University of Colorado Boulder

CSCI 3022, Intro to Data Science
Summer 2018
Tony Wong

Lecture 7: Discrete Random Variables and Their Distributions

Announcements and reminders

- HW 2 due Friday at 5 PM
- Quizlet 3 due Thursday at 1 PM

Previously, on CSCI 3022…

Definition: 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 $a_1, a_2, \dots, a_n$ or an infinite number of values $a_1, a_2, \dots$

Definition: A **probability mass function** (pmf) is the map between the random variable's values and the probabilities of those values

$$f(a) = P(X=a)$$

Definition: A **cumulative distribution function** (cdf) is a function whose value at a point *a* is the cumulative sum of probability masses up until *a*

$$F(a) = P(X \leq a)$$

Warm-up problem

Example: S'pose you roll two fair, six-sided dice. Let X be a random variable representing the maximum of the two dice. Some questions:

$X = 1, 2, 3, 4, 5, 6$

Q1: What are the possible values that X can take?

Q2: Which elements of the sample space map to which values of X?

Q3: What is the pmf of the random variable X?

$\rightarrow \Omega = \{ \underbrace{(1,1)}_{(\omega_1, \omega_2)}, (1,2) \ldots \}$

ω_1 \ ω_2	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	2	3	4	5	6
3	3	3	3	4	5	6
4	4	4	4	4	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

$\leftarrow X$

Warm-up problem

Example: S'pose you roll two fair, six-sided dice. Let X be a random variable representing the maximum of the two dice. Some questions:

Q1: What are the possible values that X can take?
Q2: Which elements of the sample space map to which values of X?
Q3: What is the pmf of the random variable X?

	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	2	3	4	5	6
3	3	3	3	4	5	6
4	4	4	4	4	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

a	1	2	3	4	5	6
$f(a)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$7/36$	$9/36$	$11/36$

Warm-up problem

Example: S'pose you roll two fair, six-sided dice. Let X be a random variable representing the maximum of the two dice. Some questions:

Q4: What is the probability that $\underbrace{\text{X is an even number}}_{E}$?

a	1	2	3	4	5	6
$f(a)$	1/36	3/36	5/36	7/36	9/36	11/36

$P(E) = P(X=2 \cup X=4 \cup X=6)$

disjoint: $\cup = +$

$= P(X=2) + P(X=4) + P(X=6)$

$= \frac{3}{36} + \frac{7}{36} + \frac{11}{36}$

$= \frac{21}{36}$

Warm-up problem

Example: S'pose you roll two fair, six-sided dice. Let X be a random variable representing the maximum of the two dice. Some questions:

Q5: What is the probability that X is 3 or smaller? $= F(3) = P(X \le 3)$

$= \frac{1}{36} + \frac{3}{36} + \frac{5}{36}$

$= \frac{9}{36}$

a	1	2	3	4	5	6
$f(a)$	1/36	3/36	5/36	7/36	9/36	11/36

Warm-up problem

Example: S'pose you roll two fair, six-sided dice. Let X be a random variable representing the maximum of the two dice. Some questions:

Q6: What is the complete cdf of X?

a	1	2	3	4	5	6
$f(a)$	1/36	3/36	5/36	7/36	9/36	11/36

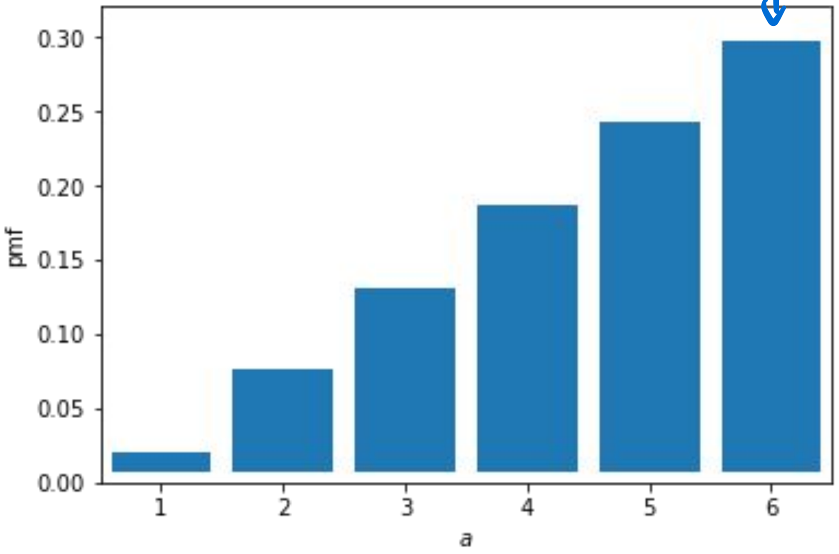
a	1	2	3	4	5	6
$F(a)$	$\frac{1}{36}$	$\frac{4}{36}$	$\frac{9}{36}$	$\frac{16}{36}$	$\frac{25}{36}$	$\frac{36}{36}$

$P(X \leq 1)$ $P(X \leq 2)$

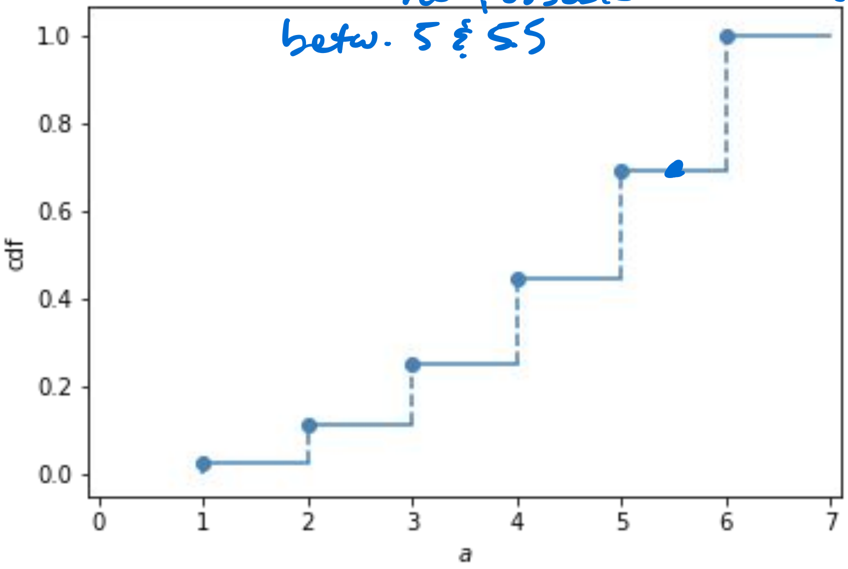
Visualizing pmfs and cdfs

Example: S'pose you roll two fair, six-sided dice. Let X be a random variable representing the maximum of the two dice.

$\frac{11}{36}$



Equally likely to get $X \le 5.5$ as you are $X \le 5$ b/c no possible outcomes betw. 5 & 5.5



Common discrete r.v. distributions

Discrete r.v.’s can be categorized into different types or classes that each **model** different real-world situations

The Bernoulli distribution

The Bernoulli distribution is used to model experiments with only two possible outcomes.

Often referred to as "success" and "failure", and encoded as 1 and 0, respectively.

Definition: A discrete random variable X has a **Bernoulli distribution** with parameter *p*, where $0 \leq p \leq 1$, if its probability mass function is given by

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

We denote this distribution by Ber(p)

$X \sim Ber(p)$

"X is a random variable w/ a Bernoulli distribution w/ parameter p"

Ex: Fair coin:

$X \sim Ber(0.5)$

The Bernoulli distribution

The Bernoulli distribution is used to model experiments with only two possible outcomes.

Often referred to as “success” and “failure”, and encoded as 1 and 0, respectively.

Definition: A discrete random variable X has a **Bernoulli distribution** with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by

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

We denote this distribution by Ber(p)

Question: Wouldn’t it be nice if we could describe the pmf with a single equation?

$$f(x) = p^x (1-p)^{1-x}$$

The Bernoulli distribution

The Bernoulli distribution is used to model experiments with only two possible outcomes.

Often referred to as “success” and “failure”, and encoded as 1 and 0, respectively.

Definition: A discrete random variable X has a **Bernoulli distribution** with parameter *p*, where $0 \leq p \leq 1$, if its probability mass function is given by

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

We denote this distribution by Ber(p)

Question: Wouldn’t it be nice if we could describe the pmf with a single equation?

→ if we have $p_X(1)=p$, and $p_X(0)=1-p$, then for x in {0, 1}, we have $p_X(x) = p^x (1-p)^{1-x}$

A counting interlude

We’ll come back to the Bernoulli distribution in a minute.   First… we ***count!***

Counting comes up all over the place in probability
(and therefore in data science, comp sci, math, physics, etc…)

Some counting is easy: how many integers are there in the interval [0, 9] ?

But we’re interested in counting problems that require a bit more thought:

- Dan, Chris, Rhonda and Tony line up at the coffee stand. How many different orders could they stand in?
- If there are 10 problems on an exam, and you need 7 correct to pass, how many different ways are there to pass?

A counting interlude

We'll talk about two important kinds of counting problems today:

1) Counting **permutations** means counting the number of ways that a set of objects can be ordered (or *permuted!*) elements are DISTINGUISHABLE

Example: Dan, Chris, Rhonda and Tony line up at the coffee stand. How many different orders could they stand in?

2) Counting **combinations** means counting the number of ways that a set of objects can be combined into subsets

Example: If there are 10 problems on an exam, and you need 7 correct to pass, how many different ways are there to pass?

A counting interlude

We'll talk about two important kinds of counting problems today:

1) Counting **permutations** means counting the number of ways that a set of objects can be ordered (or *permuted!*)

Example: Dan, Chris, Rhonda and Tony line up at the coffee stand. How many different orders could they stand in?

2) Counting **combinations** means counting the number of ways that a set of objects can be combined into subsets

INDISTINGUISHIBLE elements

Example: If there are 10 problems on an exam, and you need 7 correct to pass, how many different ways are there to pass?

Permutations

Questions:

$n=1$

- How many ways are there to order a set of 1 object? $1 = 1!$
- How many ways are there to order a set of 2 objects? $2 = 2!$
- How many ways are there to order a set of 3 objects? Think about a third object showing up to 2 of them:

The Big Question: What is a formula for the number of ways you can order *n* objects?

$\{a, b, c\}$

a b   b a

c   c

3 ways + 3 ways = $6 = 3!$

$\boxed{\text{for } n \text{ objects, there are } n! \text{ permutations of them}}$

Also: $0! = 1$ by convention

Permutations

Question: What if we have *n* objects, but want to count permutations of only *r* of them?

Example: How many 3-character strings can we make if each character is a distinct letter from the English alphabet?

$$= \underline{26} \times \underline{25} \times \underline{24} \times \underbrace{\cancel{23 \times 22 \times \cdots \times 3 \times 2 \times 1}}_{23!}$$

$$= \frac{26!}{23!}$$

Question: What is the general formula for *r*-permutations of *n* objects?

$$\boxed{P(n,r) = \frac{n!}{(n-r)!}}$$

$P_{n,r}$ or ${}_nP_r$

Permutations

Question: What if we have *n* objects, but want to count permutations of only *r* of them?

Example: How many 3-character strings can we make if each character is a distinct letter from the English alphabet?

Answer: $26 \times 25 \times 24$

Question: What is the general formula for *r*-permutations of *n* objects?

Answer: $$P(n, r) = \frac{n!}{(n-r)!}$$

Combinations

Counting **combinations** means counting the number of ways a set of objects can be combined into subsets

a, b "same" as b, a

Key difference: When counting combinations, **order does not matter**.

Example: How many 3-character **combinations** can we make if each character is a distinct letter from the English alphabet?

If we care about order, then we'd have $P(26, 3)$

→ now need to treat abc same as bac, cab, ...

But only want to count one ← 6 possible perms. of $\{c, a, b\}$

Combinations

Example: How many 3-character **combinations** can we make if each character is a distinct letter from the English alphabet?

→ Start with the number of 3-permutations of 26 letters:

$$P(26,3) = \frac{26!}{23!}$$

→ But if order doesn't matter, we are counting combinations **multiple times**

we're over-counting by a factor of $6 = 3!$

→ divide out by $3!$ to account for over-counting

→ $C(26,3) = \dfrac{26!}{23!\,3!}$

Combinations

There are many different notations for combinations. You can write the number of ways to choose *r* objects from a set of *n* objects as:

$$C(n, r) \quad \textbf{or} \quad C_{n,\,r} \quad \textbf{or} \quad \binom{n}{r} = \frac{n!}{(n-r)!\,r!}$$

↑ binomial coefficients

Example: If there are 10 problems on an exam, and you need 7 correct to pass, how many different ways are there to pass?

First: You could get 7 correct ... or 8 or 9 or 10

→ # ways = $C(10,7) + C(10,8) + C(10,9) + C(10,10)$

↑

ways to pick 7 out of 10 problems to get correct; order doesn't matter

Combinations

There are many different notations for combinations. You can write the number of ways to choose *r* objects from a set of *n* objects as:

$$C(n, r) \qquad \textbf{or} \qquad C_{n,\,r} \qquad \textbf{or} \qquad \binom{n}{r}$$

Example: If there are 10 problems on an exam, and you need 7 correct to pass, how many different ways are there to pass?

Answer: # ways = C(10, 7) + C(10, 8) + C(10, 9) + C(10, 10) = …

Combinations

Example: A coin is flipped 10 times. How many possible outcomes have exactly 2 Heads?

$$\binom{10}{2}$$

Example: A coin is flipped 10 times. How many possible outcomes have 2 Heads or fewer?

$$= \binom{10}{2} + \binom{10}{1} + \binom{10}{0}$$

$$= \frac{10!}{8!\,2!} + \frac{10!}{9!\,1!} + \frac{10!}{10!\,0!}$$

Sum of Bernoulli random variables

Example: S'pose you show up to a quiz completely unprepared. Oh no! The quiz has 5 questions, each with 4 multiple choice options. You decide to guess each answer in a completely random way. What is the probability that you get 3 questions correct?

Define: $R_i = \begin{cases} 0 & \text{if problem } i \text{ is wrong} \\ 1 & \text{if } \ldots \text{ right} \end{cases}$

What dist. for R_i?

$R_i \sim Ber(p = \frac{1}{4})$

Sum of Bernoulli random variables

Example: S'pose you show up to a quiz completely unprepared. Oh no! The quiz has 5 questions, each with 4 multiple choice options. You decide to guess each answer in a completely random way. What is the probability that you get 3 questions correct?

For $i = 1, 2, 3, 4, 5$ let $$R_i = \begin{cases} 1 & \text{if the } i^{th} \text{ answer is correct} \\ 0 & \text{if the } i^{th} \text{ answer is incorrect} \end{cases}$$

Question: What can you say about R_i ?

Sum of Bernoulli random variables

Example: S'pose you show up to a quiz completely unprepared. Oh no! The quiz has 5 questions, each with 4 multiple choice options. You decide to guess each answer in a completely random way. What is the probability that you get 3 questions correct?

→ Let r.v. X = # correct answers. → $X = R_1 + R_2 + R_3 + R_4 + R_5$

Question: What values can X take?

$X = 0, 1, 2, \ldots, 5$

Question: What is the probability that you get 0 problems correct? :(

$p = 0.25$ to get one correct

$\rightarrow (1-p)(1-p)(1-p)(1-p)(1-p) = 0.75^5$

Question: What is the probability that you get 0 problems correct? :(

Answer:
$$\begin{aligned} P(X=0) &= P(R_1=0, R_2=0, R_3=0, R_4=0, R_5=0) \\ &= P(R_1=0)P(R_2=0)P(R_3=0)P(R_4=0)P(R_5=0) \\ &= \left(\frac{3}{4}\right)^5 \end{aligned}$$

Question: What is the probability that you get exactly 1 problem correct?

5 ways

$R_1 = 1$ & $R_2 = R_3 = R_4 = R_5 = 0 \longrightarrow \underline{p(1-p)^4}$

$R_2 = 1$ & $R_1 = R_3 = R_4 = R_5 = 0 \longrightarrow p(1-p)^4$

account for choosing 1 problem to get correct out of 5

$= \binom{5}{1} p(1-p)^4$

Sum of Bernoulli random variables *(Example, cont.)*

Question: What is the probability that you get 0 problems correct? :(

Answer:
$$\begin{aligned} P(X=0) &= P(R_1=0, R_2=0, R_3=0, R_4=0, R_5=0) \\ &= P(R_1=0)P(R_2=0)P(R_3=0)P(R_4=0)P(R_5=0) \\ &= \left(\frac{3}{4}\right)^5 \end{aligned}$$

Question: What is the probability that you get exactly 1 problem correct?

→ P(X=1) = ???

→ Could have gotten Q1 correct → $P(R_1=1, \text{others} = 0) = (1/4)(3/4)^4$
→ Could have gotten Q2 correct → $P(R_2=1, \text{others} = 0) = (3/4)(1/4)(3/4)^3 = (1/4)(3/4)^4$
→ … *and so on* … **$P(X=1) = 5\cdot(1/4)\cdot(3/4)^4$**

Sum of Bernoulli random variables

Question: What is the probability that you get k problems correct out of n problems total?
(k some ≥ 0, integer)

k successes out of n Bernoulli trials

Answer: $$p_X(k) = \binom{n}{k} p^k \, (1-p)^{n-k}$$

Where the combination (or binomial coefficient) is $$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

The Binomial distribution

Question: What is the probability that you get k problems correct out of n problems total?
(k some ≥ 0, integer)

Answer: $$p_X(k) = \binom{n}{k} p^k \, (1-p)^{n-k}$$

Definition: A discrete r.v. X has a **binomial distribution** with parameters *n* and *p*, where *n* = 1, 2, … and 0 ≤ *p* ≤ 1, if its probability mass function is given by

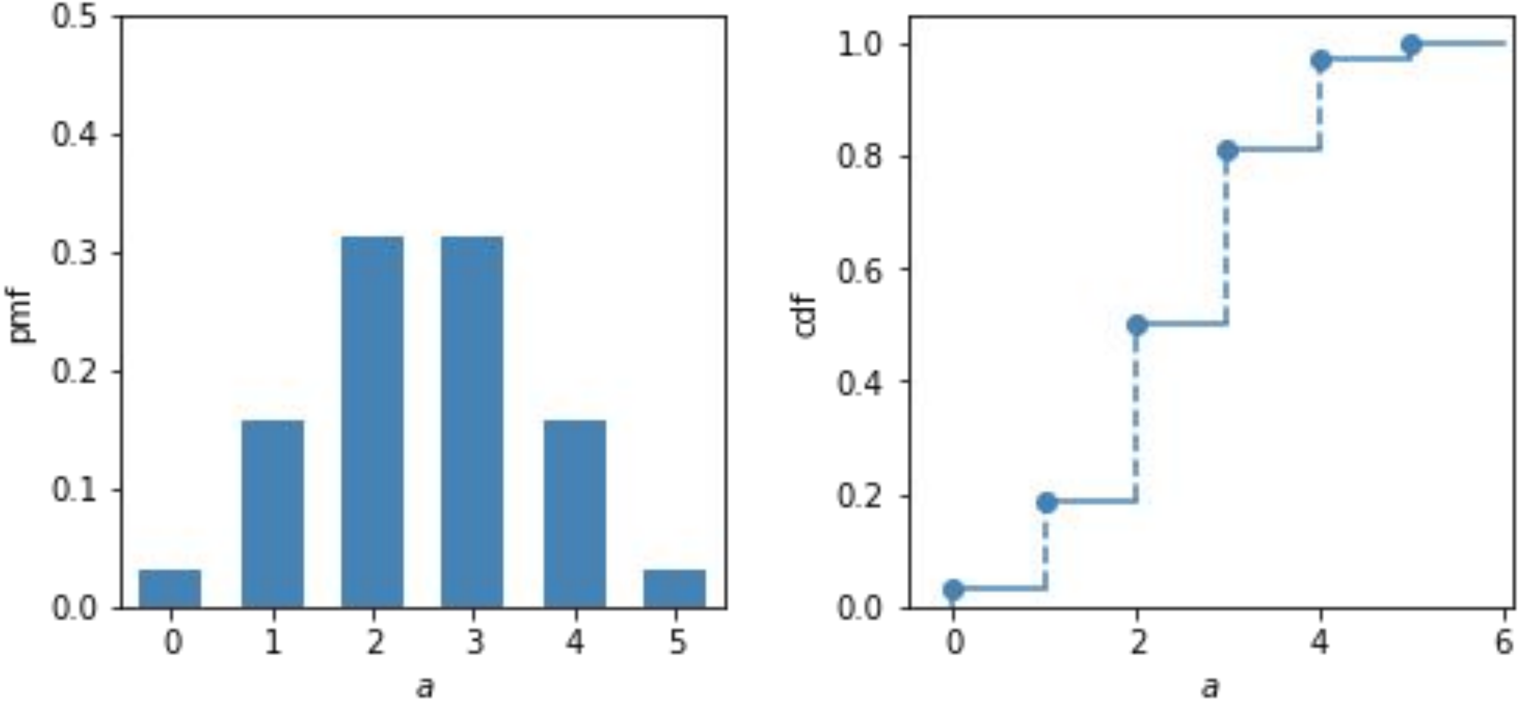
$$p_X(k) = P(X = k) = \binom{n}{k} p^k \, (1-p)^{n-k} \qquad \text{for } k = 0, 1, 2, \ldots, n$$

We denote this distribution by Bin(n, p)

$X \sim Bin(n, p)$

The Binomial distribution

Example: n=5, p=0.5

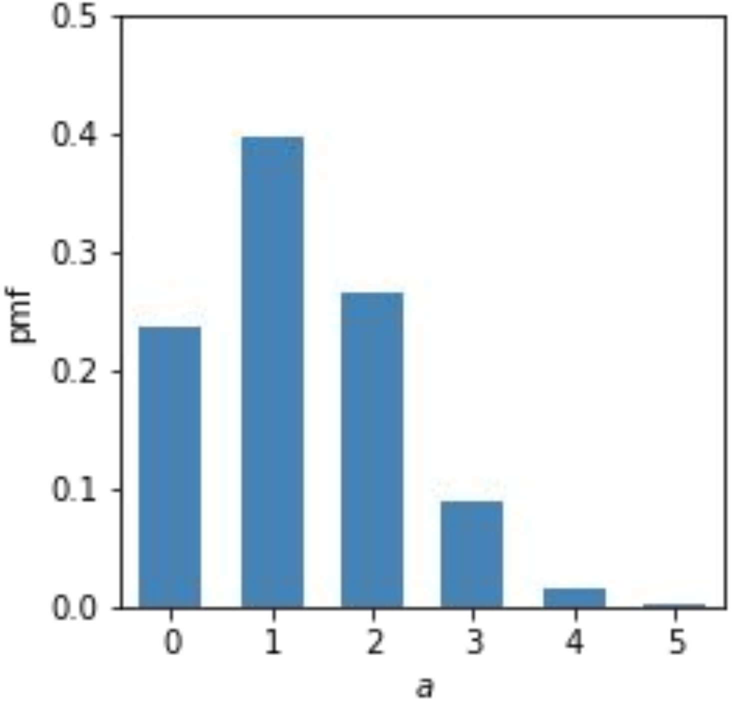


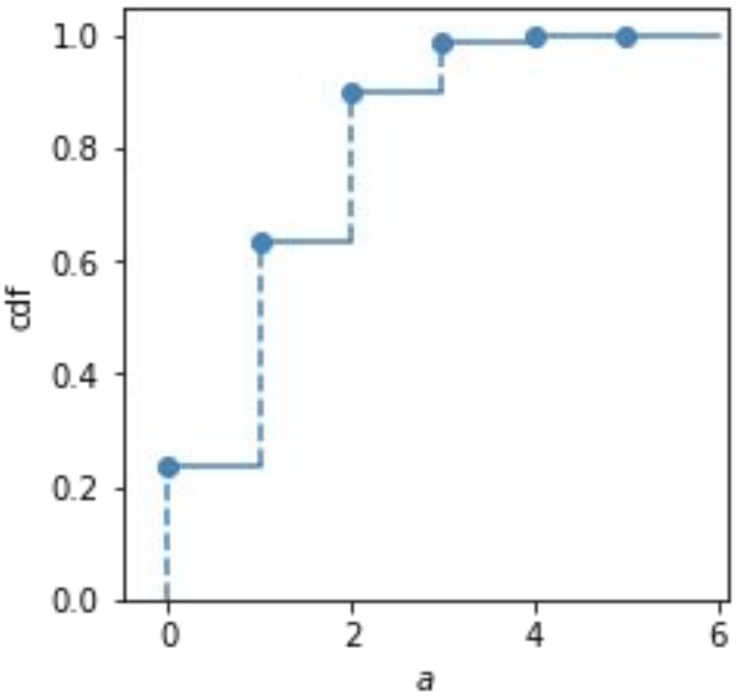
The Binomial distribution

Example: $n=5$, $p=0.25$

The Binomial distribution

Example: n=5, p=0.25



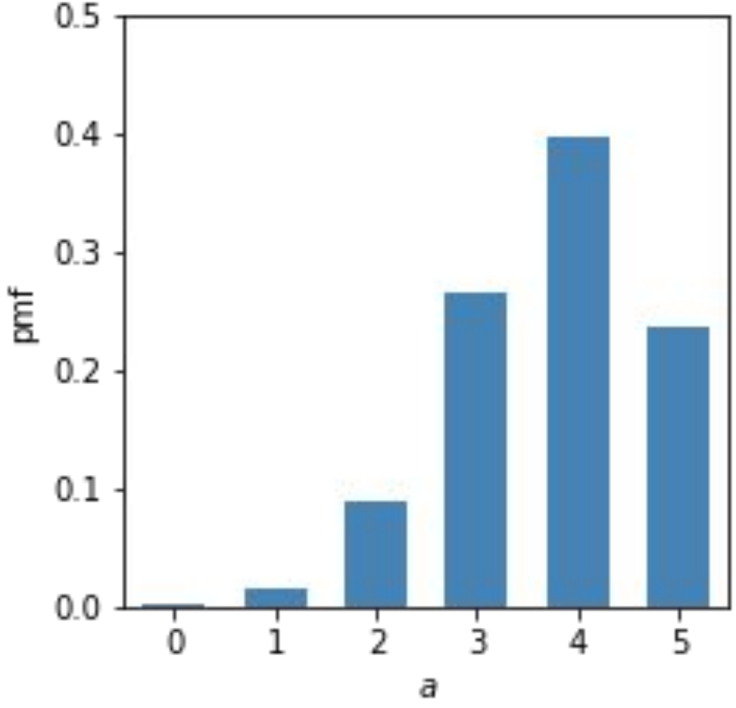


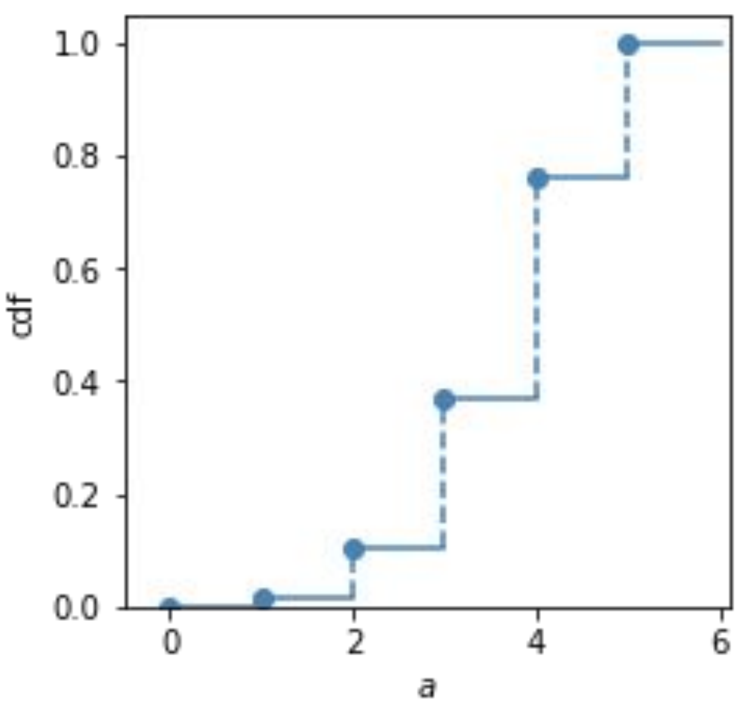
The Binomial distribution

Example: $n=5$, $p=0.75$

The Binomial distribution

Example: n=5, p=0.75





The Binomial distribution

What **assumptions** did we make in going from Ber(p) to Bin(n, p) ?

The Binomial distribution

What **assumptions** did we make in going from Ber(p) to Bin(n, p) ?

- Each of the n Bernoulli trials are independent

- Each of the Bernoulli trials has the same probability of success p

The Most Boring (but Common) Distribution of Them All

What is the distribution of a fair die?

The Most Boring (but Common) Distribution of Them All

What is the distribution of a fair die?

$p(1) = p(2) = \dots = p(6) = \frac{1}{6}$

Definition: A discrete r.v. X has a **discrete uniform distribution** with parameters *a*, *b*, and *n*=*b*-*a*+1 if

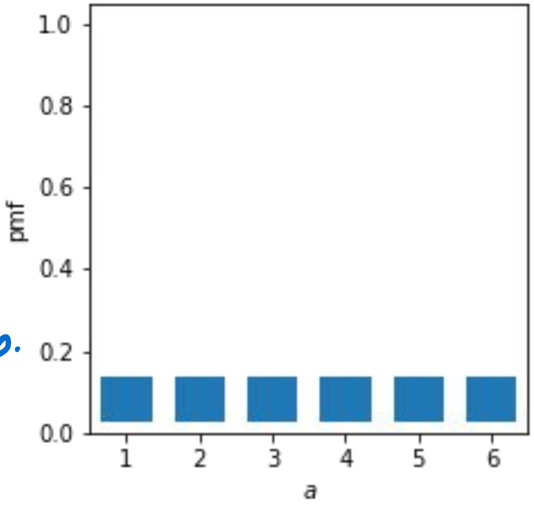
$$p_X(k) = \frac{1}{n} \quad \text{for} \quad k = a, a+1, a+2, \dots, b$$

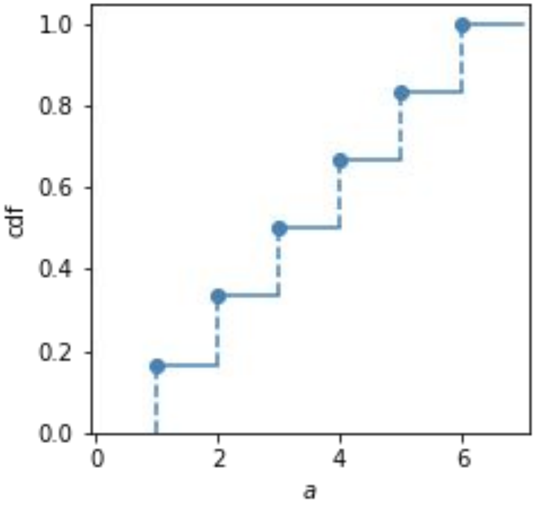
n = # possible outcomes

↓

each outcome has prob. $\frac{1}{n}$

each x[i] has prob. $\frac{1}{\text{len}(x)}$





Okay! Let's get to work!

Get in groups, get out laptops, and open **nb07** notebook

Lets…

- See some more examples of computing pmfs and dfs
- Look at some more examples of the Binomial distribution
- Learn how to sample from the Bernoulli and Binomial distributions in Numpy