## Foundations of Data Analysis

Broadly speaking we will cover five topics:

- Probability
- Bayesian inference
- Maximum likelihood inference
- Hypothesis testing
- Regression

Probability: a mathematical framework for reasoning about uncertainty

- Probabilistic models
  - Sample space
  - Probability function

Sample space:

Given an "experiment" (some process of observation):

- Sample space $S$ = set of all possible outcomes

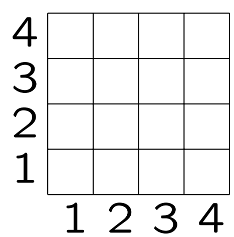
- Set $S$ must be:
  - Mutually exclusive
  - Collectively exhaustive

## Sample space: Discrete example

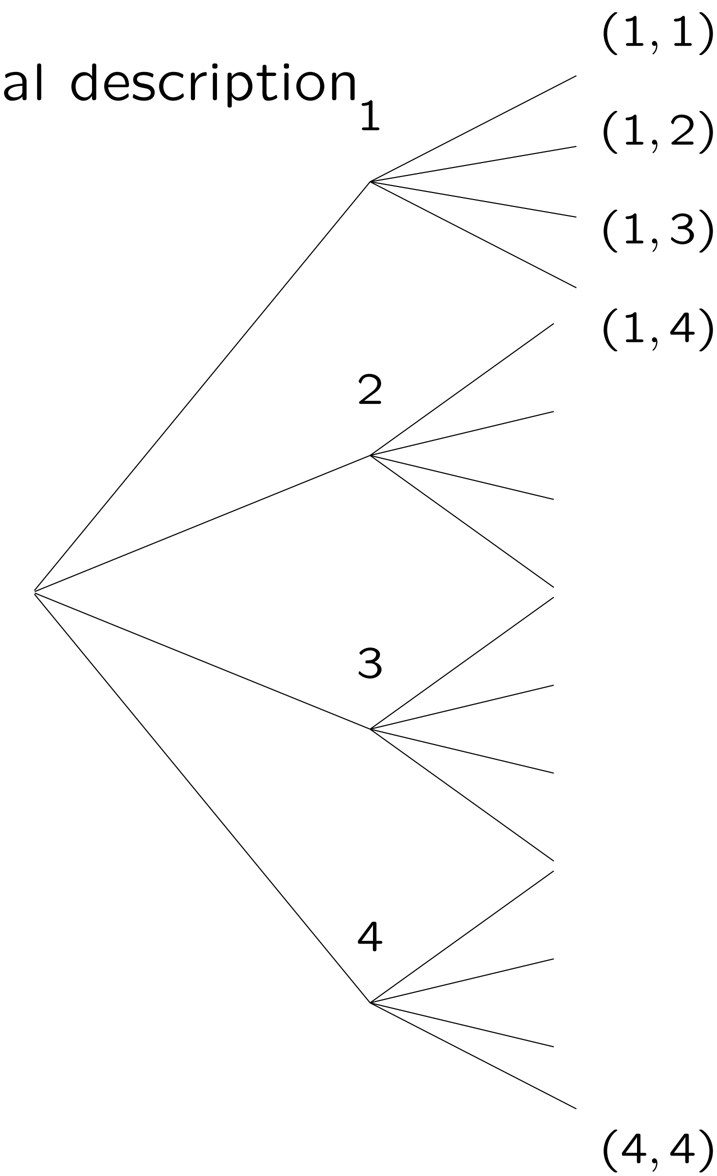
Two rolls of tetrahedral die

Sample space vs. sequential description

$Y =$ 2nd roll

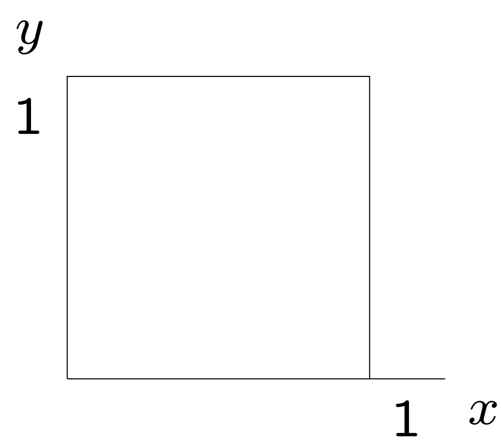



$X =$ 1st roll



## Sample space: Continuous example

Unit square: $S = \{(x, y) | 0 \leq x, y \leq 1\}$



Event: subset of the sample space (set of outcomes)

For any events $A$ and $B$:

- $A^c$ = complement of $A$ ("not $A$")

- $A \cap B$ = intersection of $A$ and $B$

- $A \cup B$ = union of $A$ and $B$

- $A$ and $B$ are mutually exclusive if $A \cap B = \emptyset$
  - Example: $(A \cap B^c) \cup (A \cap B) = A$

## Probability

Def: For an experiment with sample space $S$, probability is a function that assigns a number $P(A)$ to an event $A \subseteq S$ so that the following axioms hold:

1. Nonnegativity: $0 \leq P(A) \leq 1$

2. Normalization: $P(S) = 1$

3. Additivity: If $A \cap B = \emptyset$, $P(A \cup B) = P(A) + P(B)$

Are these axioms going to be enough?

Generalization of Axiom 3: If $A_1, \ldots, A_n$ disjoint,
$P(A_1 \cup \ldots \cup A_n) = P(A_1) + \ldots + P(A_n)$
Proof: By induction on $n$

Special case: $S$ consists of finite # of possible outcomes

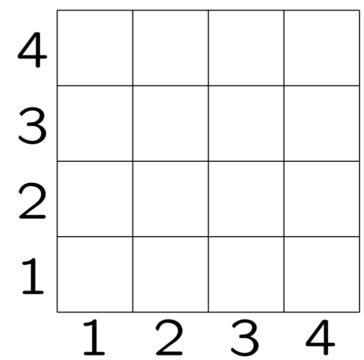
Then if $s_1, s_2, \ldots, s_k \in S$,

$$\begin{aligned}
P(\{s_1, s_2, \ldots, s_k\}) &= P(\{s_1\}) + P(\{s_2, \ldots, s_k\}) \\
&\ \ \vdots \\
&= P(\{s_1\}) + P(\{s_2\}) + \ldots + P(\{s_k\}) \\
&= P(s_1) + P(s_2) + \ldots P(s_k) \text{ (drop \{ \})}
\end{aligned}$$

Do all subsets of any sample space have probabilities?
(No, but not see these sets in this course)

## Probability: Example with finite sample space

$Y =$ 2nd roll




$X =$ 1st roll

Let every outcome have probability 1/16:

- $P(\{X = 1\}) =$

- $P(X + Y \text{ is odd}) =$

- $P(\min(X, Y) = 2) =$

## Discrete uniform probability

- Let all outcomes be equally likely

- Then

$$P(A) = \frac{\#\text{ elements of } A}{\text{total }\#\text{ of sample points}}$$

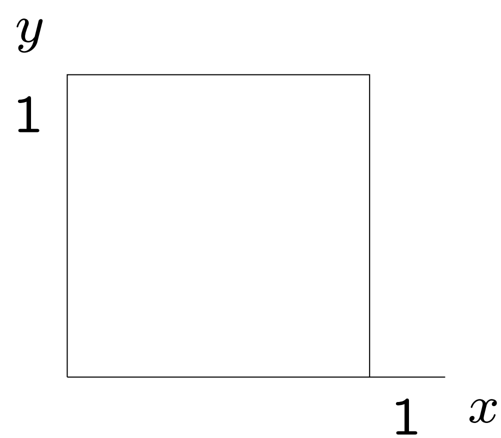
- Computing probabilities reduces to counting

- Defines fair coins, fair dice, shuffled decks of cards

## Continuous uniform probability

Unit square: $S = \{(x, y) | 0 \leq x, y \leq 1\}$

Uniform probability: $P(A) = \text{area}(A)$, $A \subseteq S$




- $P((X, Y) = (0.5, 0.75)) =$

- $P(X + Y \leq 0.5) =$

## Probability: Example with countably infinite sample space

- Flip fair coin until a tail occurs; outcome # of flips
- Sample space: $\{1, 2, 3, \ldots\}$
- $P(n) = (\frac{1}{2})^n$, $n = 1, 2, 3, \ldots$
- Find $P$(odd # of flips)
- $P(\{1, 3, \ldots\}) = P(1) + P(3) + \ldots = \frac{1}{2} + (\frac{1}{2})^3 + \ldots = \frac{2}{3}$

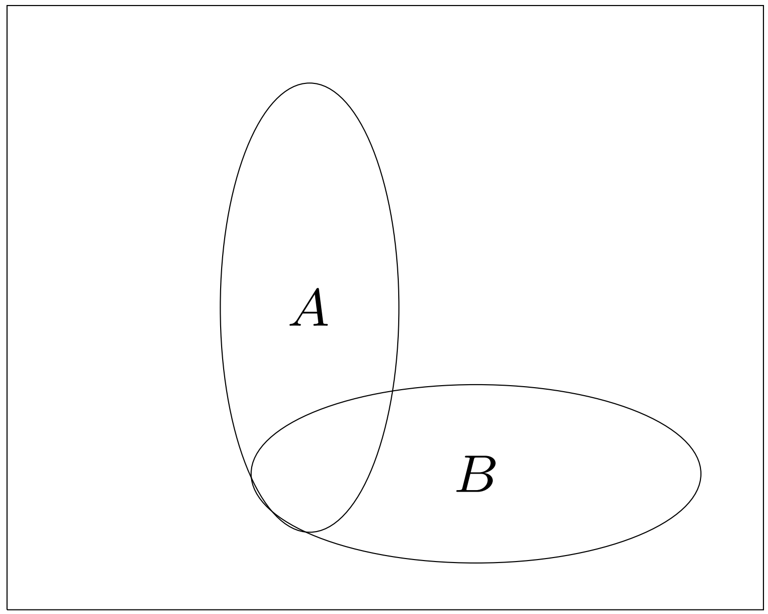
Introduce stronger form of Axiom 3:

Axiom 3: Countable additivity

If $A_1, A_2, \ldots$ is a sequence of disjoint events, then

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$$

Conditional probability




$P(A|B) =$ probability of $A$, given that $B$ occurred

$B$ is the new sample space

Def: (Conditional probability) If $P(B) > 0$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

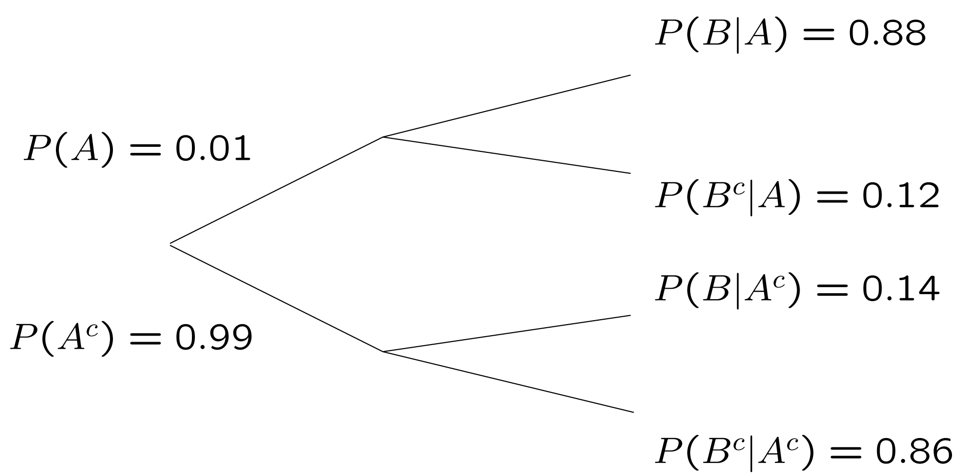
Multiplication rule:

$$\begin{aligned} P(A \cap B) &= P(B) \cdot P(A|B) \text{ (even if } P(B) = 0) \\ &= P(A) \cdot P(B|A) \end{aligned}$$

## Probability models based on conditional probabilities

Event $A$: Subject is lying

Event $B$: Polygraph test is positive



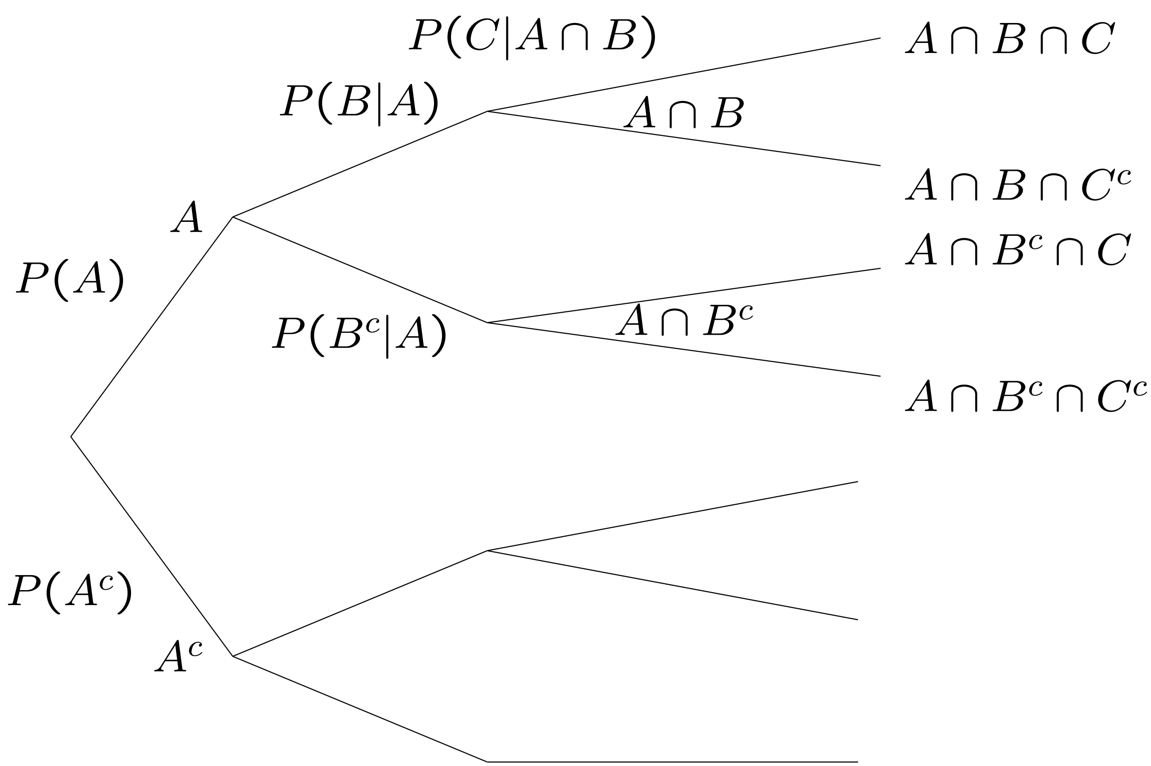

- $P(A \cap B) =$

- $P(B) =$

- $P(A|B) =$

## Multiplication rule

$$P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|A \cap B)$$



Law of total probability

- Divide and conquer

- Partition sample space into disjoint events $A_1, \ldots, A_n$

- Know $P(A_i)$ and $P(B|A_i)$ for every $i$

- Then can compute $P(B)$:

$$
\begin{aligned}
P(B) = & \quad P(A_1) \cdot P(B|A_1) \\
& + \; P(A_2) \cdot P(B|A_2) \\
& \qquad \vdots \\
& + \; P(A_n) \cdot P(B|A_n)
\end{aligned}
$$

Bayes' law

- Know "prior" probabilities $P(A_i)$ for each $i$

- Know $P(B|A_i)$ for each $i$

- Wish to compute $P(A_i|B)$, i.e., revise probabilities $P(A_i)$, given that $B$ occurred:

$$
\begin{aligned}
P(A_i|B) &= \frac{P(A_i \cap B)}{P(B)} \\
&= \frac{P(A_i) \cdot P(B|A_i)}{P(B)} \\
&= \frac{P(A_i) \cdot P(B|A_i)}{\sum_j P(A_j) \cdot P(B|A_j)}
\end{aligned}
$$

## Independence

- Intuitively, $A$ is independent of $B$ if occurrence of $B$ provides no information about $A$'s occurrence, i.e.,

  $$P(A|B) = P(A)$$

  If

  $$P(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  then

  $$P(A \cap B) = P(A) \cdot P(B)$$

- Def: If $P(A \cap B) = P(A) \cdot P(B)$, then $A$ and $B$ are independent events

- Symmetric in $A$ and $B$: $P(A|B) = P(A) \Rightarrow P(B|A) = P(B)$

- If $A$ and $B$ are disjoint events, can $A$ and $B$ be independent?

- Are $A$ and $A^c$ independent?

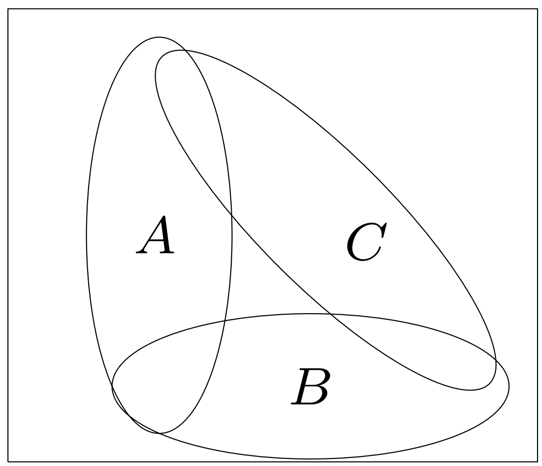
- If $A$ and $B$ are independent, are $A$ and $B^c$ independent?

## Conditioning may affect independence

- Given an event $C$, $A$ and $B$ are conditionally independent if

$$P(A \cap B \;|C) = P(A|C) \cdot P(B|C)$$

- Suppose $A$ and $B$ are independent




If $C$ is known to occur, are $A$ and $B$ independent?

Independence of multiple events

Def: Events $A_1, A_2, \ldots, A_n$ are independent if

$$P(A_i \cap A_j \cap \cdots \cap A_t) = P(A_i) \cdot P(A_j) \cdots P(A_t)$$

for any distinct indices $i, j, \ldots, t$ chosen from $\{1, \ldots, n\}$

i.e., the occurrence or nonoccurrence of any number of the events carries no information on the remaining events

## Random variables

Def: A random variable is a function from the sample space $S$ to the real numbers $\mathbb{R}$

- assigns a value (number) to each possible outcome of sample space

- discrete r.v.: finite or countable # of values

- continuous r.v.: values form a set of real #'s

- Notation:

    - random variable $X$

    - numerical value $x$

Example: Suppose r.v. $X =$ "total # of Heads" in two independent coin flips

Can make a table that displays $X$ as a function on $S$:

| $s \in S$ | $TT$ | $HT$ | $TH$ | $HH$ |
|---|---|---|---|---|
| $X(s)$ | 0 | 1 | 1 | 2 |

## Indicator random variable

For an event $A$, define indicator random variable $I_A$ by

$$I_A(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{if } s \notin A \end{cases}$$

- $I_A$ is 1 if $A$ occurs, 0 if $A$ does not

- Use indicator r.v.'s to do counting:

  Example: let $Y = \#$ of heads in $n$ coin flips

  $Y = X_1 + X_2 + \ldots + X_n$, where $X_i$ is indicator of event "heads on $i$th toss"

  $X_i$ counts 1 for every head, 0 for every tail

## Probability mass function (PMF)

Probability mass function (PMF) $p_X(x)$ of a discrete r.v. $X$ is a function that assigns to each possible value $x$ of $X$ its probability:

$$\begin{aligned} p_X(x) &= P(X = x) \\ &= P(\{s \in S : X(s) = x\}) \end{aligned}$$

Note: $p_X(x) \geq 0$, $\sum_x p_X(x) = 1$

Why?

Example:

- $X$: # heads in 2 independent flips of fair coin

- table:

| $x$ | 0 | 1 | 2 |
|---|---|---|---|
| $p_X(x)$ | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |

- PMF of $X$ is

$$p_X(x) = \begin{cases} 1/4 & \text{if } x = 0 \text{ or } x = 2 \\ 1/2 & \text{if } x = 1 \\ 0 & \text{o.w.} \end{cases}$$

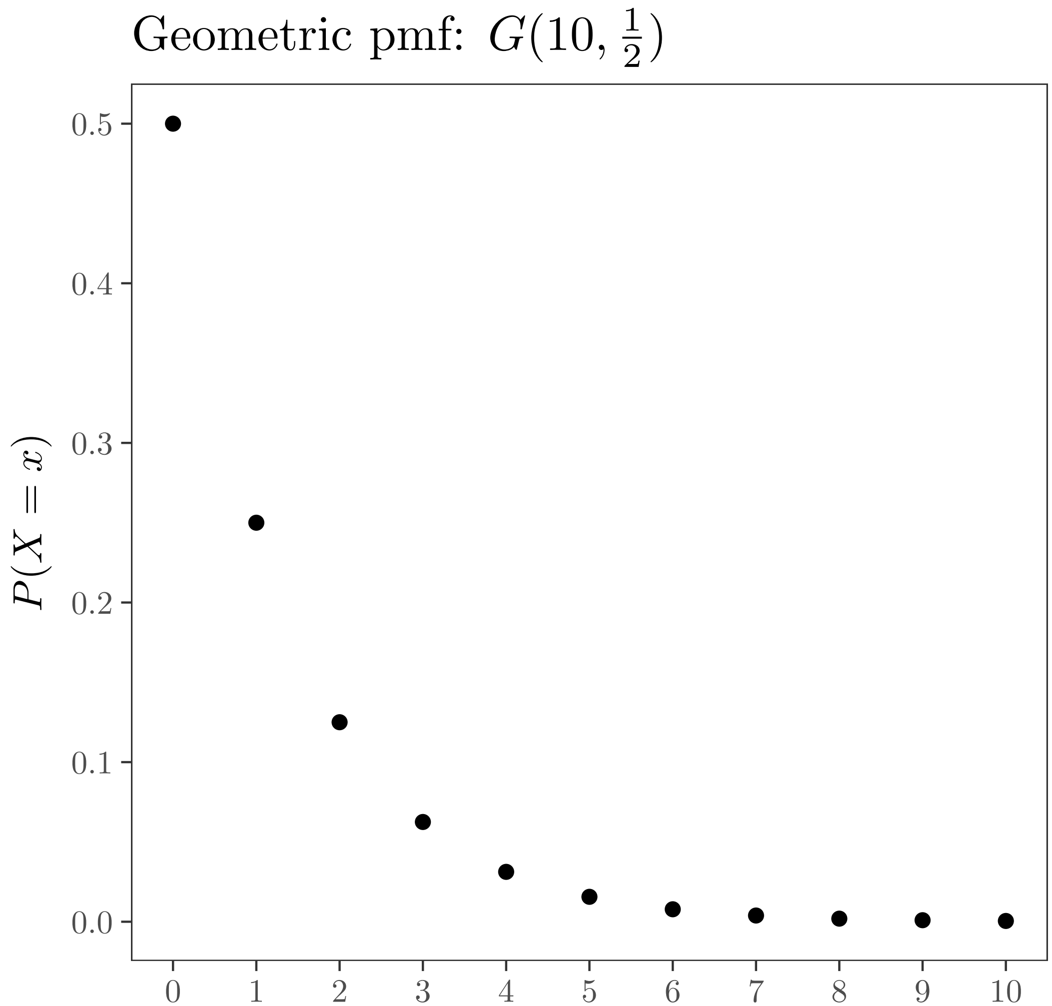
Example: flip a coin repeatedly until obtain a head

- $X = \#$ of coin flips until 1st head

- calculate PMF of $X$:
  assume independent flips and $P(H) = p > 0$

$$\begin{aligned} p_X(k) &= P(X = k) \\ &= P(TT \cdots TH) \\ &= (1-p)^{k-1}p, \ \ k = 1, 2, \ldots \end{aligned}$$

- geometric PMF: $p_X(k) = (1-p)^{k-1}p, \ \ k = 1, 2, \ldots$



Geometric pmf: $G(10, \frac{1}{2})$
$P(X = x)$
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
9
10

## How to compute the PMF $p_X(x)$

For each possible value $x$ of $X$:

- collect all possible outcomes for which $X = x$
- add their probabilities to obtain $p_X(x)$

Example: Binomial PMF

$X$: # heads in $n$ independent coin flips

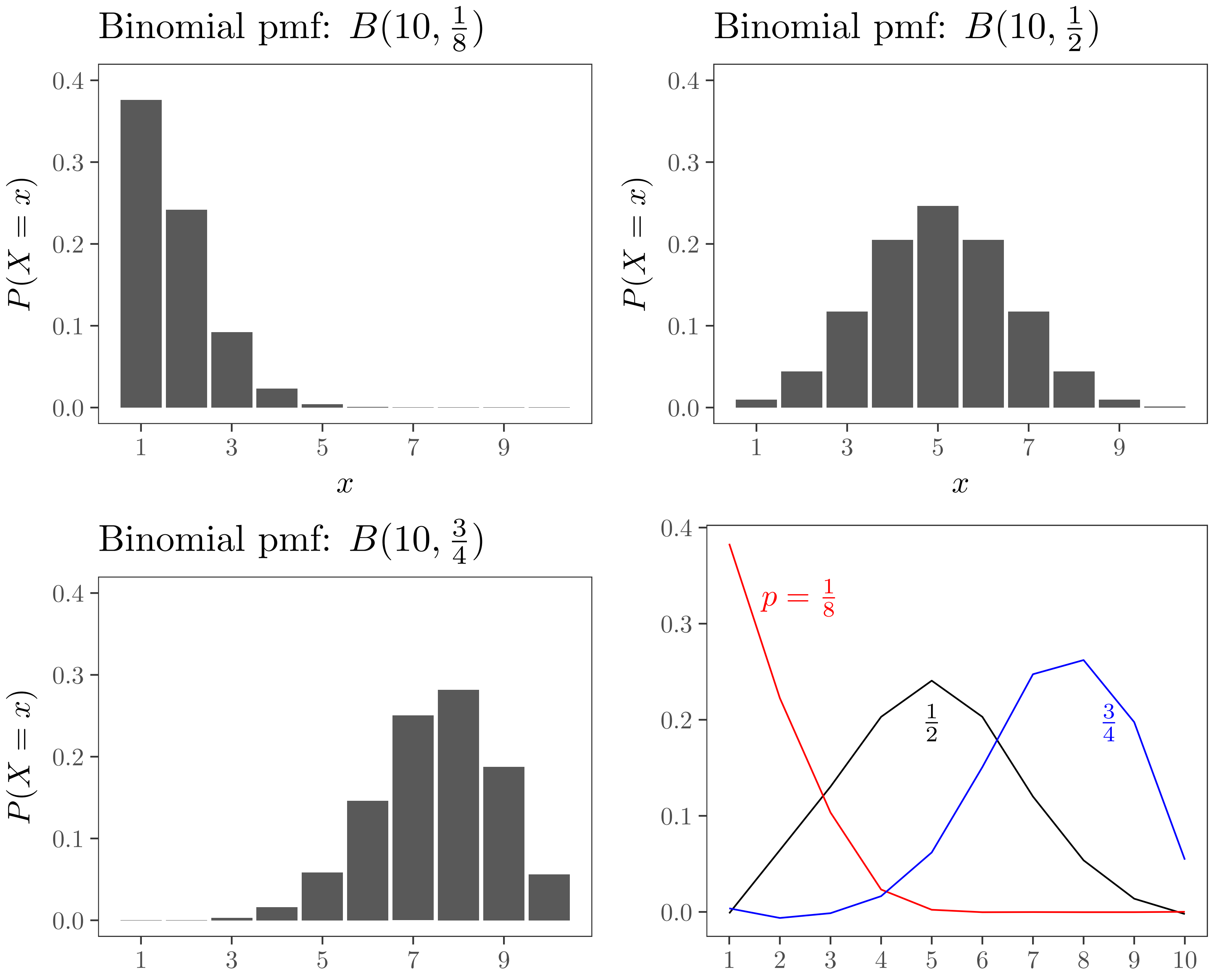
$P(H) = p$

let $n = 4$:

$$\begin{aligned} p_X(2) &= P(X = 2) \\ &= P(HHTT) + P(HTHT) + P(HTTH) \\ &\quad + P(THHT) + P(THTH) + P(TTHH) \\ &= 6p^2(1-p)^2 \\ &= \binom{4}{2} p^2 (1-p)^2 \end{aligned}$$

In general, for $n$ trials, probability of $k$ successes:

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




$\binom{n}{k}p^k(1-p)^{n-k}$ for $1, \cdots$ to $k$

Bernoulli random variable

Bernoulli r.v. describes success or failure in a single trial:

$$X = \begin{cases} 1 & \text{if success} \\ 0 & \text{o/w} \end{cases}$$

Its PMF is

$$p_X(k) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}$$

## Cumulative distribution function (CDF)

The cumulative distribution function (CDF) $F_X$ of a r.v. $X$ is the function $F_X : \mathbb{R} \to [0,1]$ with

$$F_X(x) = P(X \leq x) \quad \text{(for } x \in \mathbb{R}\text{)}$$

i.e., the probability of event $\{X \leq x\}$ where $\{X \leq x\}$ is the subset $\{s \in S : X(s) \leq x\}$

If $X$ is a discrete r.v.

$$F_X(x) = P(X \leq x) = \sum_{k \leq x} p_X(k)$$

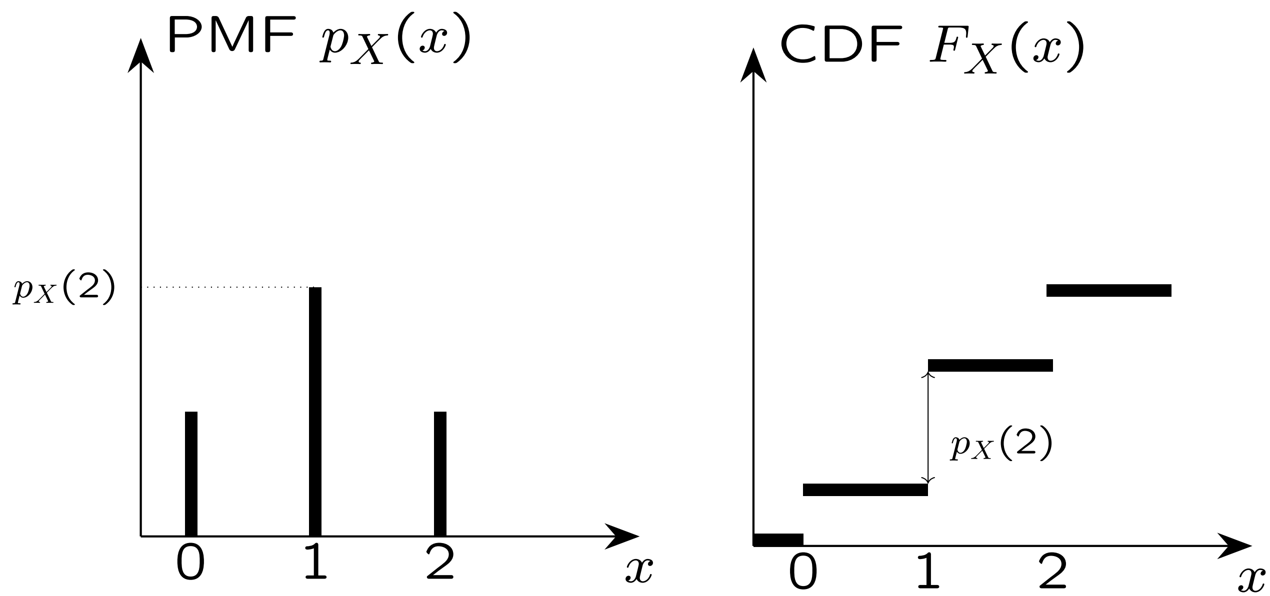
Example: $X$: # heads in 2 independent flips of fair coin

| $x$ | 0 | 1 | 2 |
|---|---|---|---|
| $p_X(x)$ | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |
| $F_X(x)$ | $\frac{1}{4}$ | $\frac{1}{4}+\frac{1}{2}=\frac{3}{4}$ | $\frac{3}{4}+\frac{1}{4}=1$ |

- What is $\{X \leq 0\}$? What is $P(\{X \leq 0\})$?

- What is $\{X \leq 1.5\}$? What is $P(\{X \leq 1.5\})$?

- What is $\{X \leq 3\}$? What is $P(\{X \leq 3\})$?



PMF $p_X(x)$
$p_X(2)$
0
1
2
$x$
CDF $F_X(x)$
$p_X(2)$
0
1
2
$x$

- CDF is related to PMF by the formula

$$F_X(x) = P(X \leq x) = \sum_{k \leq x} p_X(k)$$

  and has staircase form, with jumps at the values of positive PMF

- Size of jump at each $x$ equals $p_X(x) = P(X = x)$

- PMF can be obtained from the CDF: $p_X(k) = F_X(k) - F_X(k-1)$

## Continuous random variables and PDFs

A random variable is continuous if it can take any value within a finite or infinite interval of the real number line (and $\therefore$ its values cannot be listed sequentially)

Example: uniform random variable on interval $[0, 1]$

- model for what mean by "choose a number at random between 0 and 1"

- any real number in $[0, 1]$ is possible outcome

- "at random" means any two subintervals of same length have same probability

- probability that $X$ is in any subinterval of length $l$ equals $l$

## Probability density function (PDF)

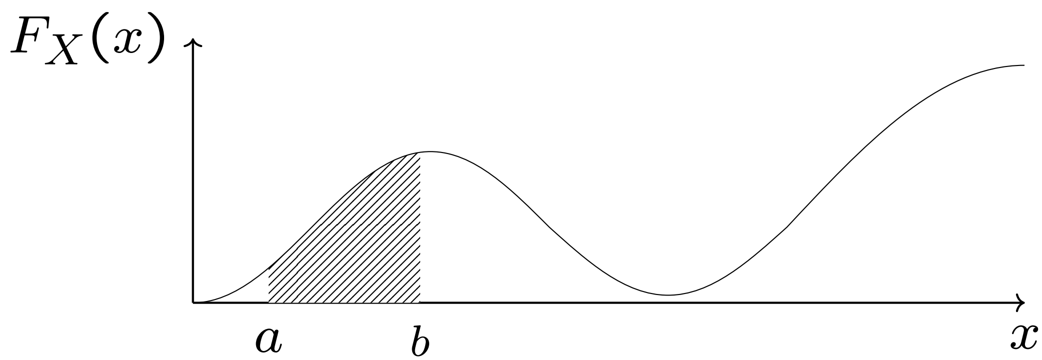
- Because the values of a continuous random variable cannot be listed, their probabilities cannot be listed

- Describe probabilities by density function $f(x)$ with properties:

  - $f(x) \geq 0$
  - $f$ is piecewise continuous
  - $\int_{-\infty}^{\infty} f_X(x)dx = 1$

<u>Def</u>: The <u>probability density function (PDF)</u> of a continuous r.v. $X$ is a function $f_X(x)$ such that for any real numbers $a < b$, the probability that $X$ falls in interval $[a, b]$ is the area under the PDF between $a$ and $b$:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$




shaded area is $\int_a^b f(x)$, the probability that $a \leq X \leq b$

- For any single value $a$, $P(X=a) = \int_{-a}^{a} f_X(x)dx = 0$

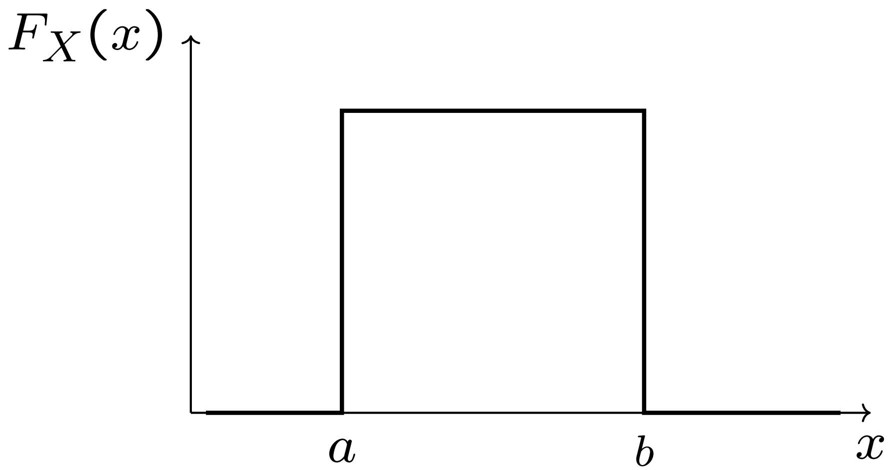
- For this reason, including or excluding the endpoints of an interval has no effect on its probability:

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

  Note: this is not true for a discrete r.v.

- To qualify as a PDF, $f_X(x) \geq 0$ for every $x$ and $\int_{-\infty}^{\infty} f_X(x)dx = P(-\infty < X < \infty) = 1$, i.e., entire area under graph of PDF must equal 1

Example: uniform r.v. $X$ on general interval $[a, b]$

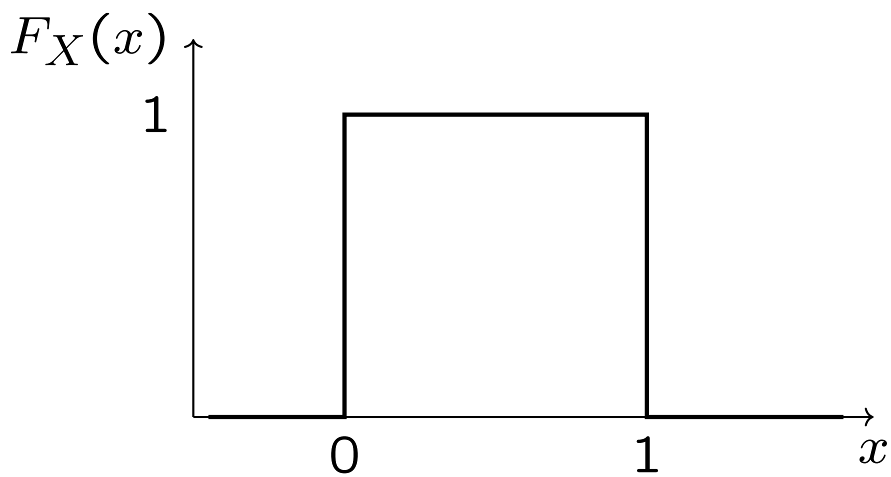



PDF of $X$ is:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & a \leq x \leq b \\ 0 & x < a \text{ or } x > b \end{cases}$$

Why?

Example: uniform r.v. $X$ on interval $[0, 1]$




PDF is:

$$f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & x < 0 \text{ or } x > 1 \end{cases}$$

## CDF of continuous random variable

The CDF $F_X$ of a continuous r.v. $X$ is defined in the same way as for a discrete r.v.:

$$F_X(x) = P(X \leq x) \text{ for all } x$$

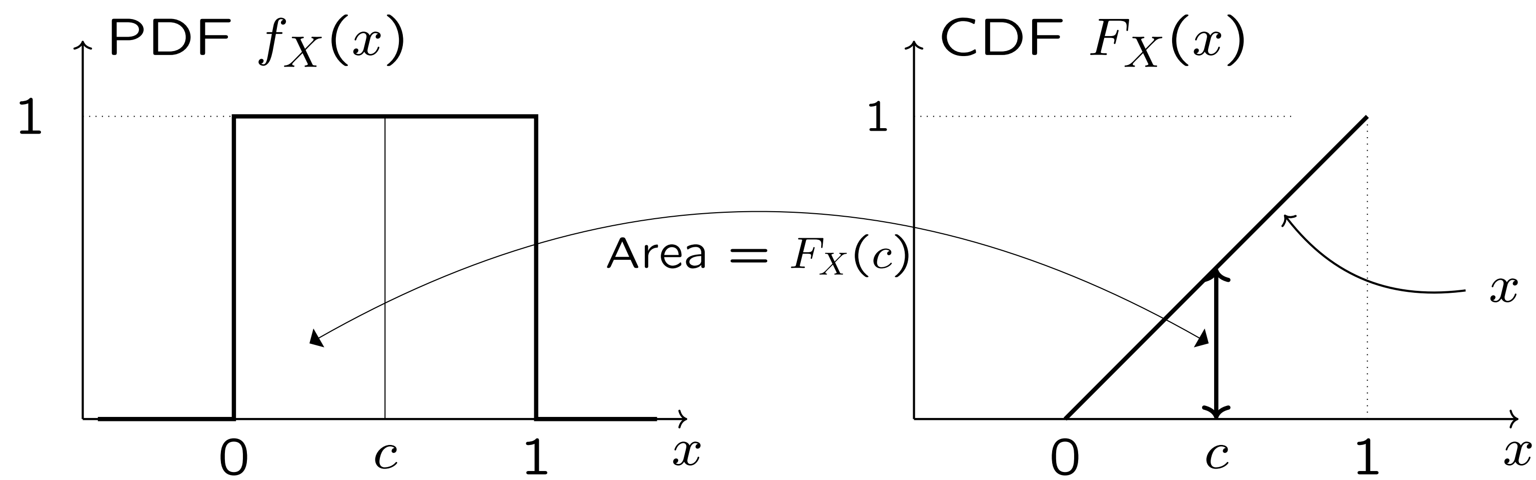
$F_X(x)$ can be expressed in terms of the PDF:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

The CDF can be used to evaluate the probability that $X$ falls in an interval:

$$P(a < X \leq b) = \int_a^b f_X(t)dt = F_X(b) - F_X(a)$$

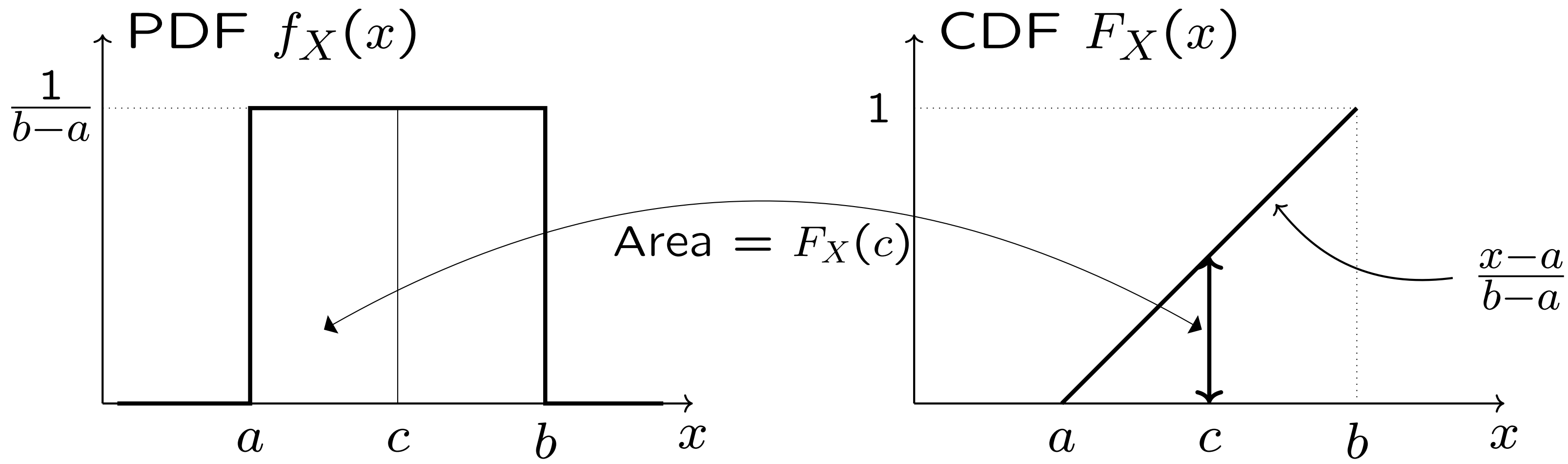
Example: uniform r.v. $X$ on interval $[0, 1]$




CDF of uniform r.v. on $[0, 1]$ is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

Example: uniform r.v. $X$ on general interval $[a,b]$




CDF of uniform r.v. on $[a,b]$ is

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$

CDF is related to the PDF by the formula

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)dt$$

and has no jumps, i.e. it is continuous

Thus PDF $f_X$ can be obtained from the CDF by differentiation:

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Acknowledgements: Probability & Statistics with R by
Michael Akritas (Pearson 2016)
MIT OCW 6.436J (Fall 2008)
University of Chicago STAT 24400 (Autumn 2016)