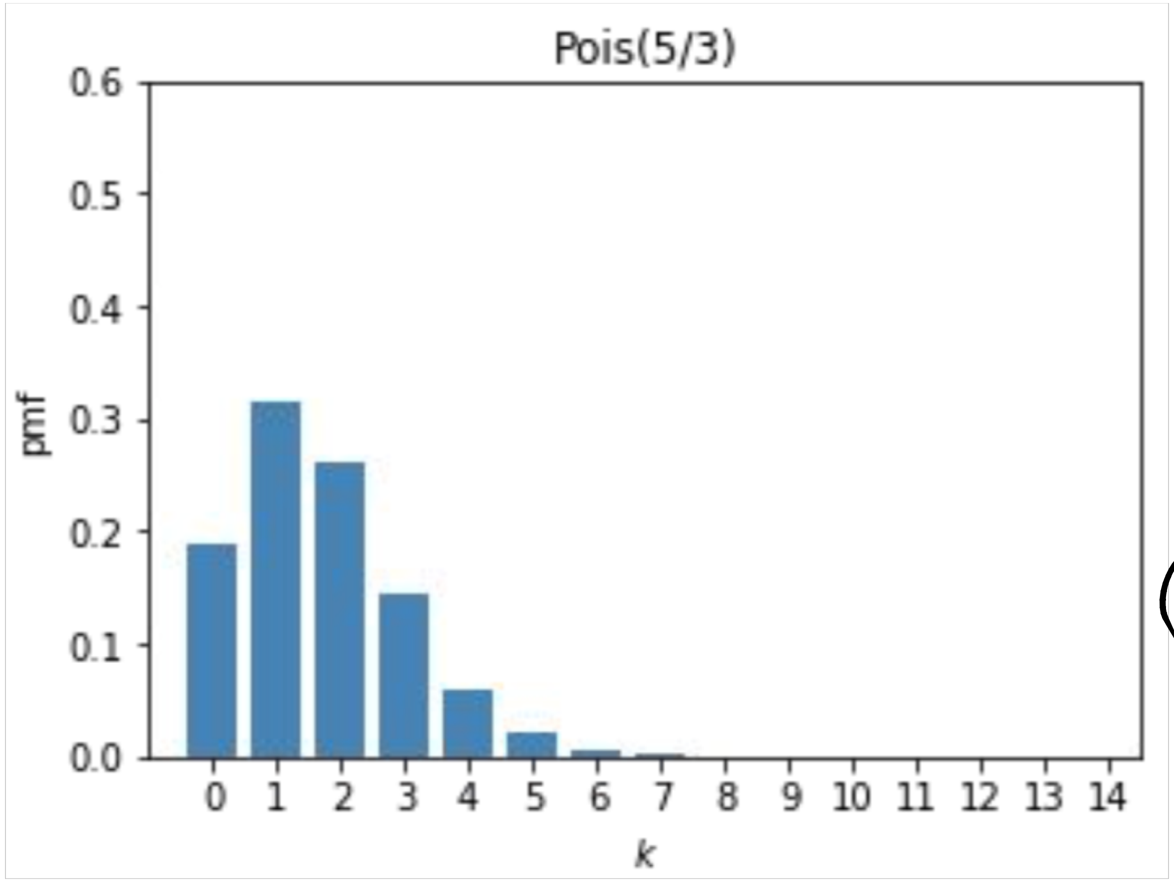
Example: A grocery store check-out line moves people through at an average rate of 1 customer per 3 minutes. What is the probability that they check-out:

(i) 1 customer in 5 minutes? (ii) 3 customers in 5 minutes? (iii) 10 customers in 5 minutes?

5 minutes:
5 time chunks.

$\mu = n \cdot P$

$\mu = 5 \cdot \frac{1}{3} = \frac{5}{3}$



(i) $\mu = 5/3$ $k=1$

$p_X(k) = \dfrac{e^{-\mu}\mu^k}{k!}$

$= \dfrac{e^{-5/3} \cdot \frac{5}{3}}{1}$

$\approx .315$

(ii) $\mu = 5/3$ $k = 3$

$p_X(3) = \dfrac{e^{-5/3}\left(5/3\right)^3}{3!}$

$\approx .146$

Poisson Distribution

Question: What assumptions did we implicitly make in deriving the Poisson distribution?

- Probability of observing a single event over a small interval of time is proportional to the size of the interval.

- Each event/arrival is independent.

Next Time:

- More distributions!