CSC380: Principles of Data Science

Probability 3

Xinchen Yu

Review: "probability cheatsheet"

Additivity:

For any *finite* or *countably infinite* sequence of disjoint events $E_1, E_2, E_3, \ldots$, $P\left(\bigcup_{i\geq 1} E_i\right) = \sum_{i\geq 1} P(E_i)$

Inclusion-exclusion rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Law of total probability: For events $B_1, B_2, \ldots$ that partitions Ω, $P(A) = \sum_i P(A \cap B_i)$

Conditional probability: $P(A|B) := \frac{P(A \cap B)}{P(B)}$ ($P(A|B) \neq P(B|A)$ in general)

Probability chain rule: $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$

Law of total probability + Conditional probability: $P(A) = \sum_i P(A \cap B_i) = \sum_i P(B_i)P(A|B_i) = \sum_i P(A)P(B_i|A)$

Bayes' rule: $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$

Independence: (definition) A and B are independent if $P(A, B) = P(A)P(B)$

(property) A and B are independent if and only if $P(A|B) = P(A)$ (or $P(B|A) = P(B)$)

Outline

- Random variables
- Distribution functions
  - probability mass functions (PMF)
  - cumulative distribution function (CDF)

Random Variables

Random variables (RVs)

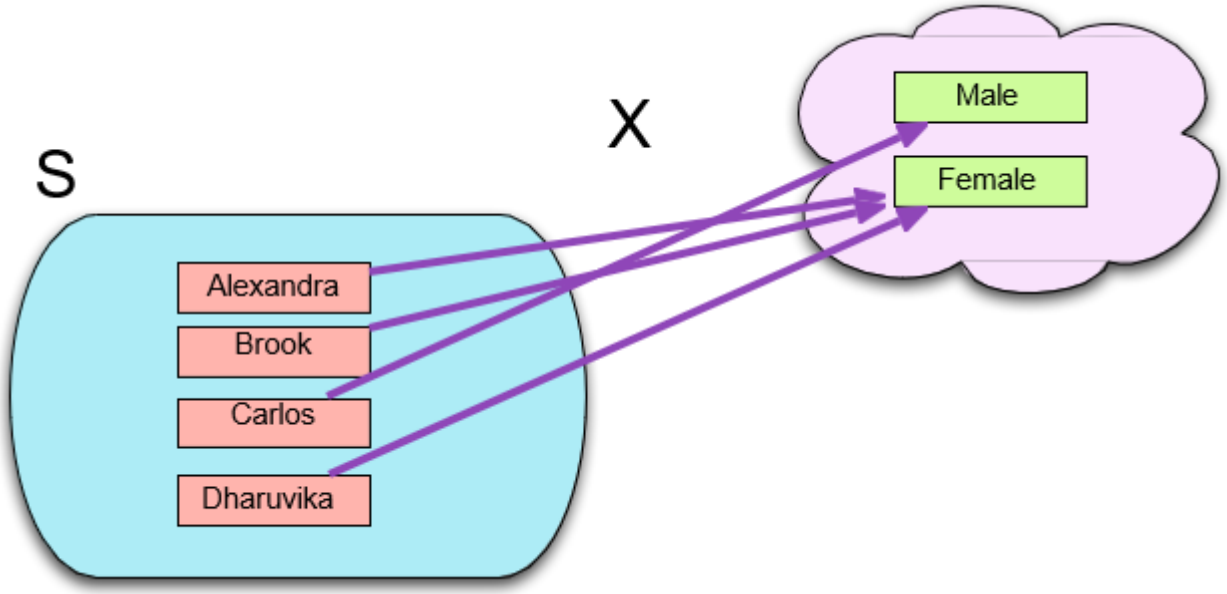
- A single random sample may have more than one characteristic that we can observe (i.e., it may be bi-/multivariate data).
- We can represent each characteristic (e.g., gender, weight, cancer status, etc.) using a separate random variable.

Random Variable

A **random variable** connects each possible outcome in the sample space to some property of interest.

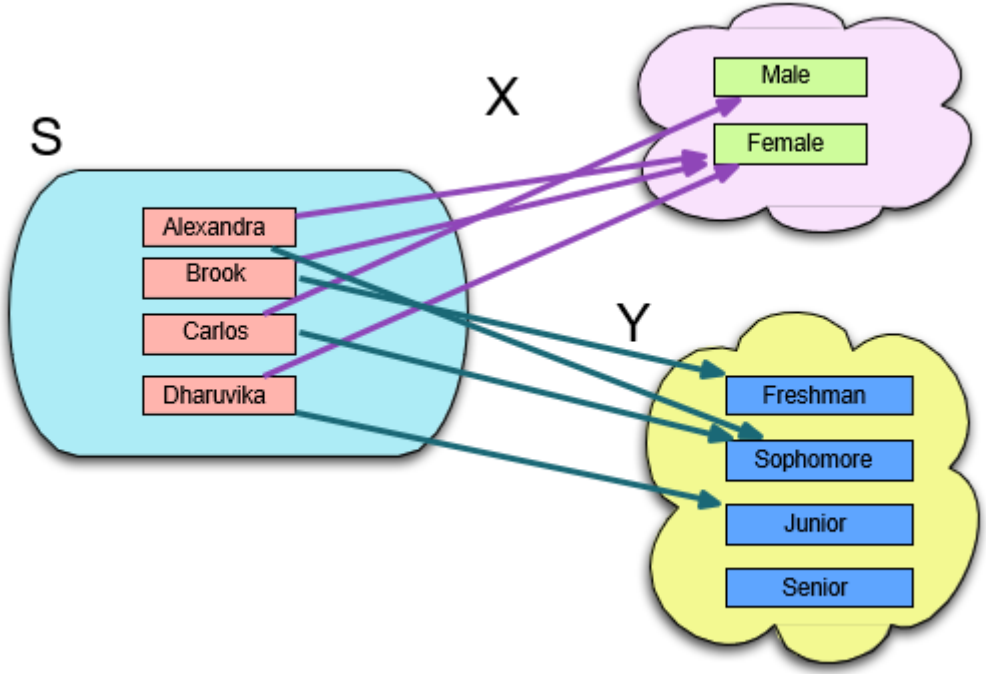
Each value of the random variable (e.g., male or female) has an associated probability.

Random Variable: Example



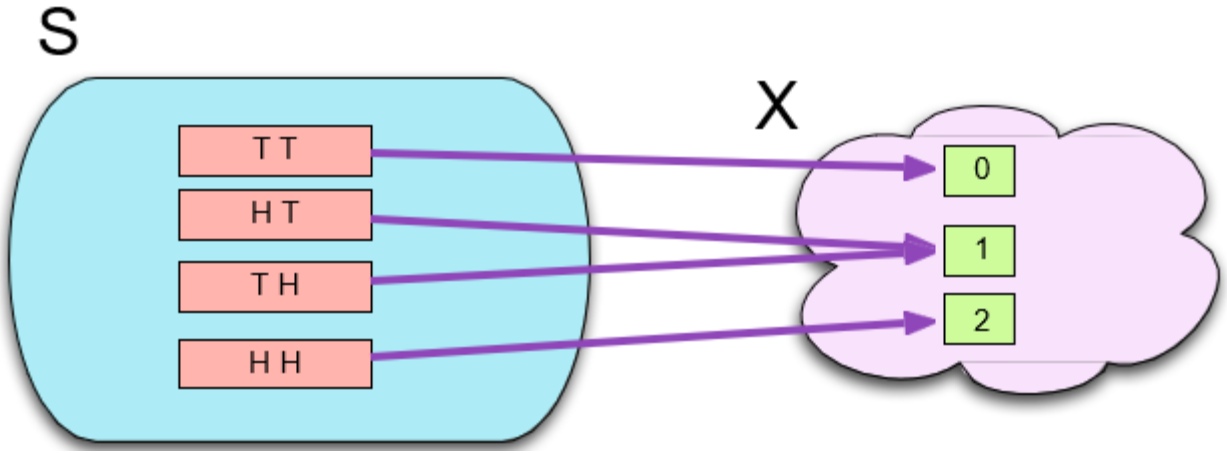
- X: people -> their genders

Random Variable: Example



- Y: people -> their class year

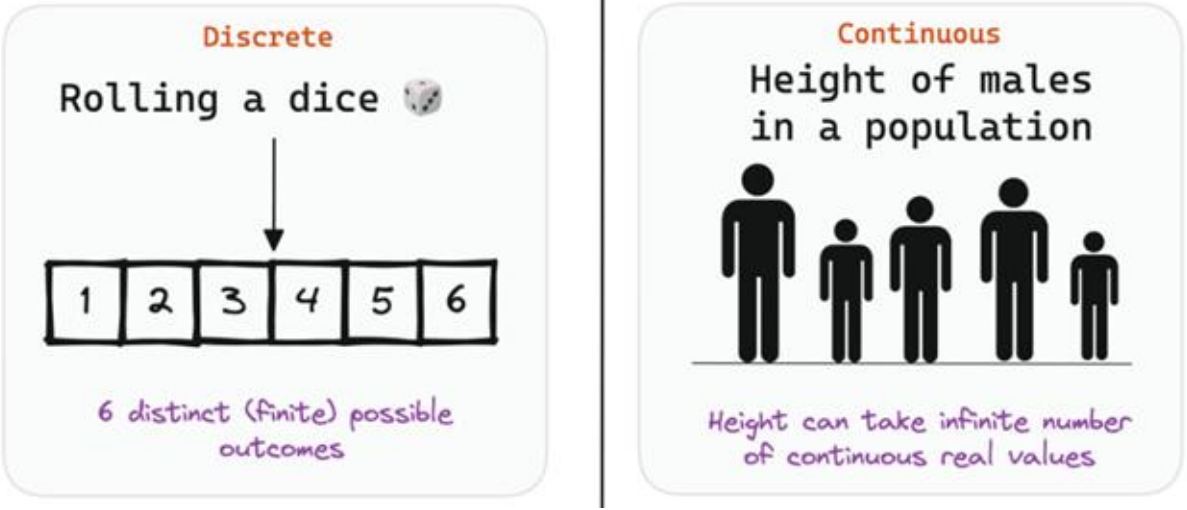
Random Variable: Example



- X: sequence of coin flips -> Number of heads

Types of Random Variables

- Discrete random variable: takes a finite or countable number of distinct values.
- Continuous random variable: takes an infinite number of values within a specified range or interval.



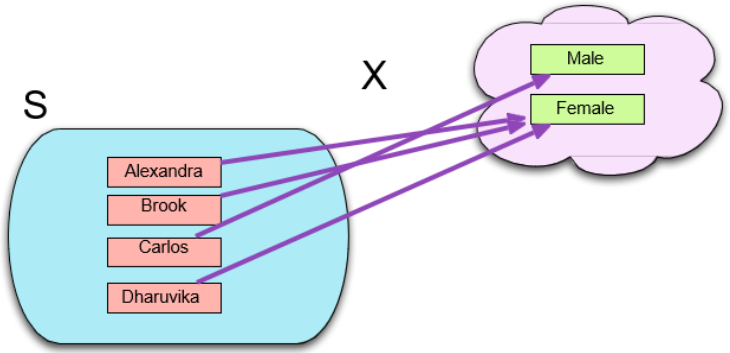
Distribution functions

Discrete distributions

- When a random variable is discrete, its *distribution* is characterized by the probabilities assigned to each distinct value.

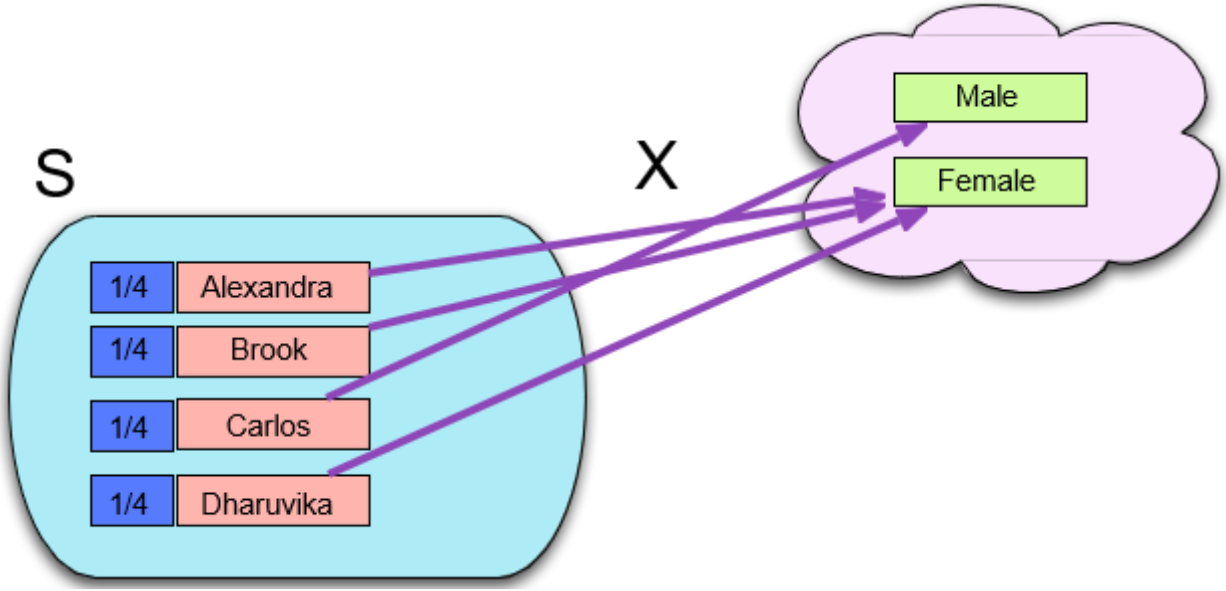
- The probability that the random variable takes a particular value comes from the probability associated with the set of individual outcomes that have that value.
  - This set is an event

- E.g. P(X = Female)



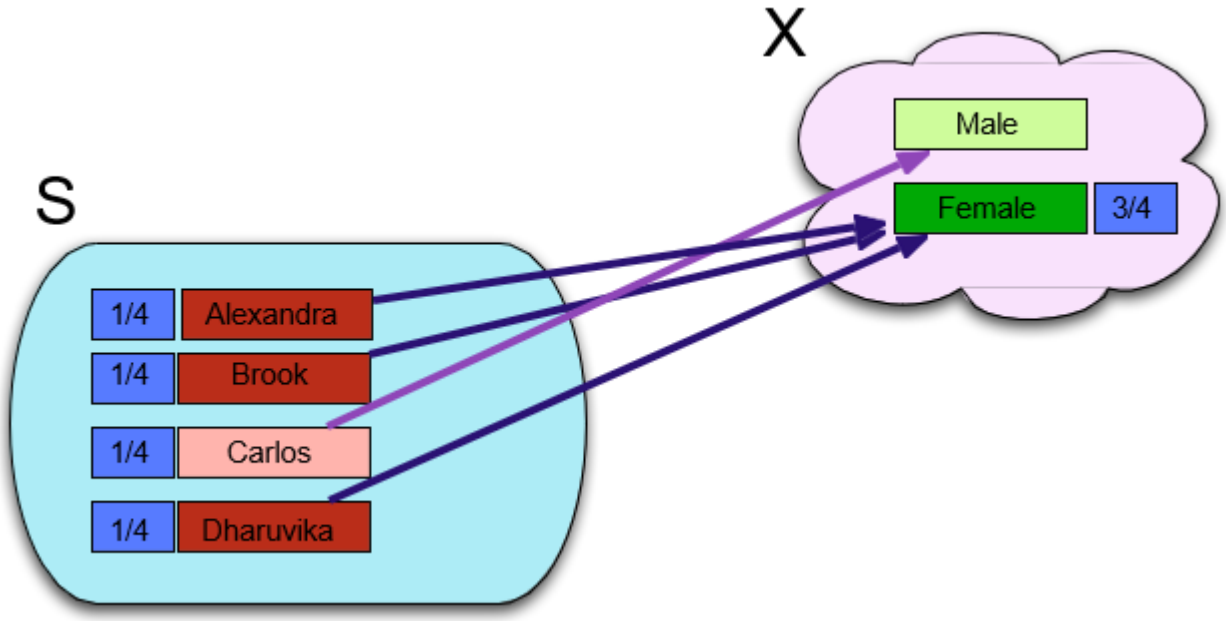
Discrete distributions

- How to find P(X = Female)?



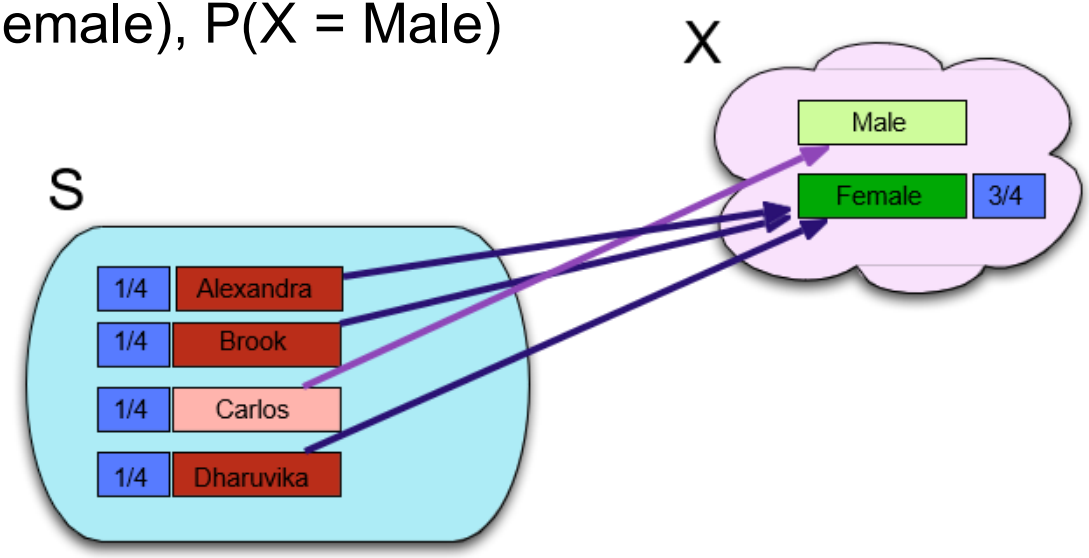
Discrete distributions

- How to find P(X = Female)?



Discrete distributions

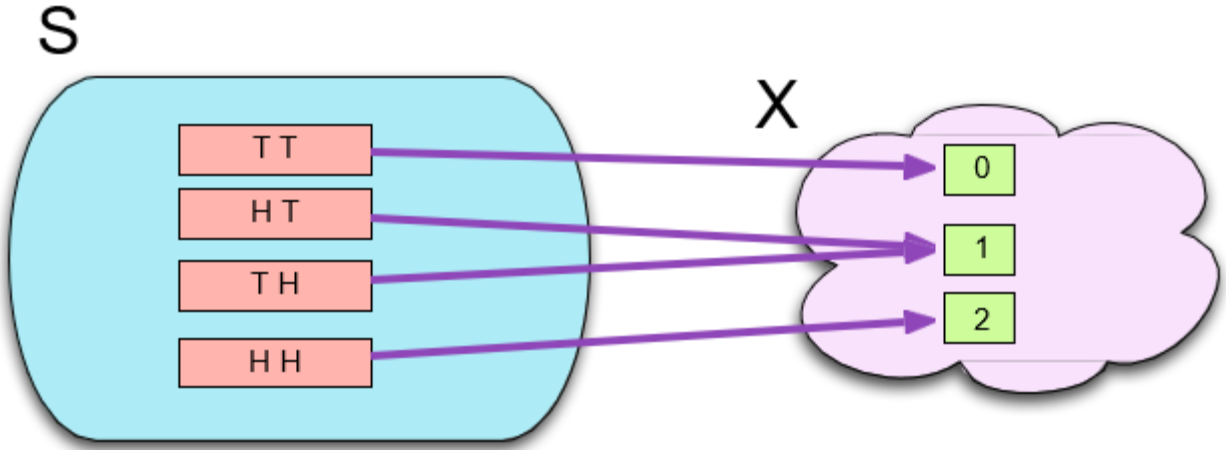
- What is the distribution of random variable X?
  - P(X = Female), P(X = Male)



x	Male	Female
$P(X = x)$	$1/4$	$3/4$

Discrete distributions

- What is the distribution of random variable X?



x	0	1	2
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Properties of Discrete Distributions

- We can write P(X = x) to mean “The probability that the random variable X takes the value x”.

- What must be true of these probabilities?

Properties of Discrete Distributions

1. Each $P(X = x)$ is a probability, so must be between 0 and 1.
2. The $P(X = x)$ must sum to 1 over all possible x values.

Probability Mass function (PMF)

The Probability Mass Function

A discrete random variable, X, can be characterized by its **probability mass function**, f (might sometimes write f_X if it's not clear from context which random variable we're talking about).

The PMF takes in values of the variable, and returns probabilities:

$$f(x) \text{ is } \textit{defined} \text{ to be } P(X = x)$$

PMF is a table

- Think of the PMF as a lookup table.

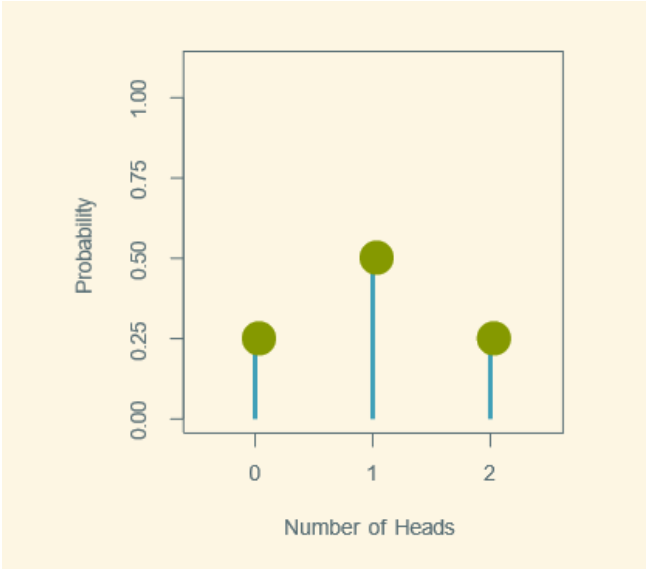
x	Male	Female
$P(X = x)$	$1/4$	$3/4$

- Best way to think of discrete random variables: they take various values, and each value has a certain probability of happening.

Visualizing discrete distributions: spike plot

Flip two coins at the same time, probability distribution of number of heads:

- Often use the spike plot
- Like a bar plot, but with probabilities, instead of frequencies or proportions, on the y-axis.



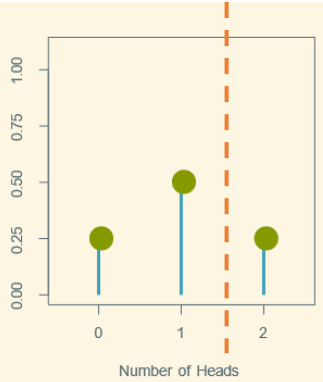
The cumulative distribution function (CDF)

- Often we are interested in the probability of falling in some range of values.
- We can use the cumulative distribution function (CDF), which gives the "accumulated probability" up to a particular value.

The Cumulative Distribution Function

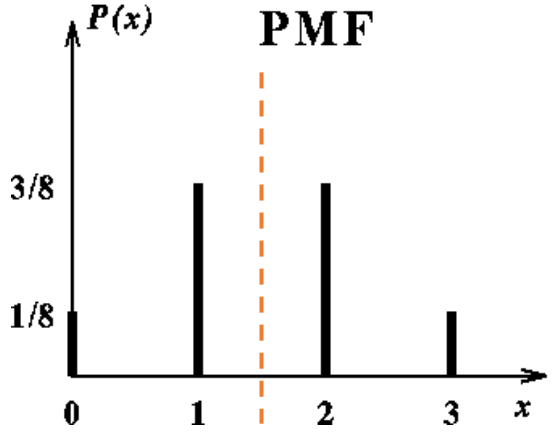
A random variable, X, can be characterized by its **cumulative distribution function**, F (or sometimes F_X if we need to be explicit), which takes values and returns *cumulative* probabilities:

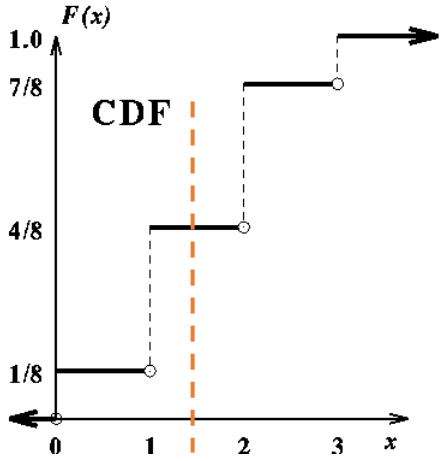
$F(x)$ is defined to be $P(X \leq x)$



Relating PMF to CDF

- How can we calculate F(x) from the PMF table f?
  - Add up all the probabilities up to and including f(x).
  - What is the value of F(−0.1) (i.e., $P(X \leq -0.1)$)? F(1.5)?

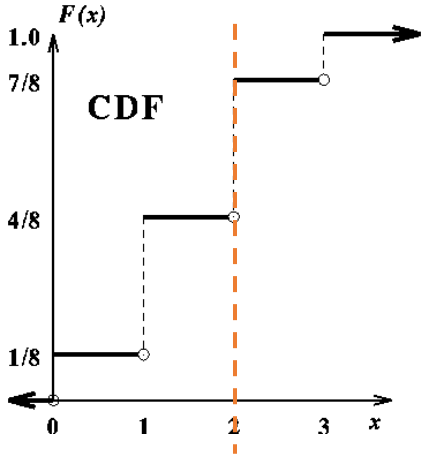




- For discrete random variables, F(x) *jumps* at locations with nonzero probability mass

Relating CDF to PMF

- How could we find f(x) from a cumulative distribution function F? e.g., f(2)?



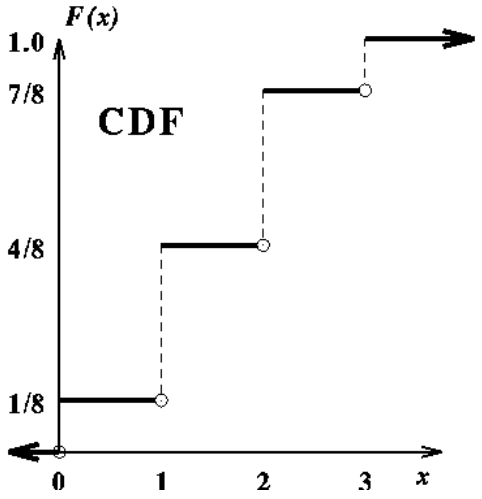
- Focus on “jumps”: f(x) = F(x) – F(jump just below x)
  - $\mathrm{f}(2) = F(2) - F(1) = \frac{7}{8} - \frac{4}{8} = \frac{3}{8}$
  - $\mathrm{f}(2.1) = F(2.1) - F(2) = \frac{7}{8} - \frac{7}{8} = 0$
  - $\mathrm{f}(1.5) = F(1.5) - F(1) = \frac{4}{8} - \frac{4}{8} = 0$

Exercise: using CDF and PMF

Given the CDF F:

- How to calculate $P(X > x)$?
  - $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$

- How about $P(X \geq x)$?
  - $P(X \geq x) = 1 - P(X < x) = 1 - (P(X \leq x) - P(X=x))$
  - $1 - F(x) + f(x)$
  - $f(x)$ can be 0 or nonzero, depending on whether x is a jump



Exercise: using CDF and PMF

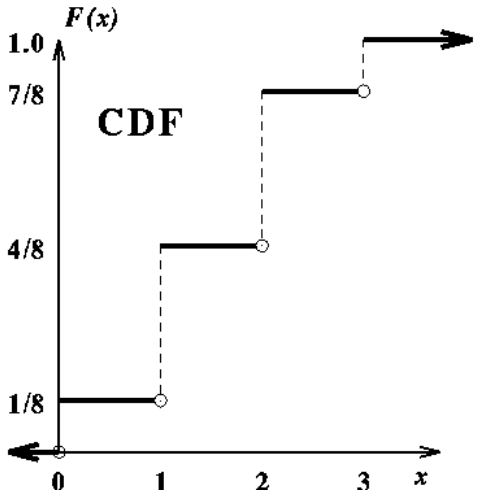
Given the CDF F:

- How to calculate $P(a < X \leq b)$?

  $= P(X \leq b) - P(X \leq a)$

  $= F(b) - F(a)$

- How to calculate $P(a < X < b)$?
  - (I'll leave this to you as an exercise..)



Transformations of random variables

- If X is a random variable, then $X + 5, 3X, X^2, \ldots$, are all random variables
- Given any transformation function f, $f(X)$ is a random variable
- How to find the PMF of $f(X)$ based on that of X?
  - First, find all values $f(X)$ can take
  - For each value c, try to find $\mathrm{P}(f(X) = c)$