

LECTURE 1: Probability models and axioms

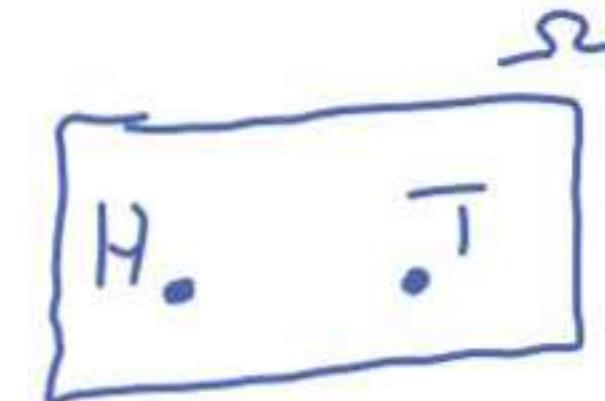
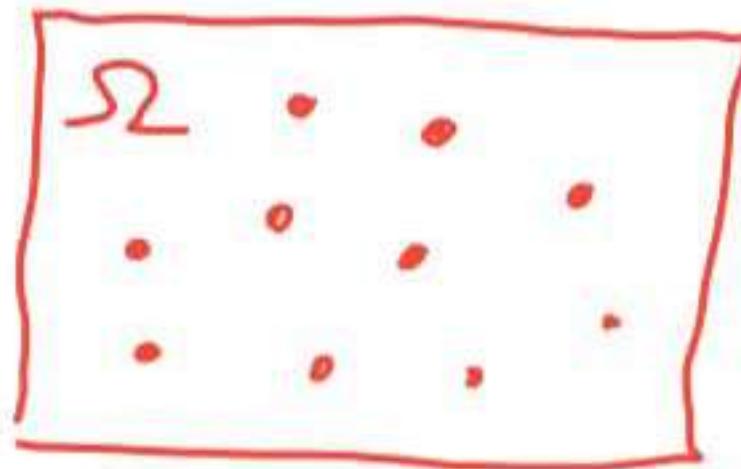
- Sample space
- Probability laws
 - Axioms
 - Properties that follow from the axioms
- Examples
 - Discrete
 - Continuous
- Discussion
 - Countable additivity
 - Mathematical subtleties
- Interpretations of probabilities

Sample space

- Two steps:
 - Describe possible outcomes
 - Describe beliefs about likelihood of outcomes

Sample space

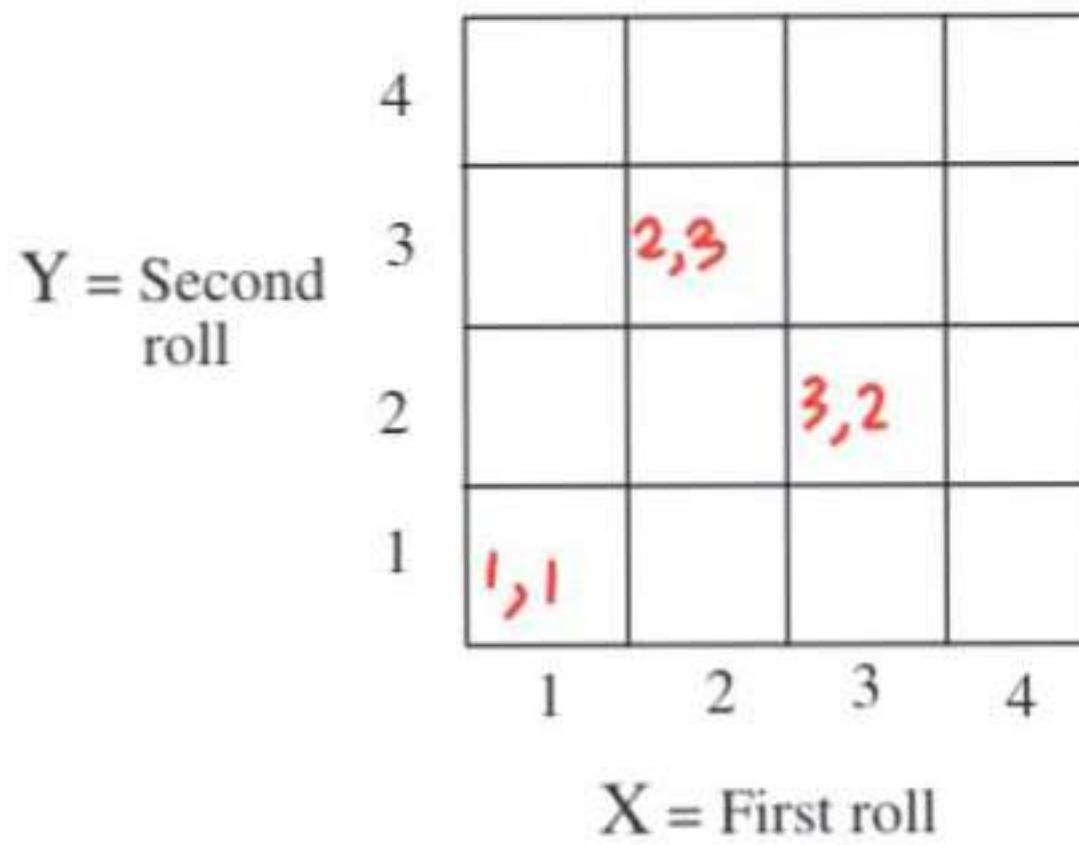
- List (set) of possible outcomes, Ω
- List must be:
 - Mutually exclusive
 - Collectively exhaustive
 - At the “right” granularity



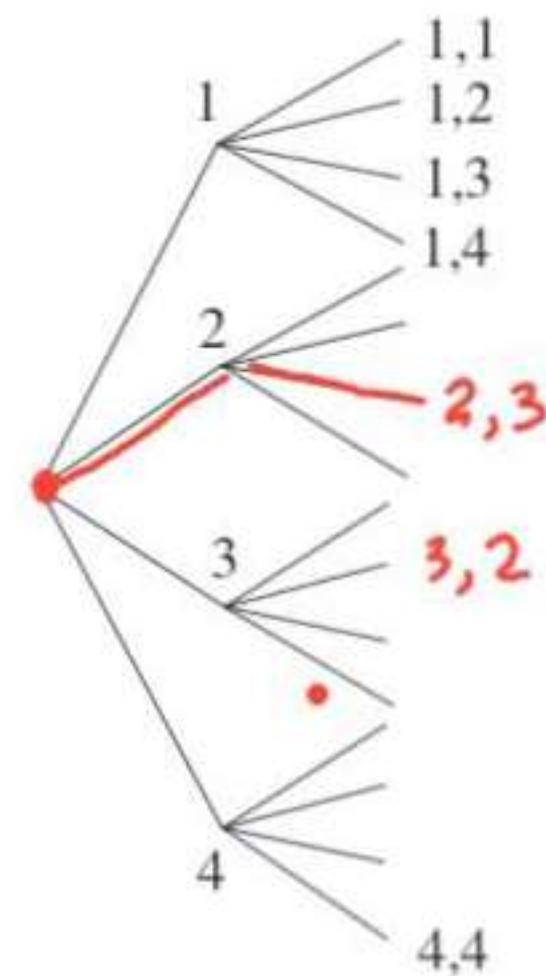
- H and rains Ω
- H and no rain
- T and rains
- T and no rain

Sample space: discrete/finite example

- Two rolls of a tetrahedral die



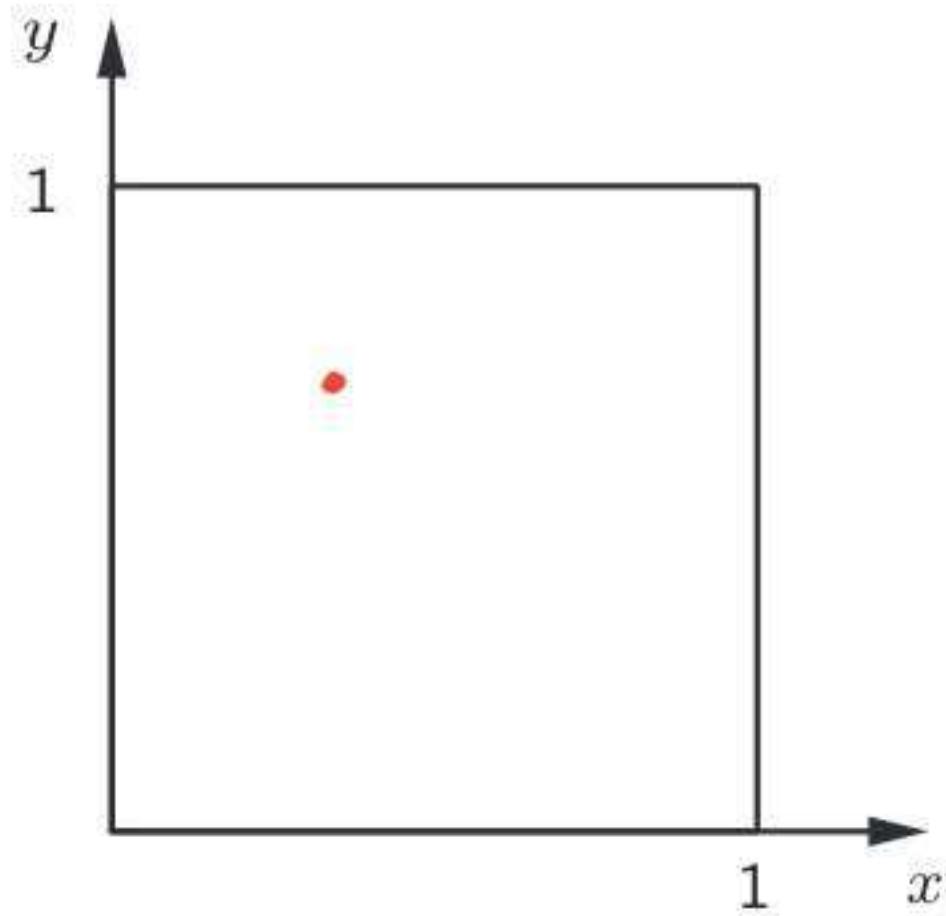
sequential description



Tree

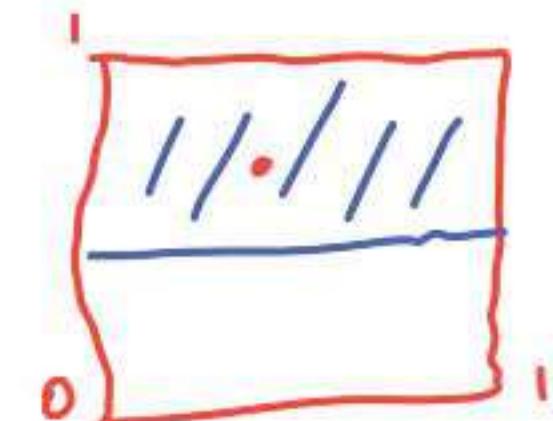
Sample space: continuous example

- (x, y) such that $0 \leq x, y \leq 1$

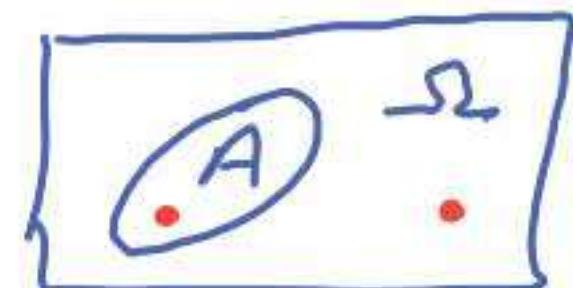


Probability axioms

- Event: a subset of the sample space
 - Probability is assigned to events



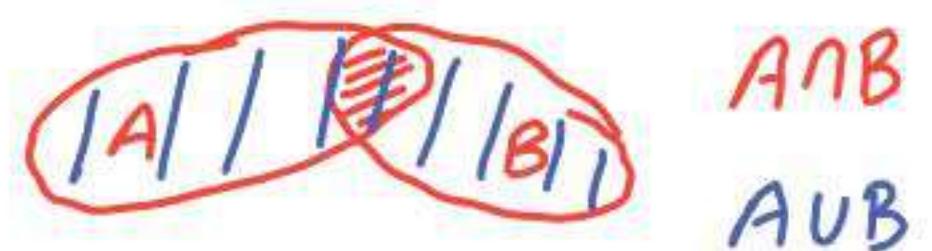
$\Omega(A)$



Axioms:

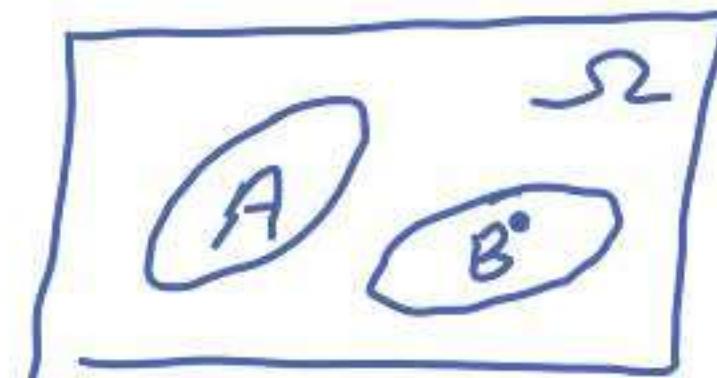
- Nonnegativity: $P(A) \geq 0$
- Normalization: $P(\Omega) = 1$
- (Finite) additivity: (to be strengthened later)
If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

empty set



$A \cap B$

$A \cup B$



Some simple consequences of the axioms

Axioms

$$P(A) \geq 0$$

$$P(\Omega) = 1$$

For disjoint events:

$$P(A \cup B) = P(A) + P(B)$$

Consequences

$$P(A) \leq 1$$

$$P(\emptyset) = 0$$

$$P(A) + P(A^c) = 1$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

and similarly for k disjoint events

$$P(\{s_1, s_2, \dots, s_k\}) = P(\{s_1\}) + \dots + P(\{s_k\})$$

$$= P(s_1) + \dots + P(s_k)$$

Some simple consequences of the axioms

Axioms

(a) $P(A) \geq 0$



$$A \cup A^c = \Omega$$

$$A \cap A^c = \emptyset$$

(b) $P(\Omega) = 1$

$$1 \stackrel{(b)}{=} P(\Omega) = P(A \cup A^c)$$

For disjoint events:

(c) $P(A \cup B) = P(A) + P(B)$

$$\stackrel{(c)}{=} P(A) + P(A^c) \stackrel{(a)}{\leq} 1$$

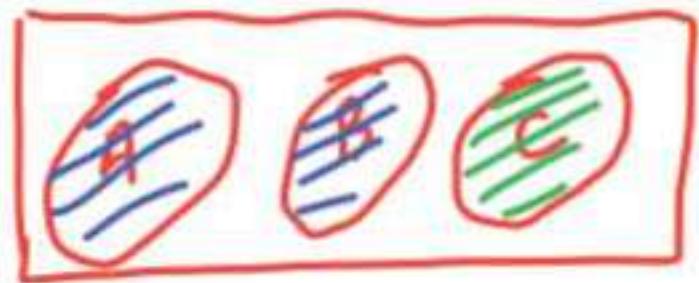
$$P(A) = 1 - \underbrace{P(A^c)}_{\leq 1} \leq 1$$

$$1 = P(\Omega) + P(\Omega^c)$$

$$1 = 1 + P(\emptyset) \Rightarrow P(\emptyset) = 0.$$

Some simple consequences of the axioms

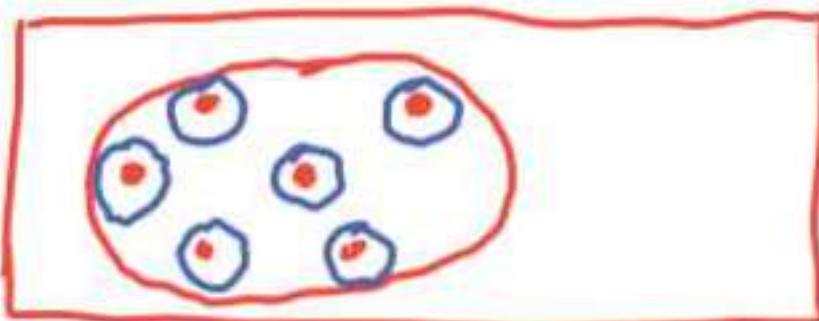
- A, B, C disjoint: $P(A \cup B \cup C) = P(A) + P(B) + P(C)$



$$\begin{aligned}P(A \cup B \cup C) &= P((A \cup B) \cup C) = P(A \cup B) + P(C) \\&= P(A) + P(B) + P(C)\end{aligned}$$

If A_1, \dots, A_k disjoint $\Rightarrow P(A_1 \cup \dots \cup A_k) = \sum_{i=1}^k P(A_i)$

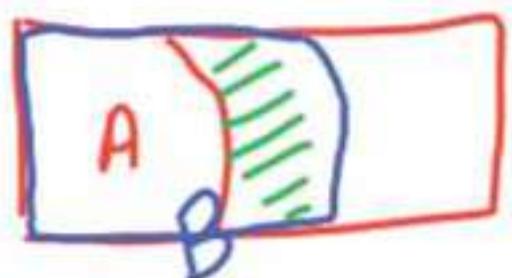
- $P(\{s_1, s_2, \dots, s_k\}) = P(\{s_1\} \cup \{s_2\} \cup \dots \cup \{s_k\})$



$$\begin{aligned}&= P(\{s_1\}) + \dots + P(\{s_k\}) \\&= P(s_1) + \dots + P(s_k)\end{aligned}$$

More consequences of the axioms

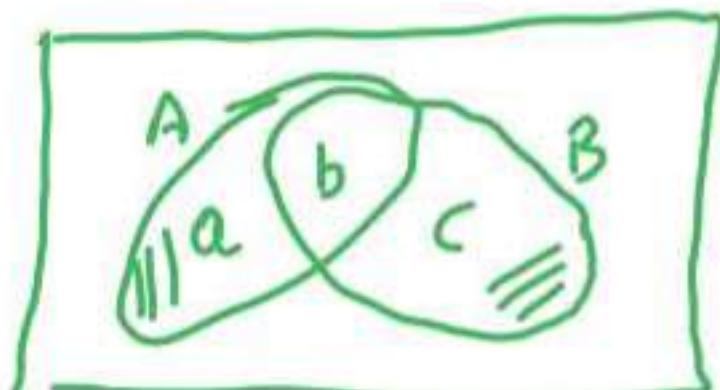
- If $A \subset B$, then $P(A) \leq P(B)$



$$B = A \cup (B \cap A^c)$$

$$P(B) = P(A) + \underline{P(B \cap A^c)} \geq P(A)$$

- $P(A \cup B) = P(A) + P(B) - \overbrace{P(A \cap B)}$



$$a = P(A \cap B^c) \quad b = P(A \cap B) \quad c = P(B \cap A^c)$$

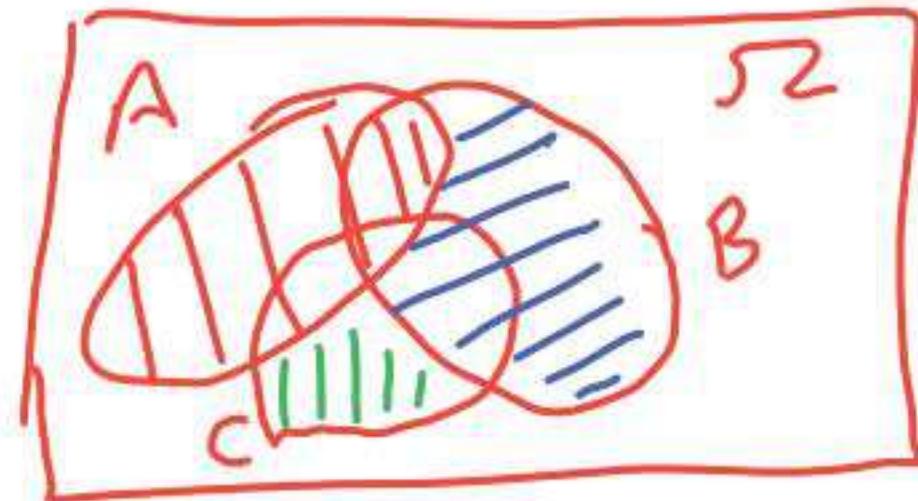
$$P(A \cup B) = a + b + c$$

$$P(A) + P(B) - P(A \cap B) = (a+b) + (b+c) - b$$

- $P(A \cup B) \leq P(A) + P(B)$ union bound = $a + b + c$

More consequences of the axioms

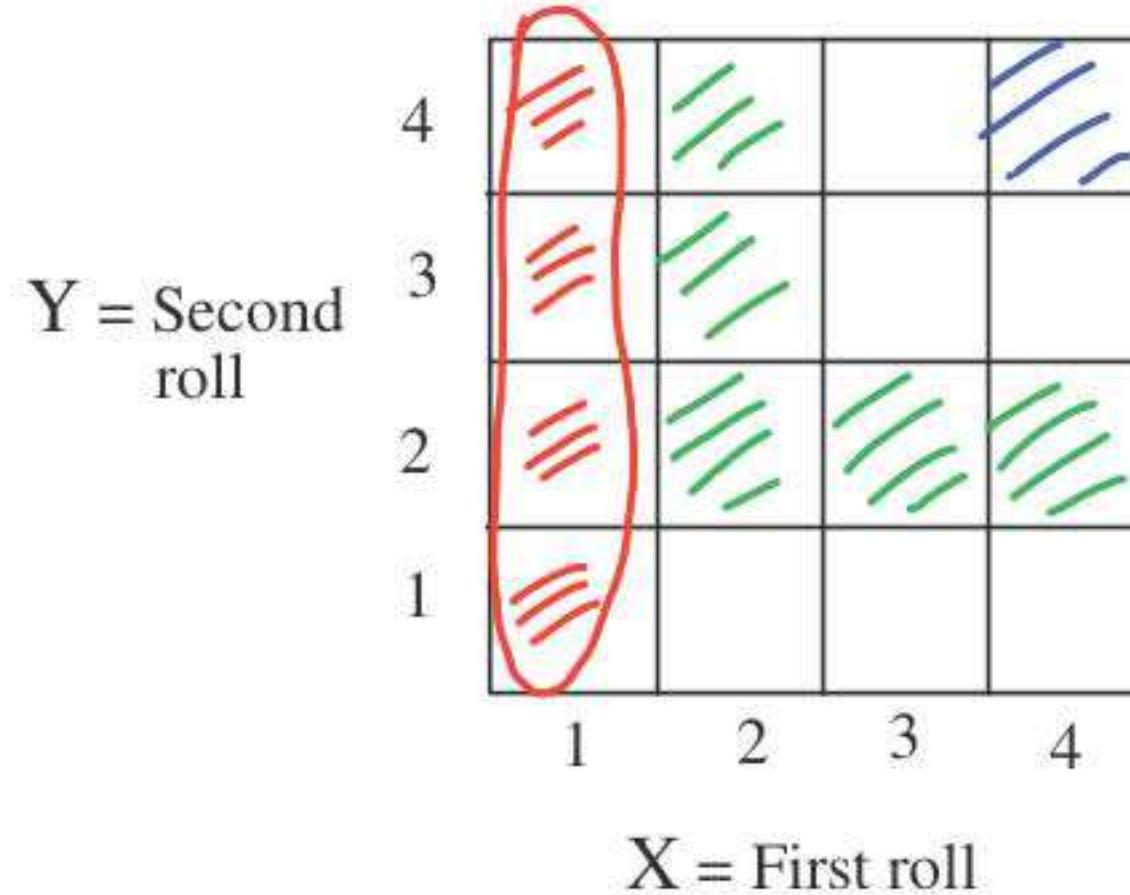
- $P(A \cup B \cup C) = P(A) + \underline{P(A^c \cap B)} + \underline{P(A^c \cap B^c \cap C)}$ • ↗



$$\begin{aligned} P(A \cup B \cup C) &= \\ &= A \cup \underline{(B \cap A^c)} \cup \underline{(C \cap A^c \cap B^c)} \end{aligned}$$

Probability calculation: discrete/finite example

- Two rolls of a tetrahedral die
- Let every possible outcome have probability $1/16$



- $P(X = 1) = 4 \cdot \frac{1}{16} = \frac{1}{4}$

Let $Z = \min(X, Y)$

~~$X = 2, Y = 3, Z = 2$~~

- $\underline{P(Z = 4)} = 1/16$

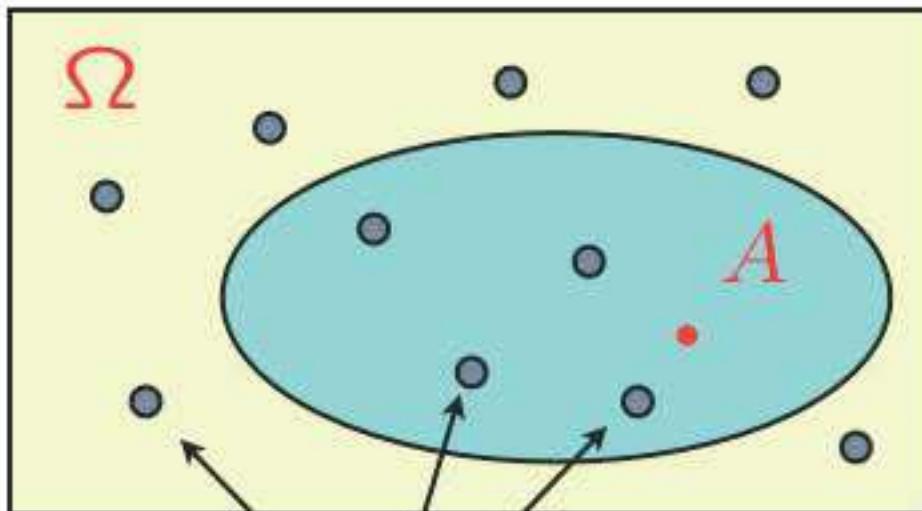
- $P(Z = 2) = 5 \cdot \frac{1}{16} .$

Discrete uniform law

✓ finite

- Assume Ω consists of n equally likely elements
- Assume A consists of k elements

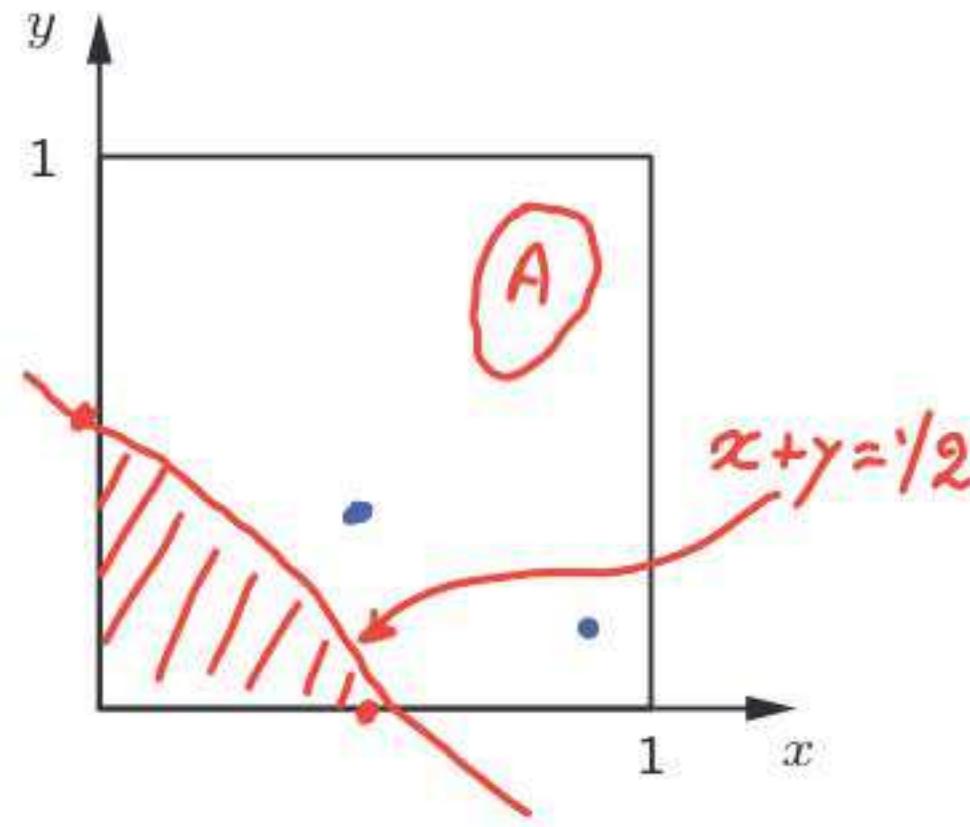
$$P(A) = k \cdot \frac{1}{n}$$



$$\text{prob} = \frac{1}{n}$$

Probability calculation: continuous example

- (x, y) such that $0 \leq x, y \leq 1$
- Uniform probability law: Probability = Area



$$P\left(\{(x, y) \mid x + y \leq 1/2\}\right) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

$$P\left(\{(0.5, 0.3)\}\right) = 0$$

Probability calculation steps

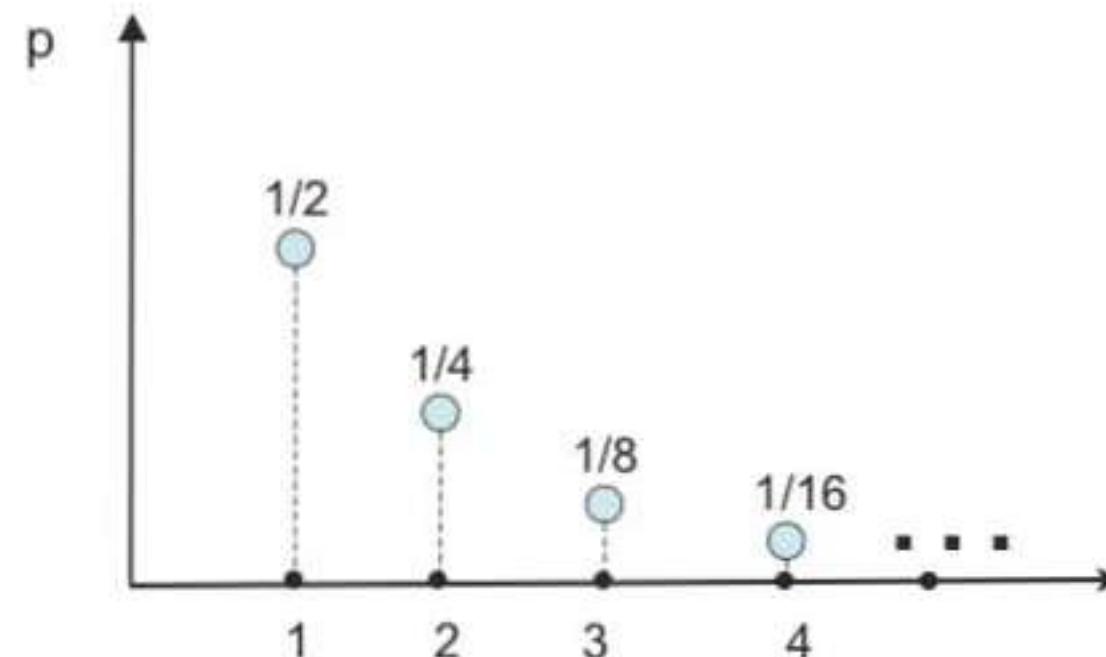
- Specify the sample space
- Specify a probability law
- Identify an event of interest
- Calculate...

Probability calculation: discrete but infinite sample space

- Sample space: $\{1, 2, \dots\}$

- We are given $P(n) = \frac{1}{2^n}, n = 1, 2, \dots$

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} \sum_{n=0}^{\infty} \frac{1}{2^n} = \frac{1}{2} \cdot \frac{1}{1 - (\frac{1}{2})} = 1$$



- $P(\text{outcome is even}) = P(\{2, 4, 6, \dots\})$

$$= P(\{2\} \cup \{4\} \cup \{6\} \cup \dots) \quad \textcircled{=} \quad P(2) + P(4) + P(6) + \dots$$

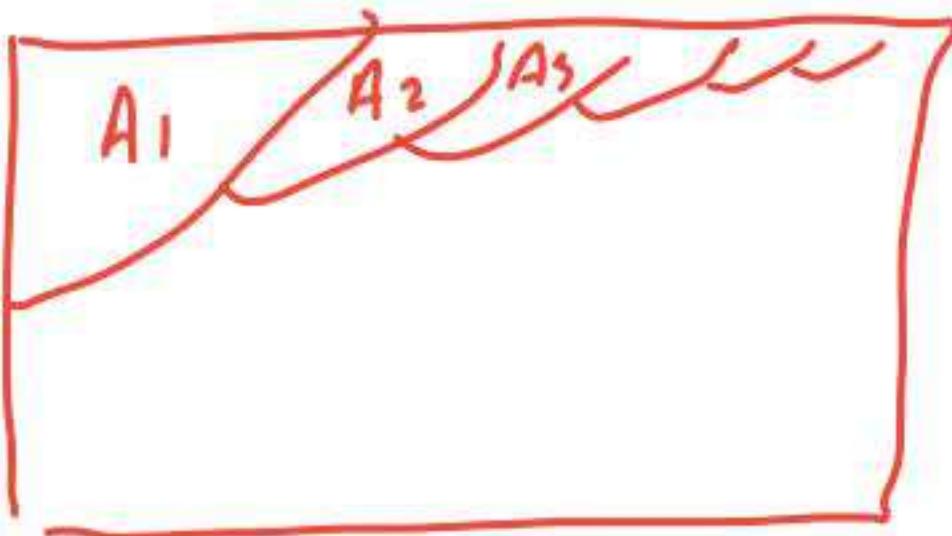
$$= \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6} + \dots = \frac{1}{4} \left(1 + \frac{1}{4} + \frac{1}{4^2} + \dots \right) = \frac{1}{4} \cdot \frac{1}{1 - \frac{1}{4}} = \frac{1}{3}$$

Countable additivity axiom

- Strengthens the finite additivity axiom

Countable Additivity Axiom:

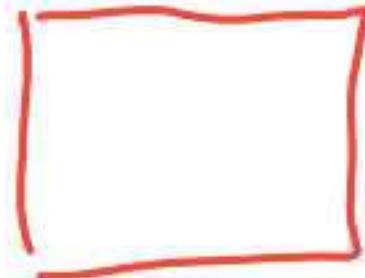
If A_1, A_2, A_3, \dots is an infinite sequence of disjoint events,
then $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$



Mathematical subtleties

Countable Additivity Axiom:

If A_1, A_2, A_3, \dots is an infinite sequence of disjoint events,
then $P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$


$$1 = P(\Omega) = P\left(\bigcup \{(x,y)\}\right) \stackrel{?}{=} \sum P(\{(x,y)\}) = \sum 0 = 0$$

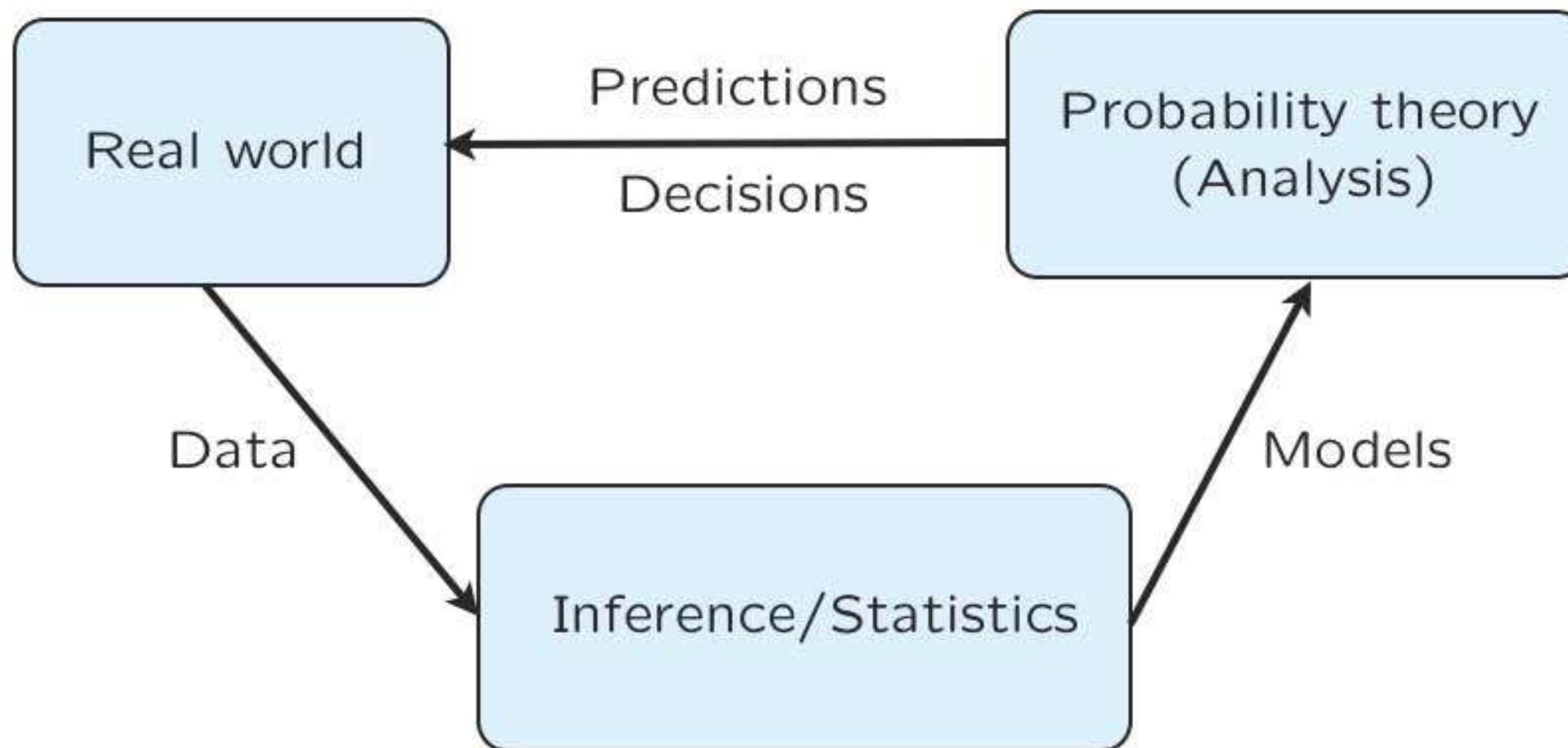
- Additivity holds only for “countable” sequences of events
- The unit square (similarly, the real line, etc.) is **not countable** (its elements cannot be arranged in a sequence)
- “Area” is a legitimate probability law on the unit square, as long as we do not try to assign probabilities/areas to “very strange” sets

Interpretations of probability theory

- A narrow view: a branch of math
 - Axioms \Rightarrow theorems “**Thm:**” “Frequency” of event A “is” $P(A)$
- Are probabilities frequencies?
 - $P(\text{coin toss yields heads}) = 1/2$
 - $P(\text{the president of ... will be reelected}) = 0.7$
- Probabilities are often interpreted as:
 - Description of beliefs
 - Betting preferences

The role of probability theory

- A framework for analyzing phenomena with uncertain outcomes
 - Rules for consistent reasoning
 - Used for predictions and decisions



MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

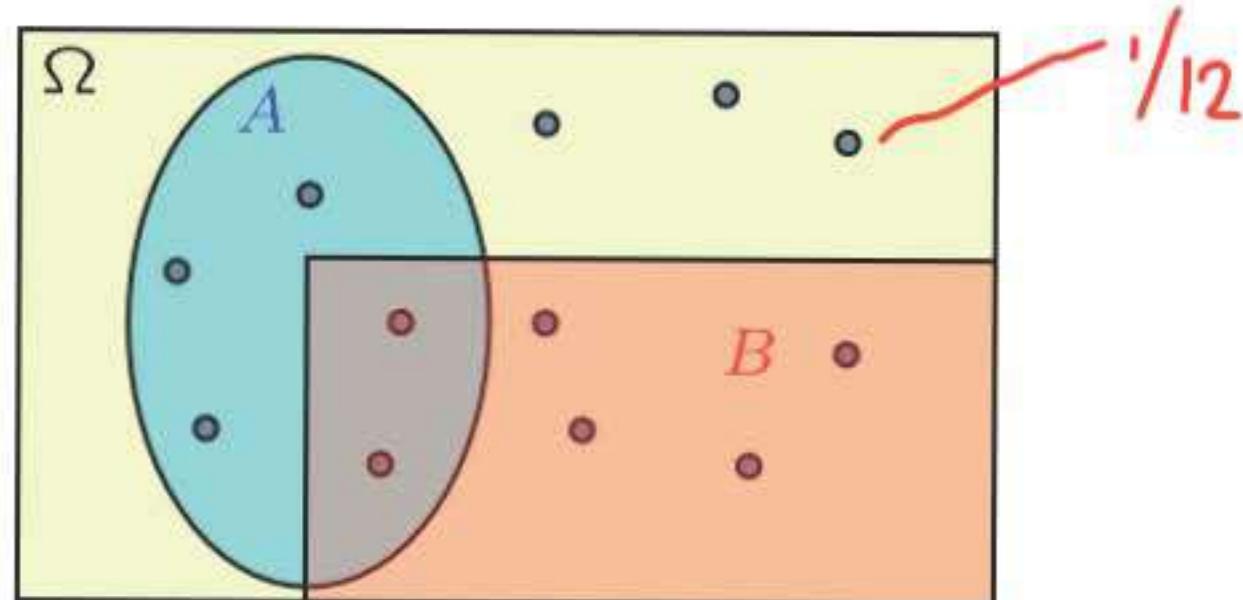
For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 2: Conditioning and Bayes' rule

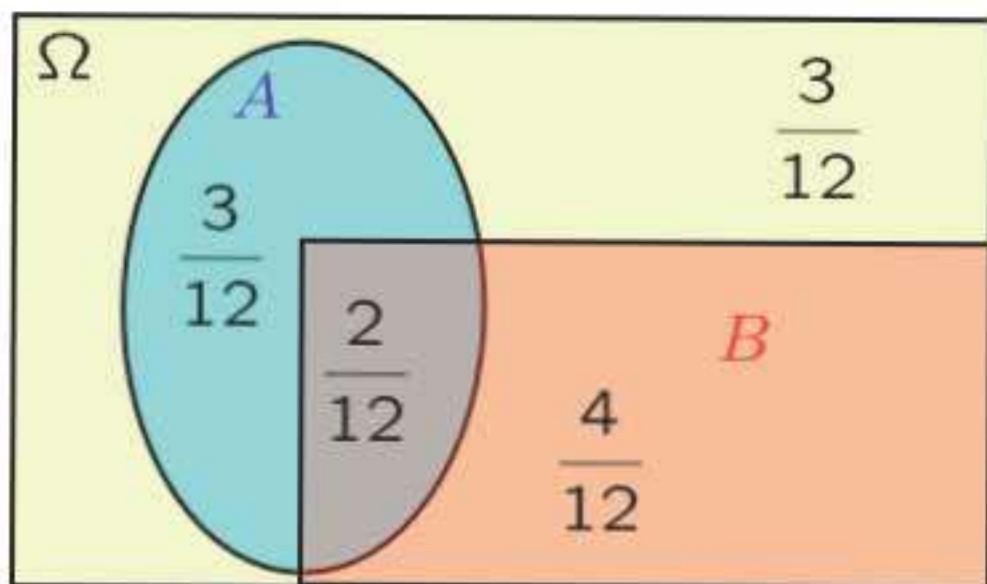
- Conditional probability
- Three **important** tools:
 - Multiplication rule
 - Total probability theorem
 - Bayes' rule (\rightarrow inference)

The idea of conditioning

Assume 12 equally likely outcomes

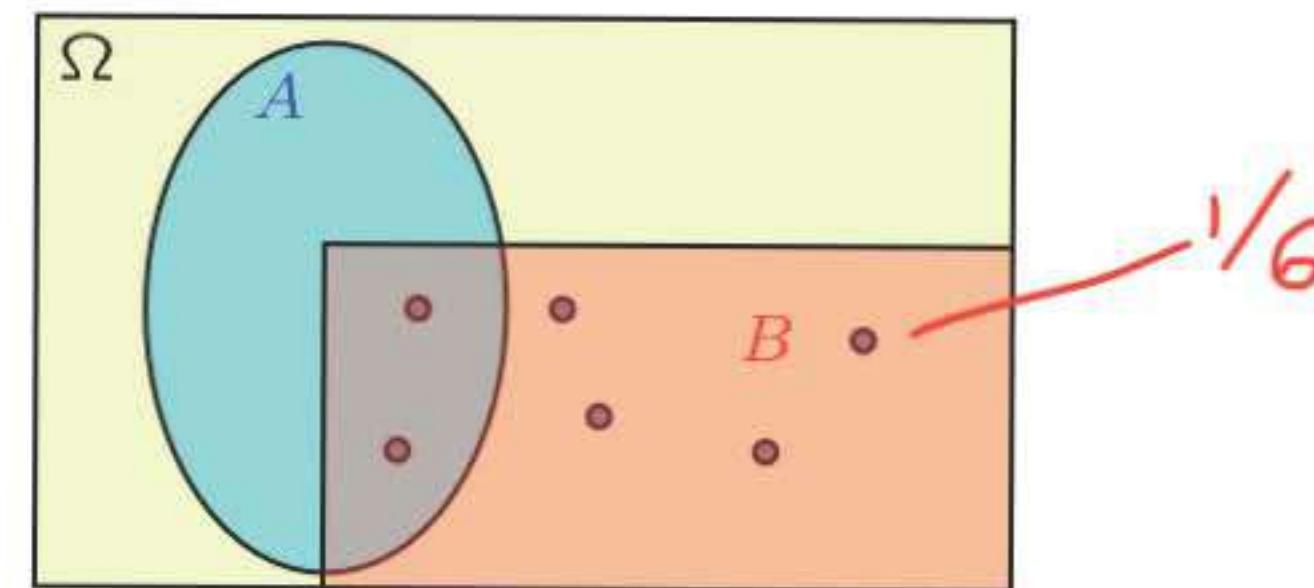


$$P(A) = \frac{5}{12} \quad P(B) = \frac{6}{12}$$

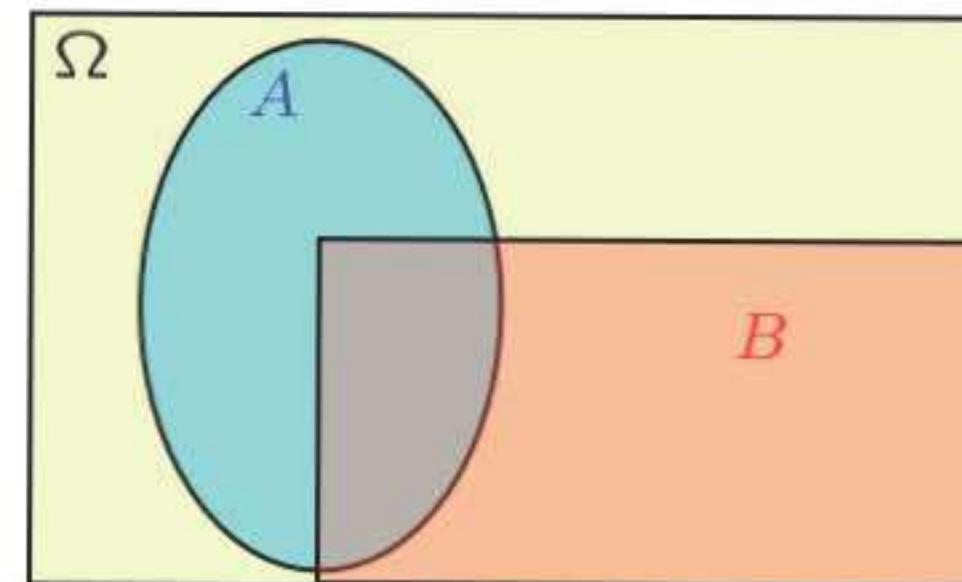


Use new information to revise a model

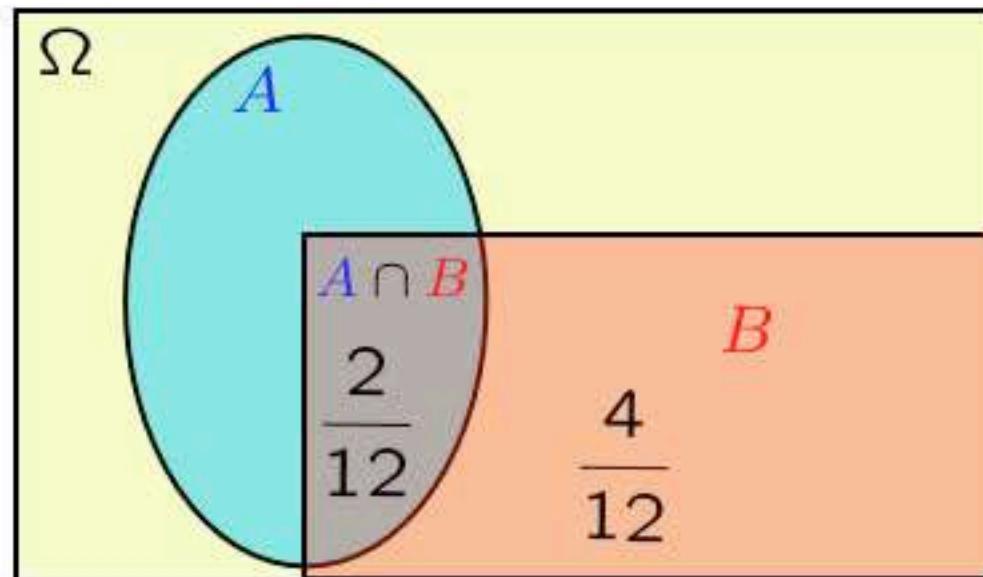
If told B occurred:



$$P(A | B) = \frac{2}{6} = \frac{1}{3} \quad P(B | B) = \underline{\underline{1}}$$



Definition of conditional probability



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

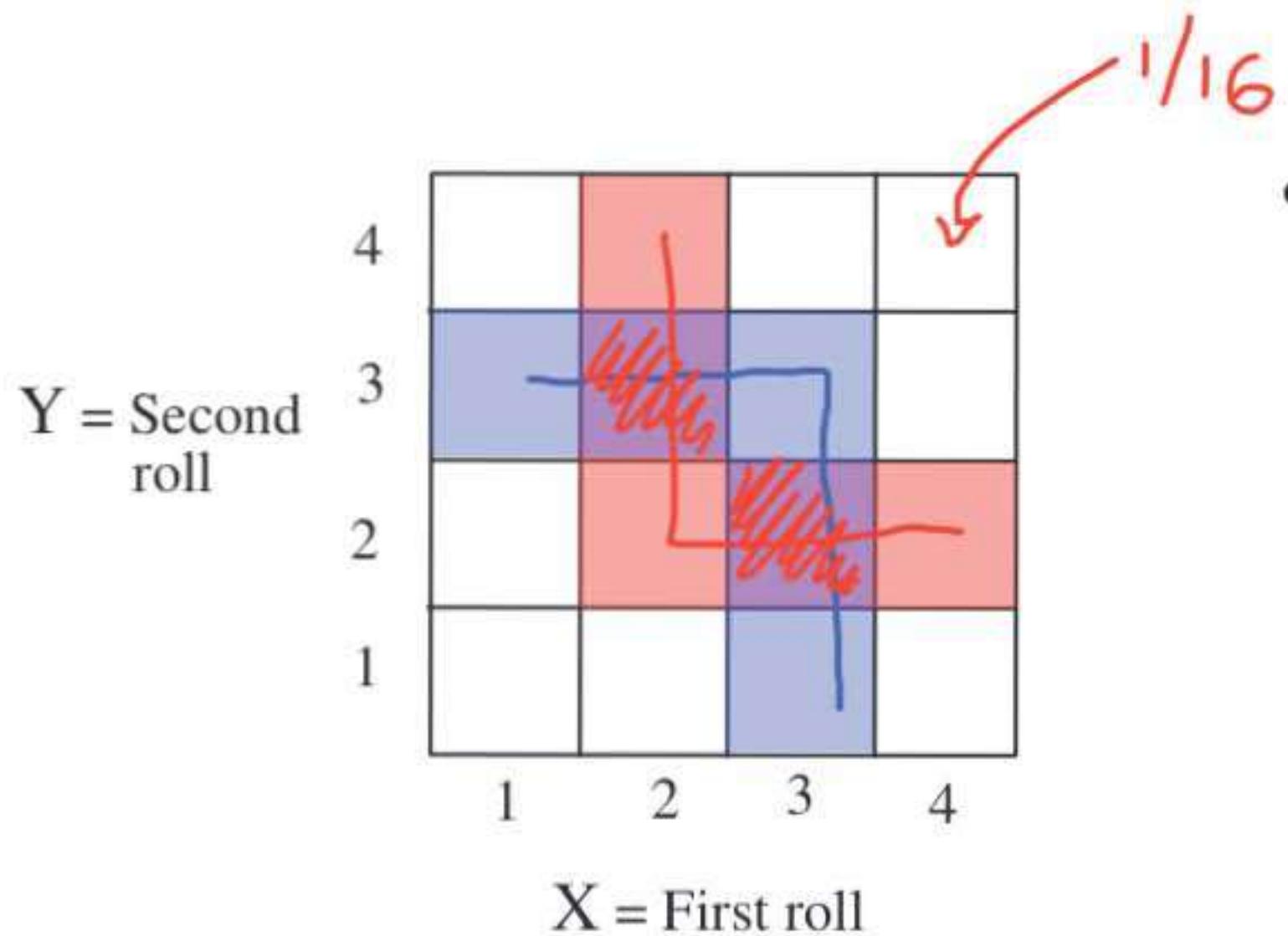
Def.

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

defined only when $P(B) > 0$

$$= \frac{2/12}{6/12} = \frac{1}{3}$$

Example: two rolls of a 4-sided die



- Let B be the event: $\min(X, Y) = 2$

Let $M = \max(X, Y)$

$$P(M = 1 | B) = 0$$

$$P(\underbrace{M = 3}_{\text{}} | B) = \frac{P(M=3 \text{ and } B)}{P(B)}$$

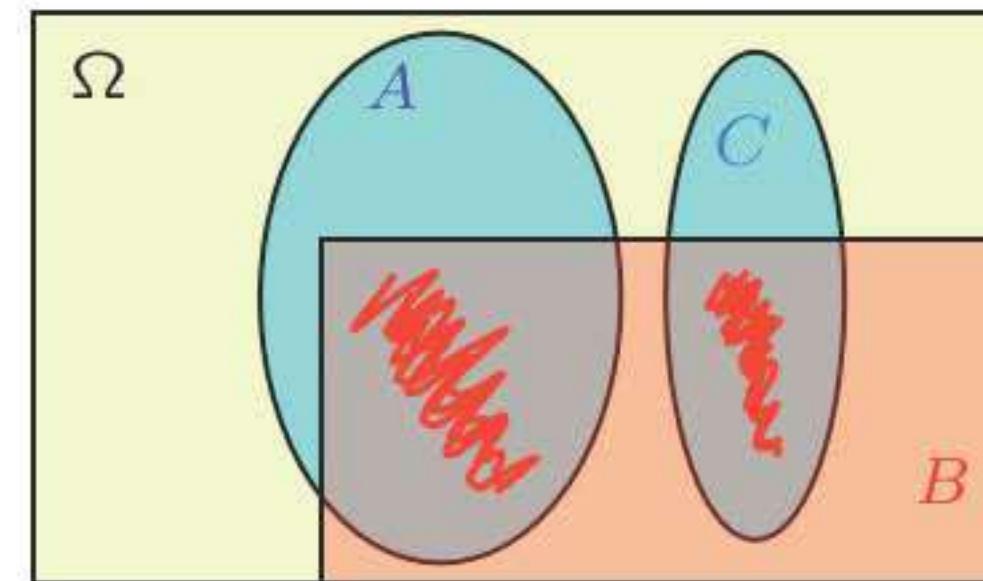
$$= \frac{2/16}{5/16} = \frac{2}{5}$$

Conditional probabilities share properties of ordinary probabilities

$$P(A | B) \geq 0 \quad \text{assuming } P(B) > 0$$

$$P(\Omega | B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

$$P(B | B) = \frac{P(B \cap B)}{P(B)} = 1$$



If $A \cap C = \emptyset$, then $P(A \cup C | B) = P(A | B) + P(C | B)$

$$= \frac{P((A \cup C) \cap B)}{P(B)} = \frac{P((A \cap B) \cup (C \cap B))}{P(B)} = \frac{P(A \cap B) + P(C \cap B)}{P(B)} =$$

$= P(A|B) + P(C|B)$ also finite
countable additivity

Models based on conditional probabilities

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

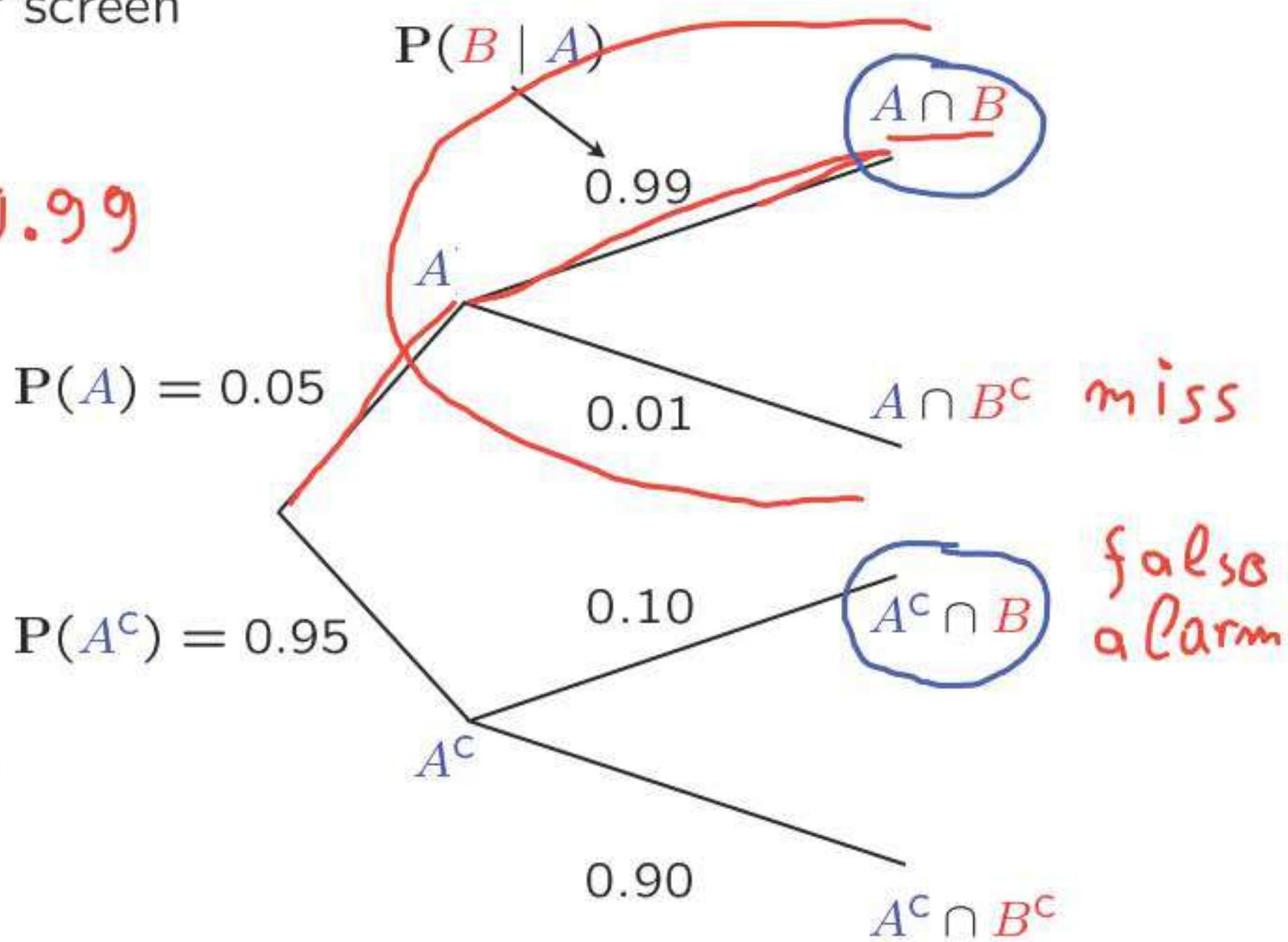
Event A : Airplane is flying above

Event B : Something registers on radar screen

- $P(A \cap B) = P(A) \cdot P(B | A) = 0.05 \cdot 0.99$

- $P(B) = 0.05 \cdot 0.99 + 0.95 \cdot 0.1 = 0.1445$

- $P(A | B) = \frac{0.05 \cdot 0.99}{0.1445} = 0.34$



The multiplication rule

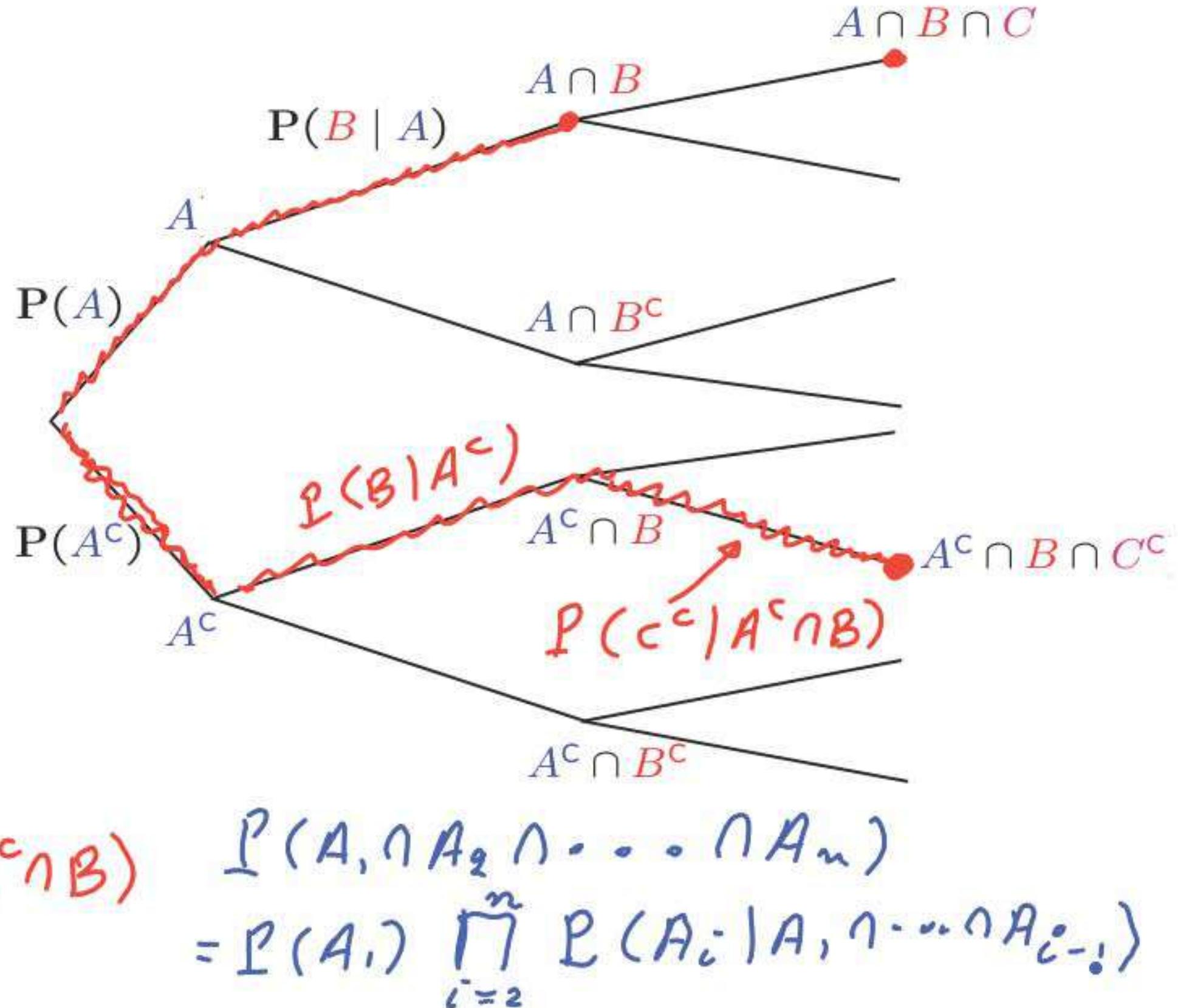
$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

$$\begin{aligned} P(A \cap B) &= P(B) P(A | B) \\ &= P(A) P(B | A) \end{aligned}$$

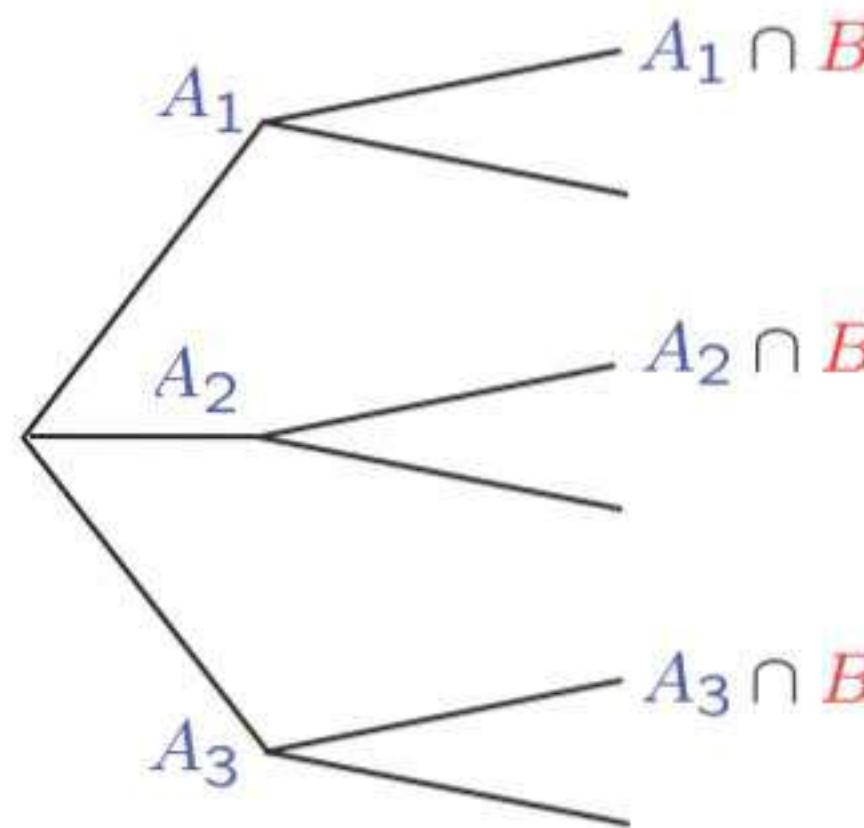
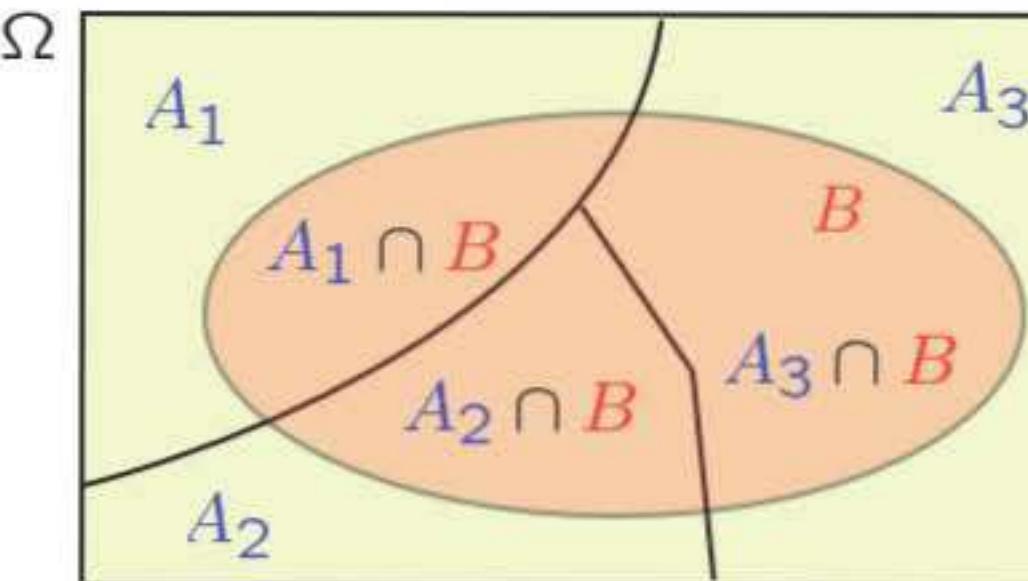
$$P(\underline{A^c \cap B} \cap \underline{C^c}) =$$

$$= P(A^c \cap B) P(C^c | A^c \cap B)$$

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



Total probability theorem



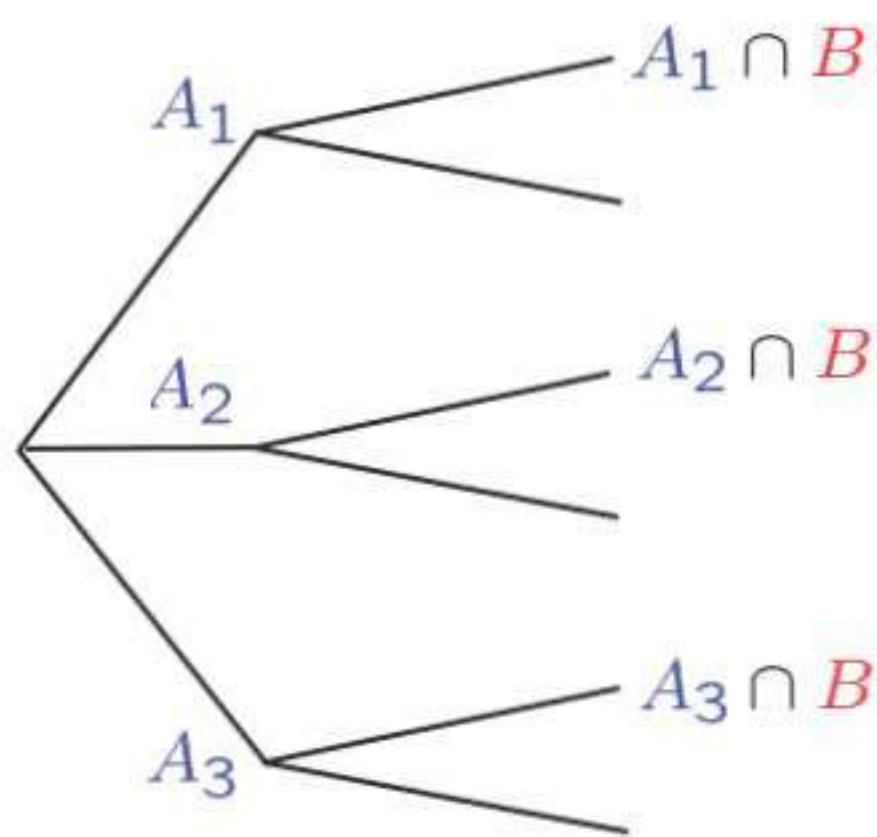
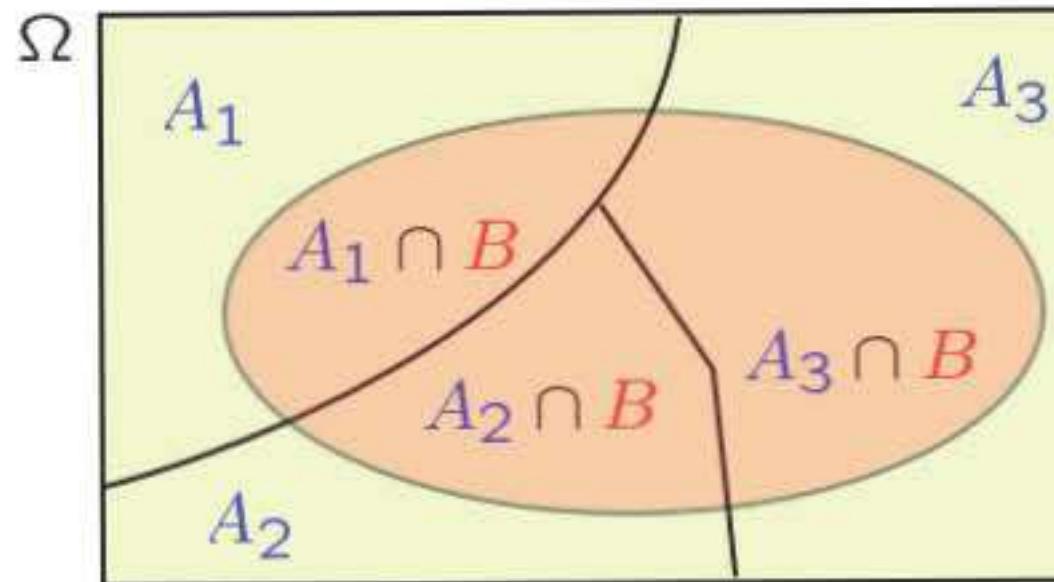
- Partition of sample space into A_1, A_2, A_3, \dots
- Have $P(A_i)$, for every i
- Have $P(B | A_i)$, for every i

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) \\ &= P(A_1)P(B | A_1) + \dots + \dots \end{aligned}$$

$$P(B) = \sum_i P(A_i) P(B | A_i)$$

weights
*weighted average
of $P(B | A_i)$*

Bayes' rule



- Partition of sample space into A_1, A_2, A_3
 - Have $P(A_i)$, for every i initial "beliefs"
 - Have $P(B | A_i)$, for every i
- revised "beliefs," given that B occurred:

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

$$P(A_i | B) = \frac{P(A_i)P(B | A_i)}{\sum_j P(A_j)P(B | A_j)}$$

Bayes' rule and inference

- Thomas Bayes, presbyterian minister (c. 1701-1761)
- “Bayes’ theorem,” published posthumously
- systematic approach for incorporating new evidence
- Bayesian inference
 - initial beliefs $P(A_i)$ on possible causes of an observed event B
 - model of the world under each A_i : $P(B | A_i)$

$$A_i \xrightarrow{\text{model}} B$$
$$P(B | A_i)$$

- draw conclusions about causes

$$B \xrightarrow{\text{inference}} A_i$$
$$P(A_i | B)$$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

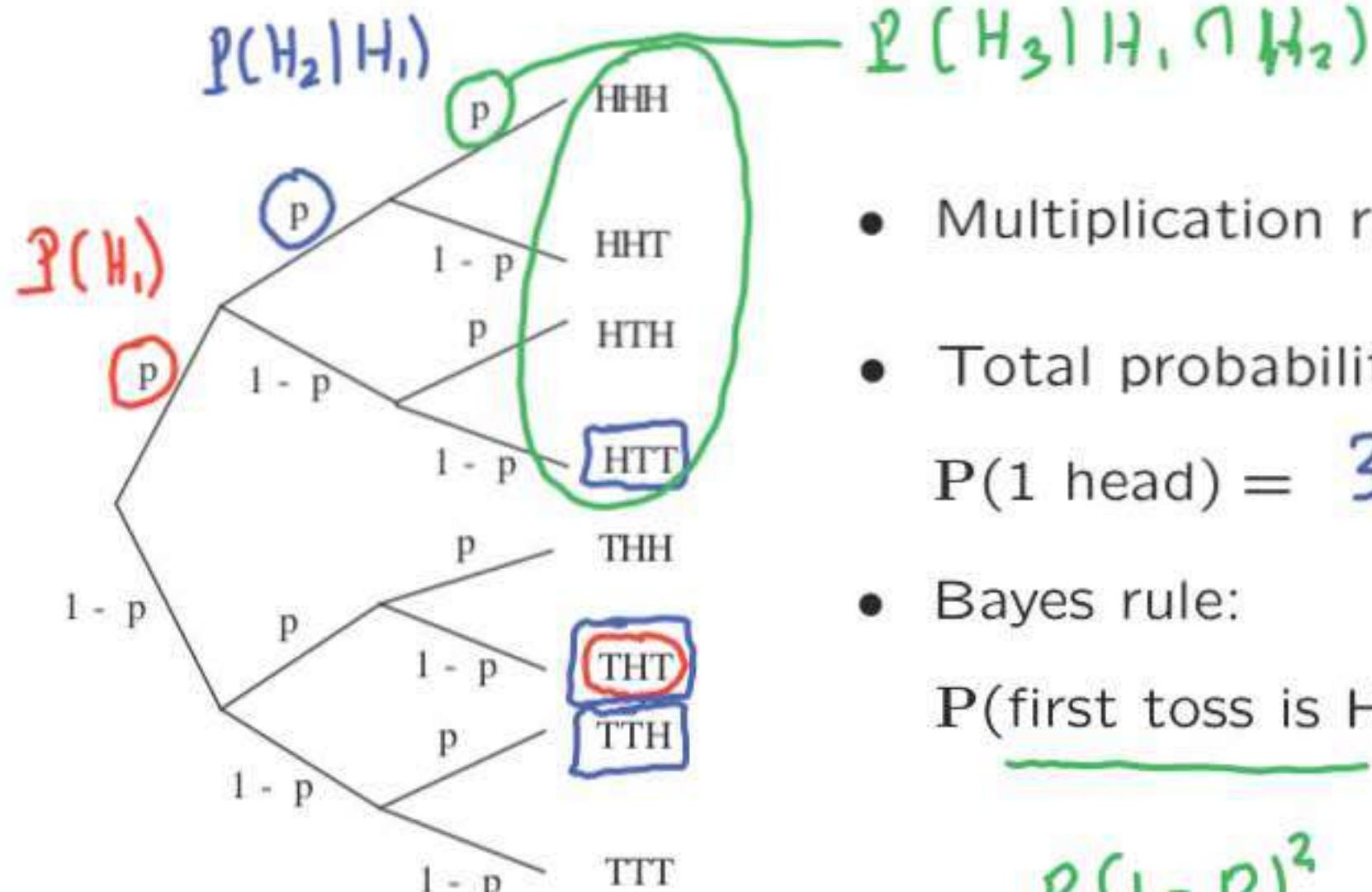
For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 3: Independence

- Independence of two events
- Conditional independence
- Independence of a collection of events
- Pairwise independence
- Reliability
- The king's sibling puzzle

A model based on conditional probabilities

- 3 tosses of a biased coin: $P(H) = p$, $P(T) = 1 - p$



$$\begin{aligned} P(H_2 | H_1) &= p = P(H_2 | T_1) \\ P(H_2) &= P(H_1) P(H_2 | H_1) \\ &\quad + P(T_1) P(H_2 | T_1) \\ &= p \end{aligned}$$

- Multiplication rule: $P(\underline{THT}) = \underline{(1-p)p(1-p)}$

- Total probability:

$$P(1 \text{ head}) = 3 p(1-p)^2$$

- Bayes rule:

$$\begin{aligned} P(\text{first toss is } H | 1 \text{ head}) &= \frac{P(H_1 \cap 1 \text{ head})}{P(1 \text{ head})} \\ &= \frac{p(1-p)^2}{3 p(1-p)^2} = \frac{1}{3} \end{aligned}$$

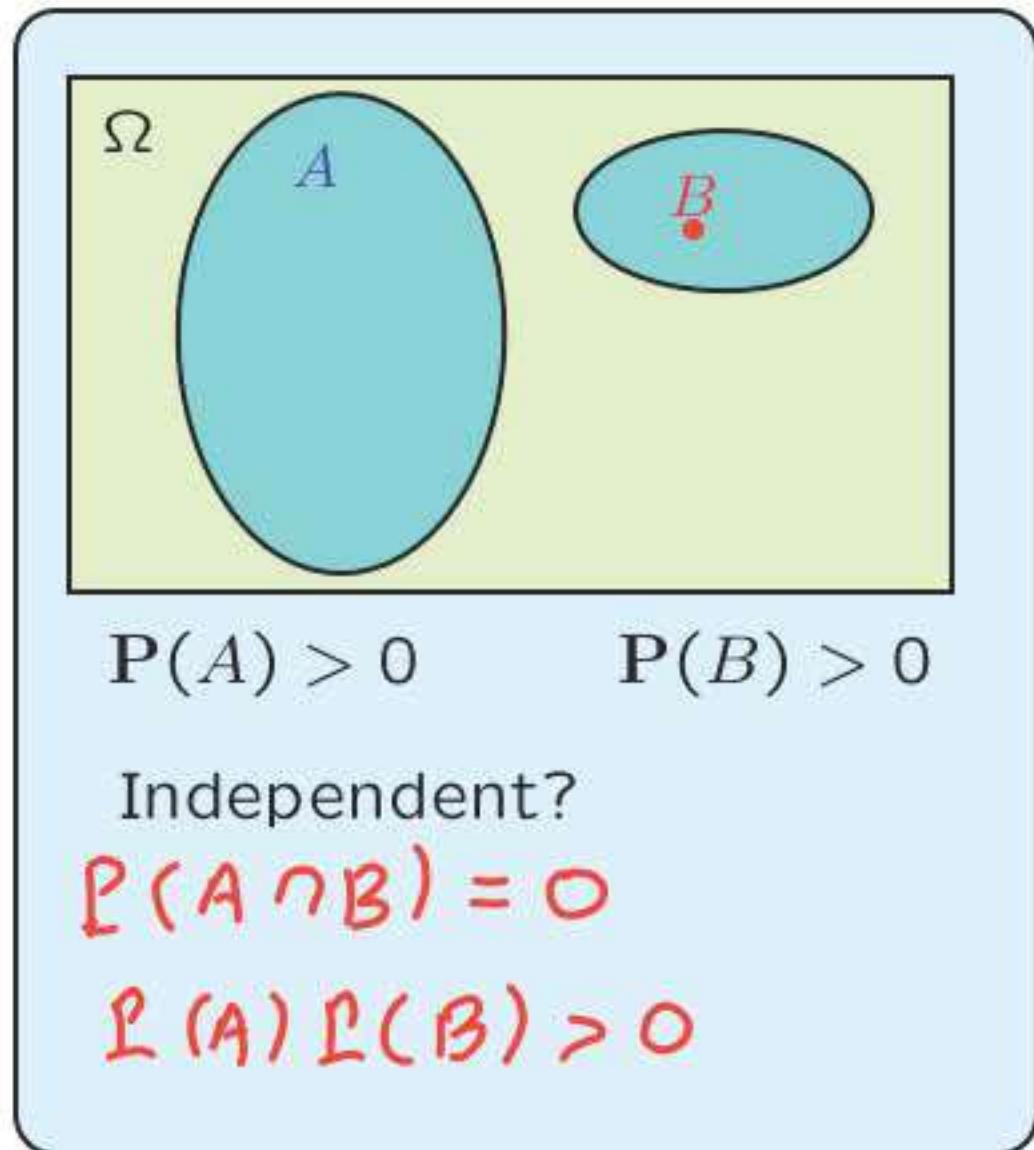
Independence of two events

- Intuitive “definition”: $P(B | A) = P(B)$
 - occurrence of A provides no new information about B

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

Definition of independence: $P(A \cap B) = P(A) \cdot P(B)$

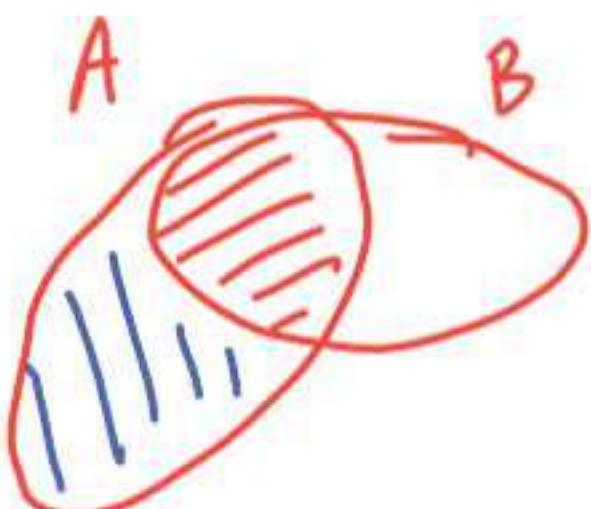
- Symmetric with respect to A and B
- implies $P(A | B) = P(A)$
- applies even if $P(A) = 0$



Independence of event complements

Definition of independence: $P(A \cap B) = P(A) \cdot P(B)$

- If A and B are independent, then A and B^c are independent.
 - Intuitive argument
 - Formal proof



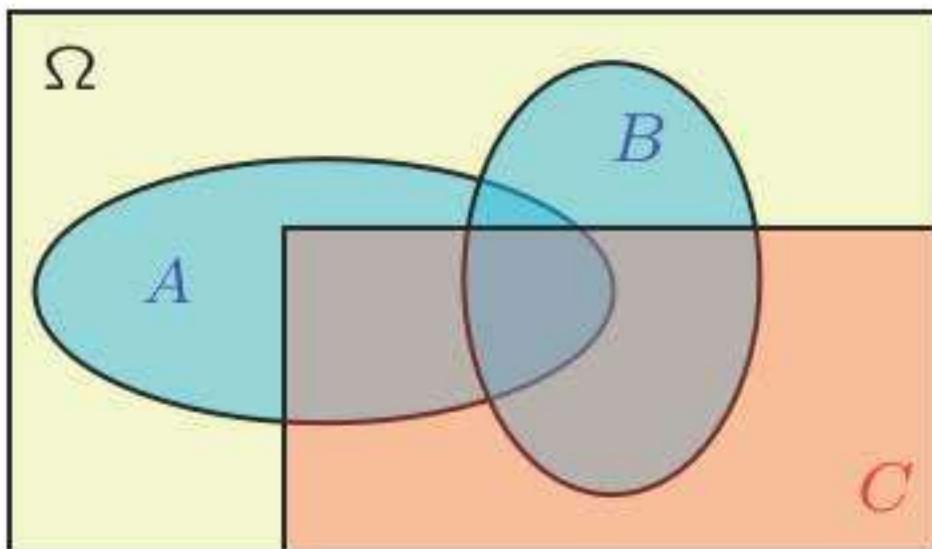
$$A = (A \cap B) \cup (A \cap B^c)$$

$$\begin{aligned}P(A) &= P(A \cap B) + P(A \cap B^c) \\&= P(A)P(B) + P(A \cap B^c)\end{aligned}$$

$$\begin{aligned}P(A \cap B^c) &= P(A) - P(A)P(B) = P(A)(1 - P(B)) \\&= P(A)P(B^c)\end{aligned}$$

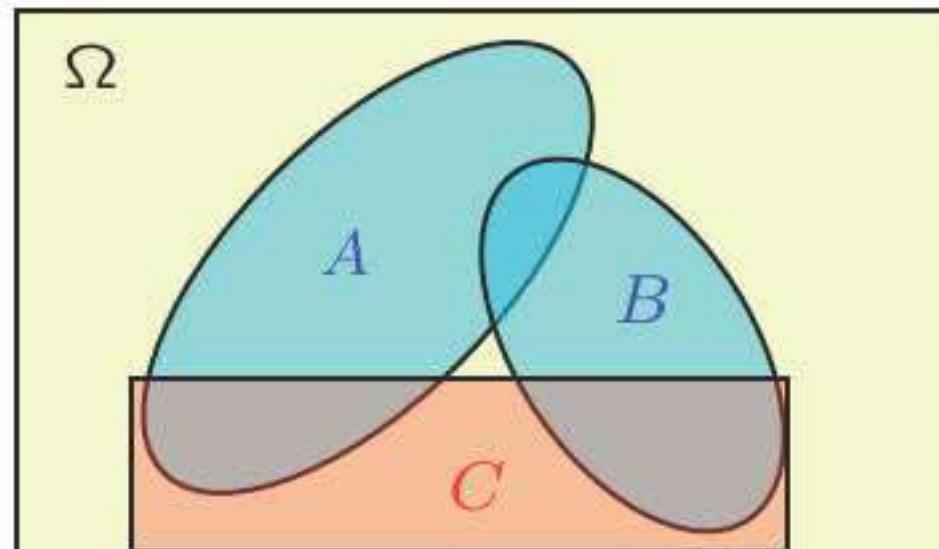
Conditional independence

- Conditional independence, given C , is defined as independence under the probability law $\mathbf{P}(\cdot | C)$



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

Assume A and B are independent



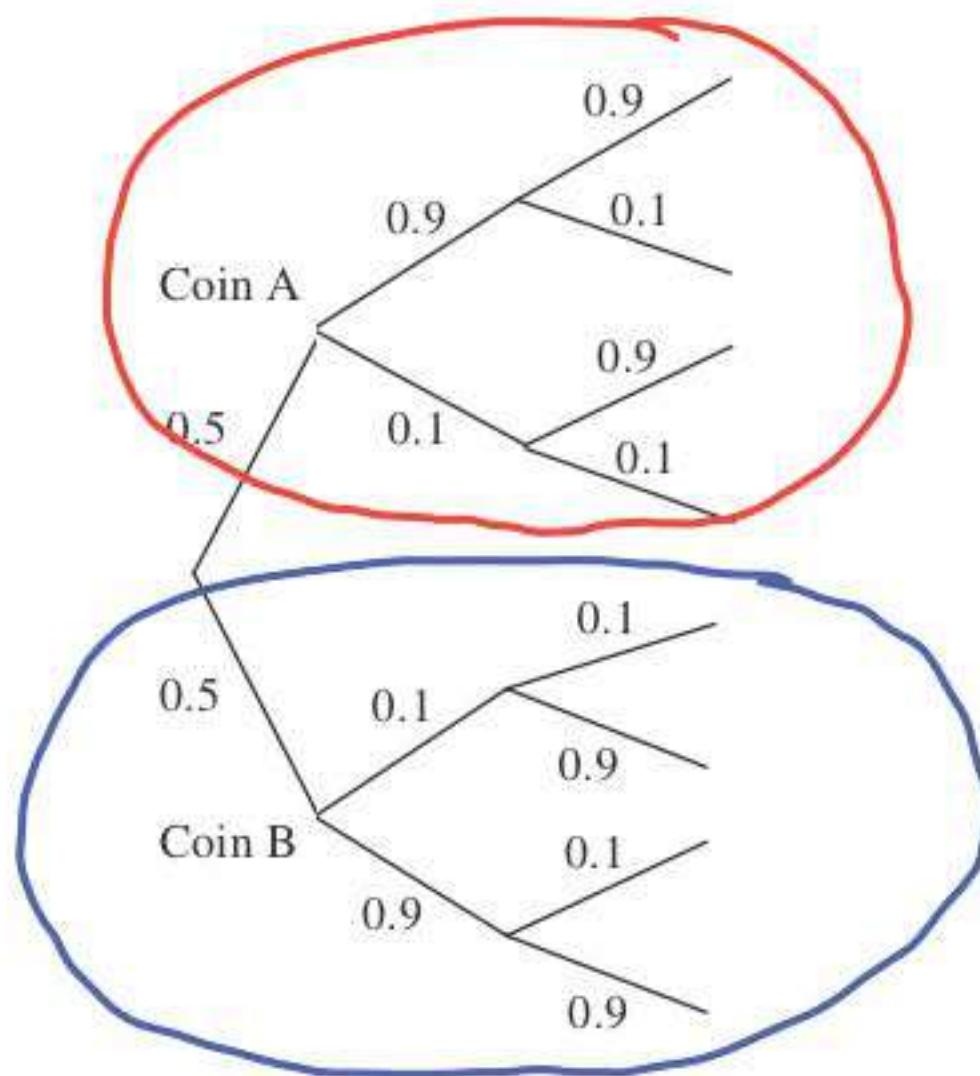
- If we are told that C occurred, are A and B independent? **No**

Conditioning may affect independence

- Two unfair coins, A and B :

$$P(H \mid \text{coin } A) = 0.9, P(H \mid \text{coin } B) = 0.1$$

- choose either coin with equal probability



- Compare:

$$\begin{aligned} P(\text{toss } 11 = H) &= P(A)P(H_{11} | A) + P(B)P(H_{11} | B) \\ &= 0.5 \times 0.9 + 0.5 \times 0.1 = 0.5 \end{aligned}$$

$$P(\text{toss } 11 = H \mid \text{first 10 tosses are heads})$$

$$\approx P(H_{11} | A) = 0.9$$

given a coin:
independent tosses

- Are coin tosses independent?

No!

Independence of a collection of events

- **Intuitive “definition”:** Information on some of the events does not change probabilities related to the remaining events

$$A_1, A_2, \dots, \text{indep} \Rightarrow P(A_3 \cap A_4^c) = P(A_3 \cap A_4^c | A_1 \cup (A_2 \cap A_5^c))$$

$$P(A_3) = P(A_3 | A_1 \cap A_2) = P(A_3 | A_1 \cap A_2^c) = P(A_3 | A_1^c \cap A_2)$$

Definition: Events A_1, A_2, \dots, A_n are called **independent** if:

$$P(A_i \cap A_j \cap \dots \cap A_m) = P(A_i)P(A_j) \cdots P(A_m) \quad \text{for any distinct indices } i, j, \dots, m$$

$n = 3$:

$$\left. \begin{array}{l} P(A_1 \cap A_2) = P(A_1) \cdot P(A_2) \\ P(A_1 \cap A_3) = P(A_1) \cdot P(A_3) \\ P(A_2 \cap A_3) = P(A_2) \cdot P(A_3) \end{array} \right\} \text{pairwise independence}$$

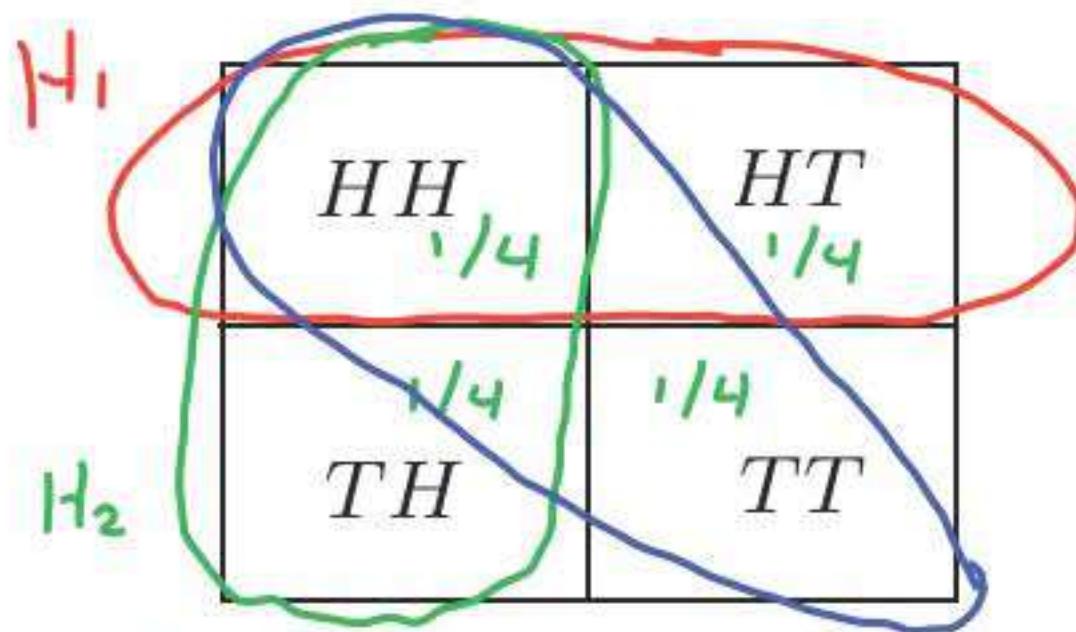
$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3)$$

Independence vs. pairwise independence

- Two independent fair coin tosses

- H_1 : First toss is H
- H_2 : Second toss is H

$$P(H_1) = P(H_2) = 1/2$$



- C : the two tosses had the same result $= \{HH, TT\}$

$$P(H_1 \cap C) = P(H_1 \cap H_2) = 1/4 \quad P(H_1) P(C) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \quad H_1, C: \text{indep.}$$

$$P(H_1 \cap H_2 \cap C) = P(HH) = 1/4 \quad P(H_1) P(H_2) P(C) = 1/8 \quad > \text{diff.}$$

$$P(C | H_1) = P(H_2 | H_1) = P(H_2) = 1/2 = P(C)$$

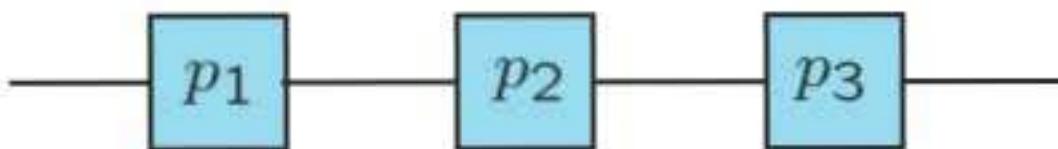
$$P(C | H_1 \cap H_2) = 1 \neq P(C) = 1/2$$

H_1 , H_2 , and C are pairwise independent, but not independent

Reliability

p_i : probability that unit i is "up"

independent units

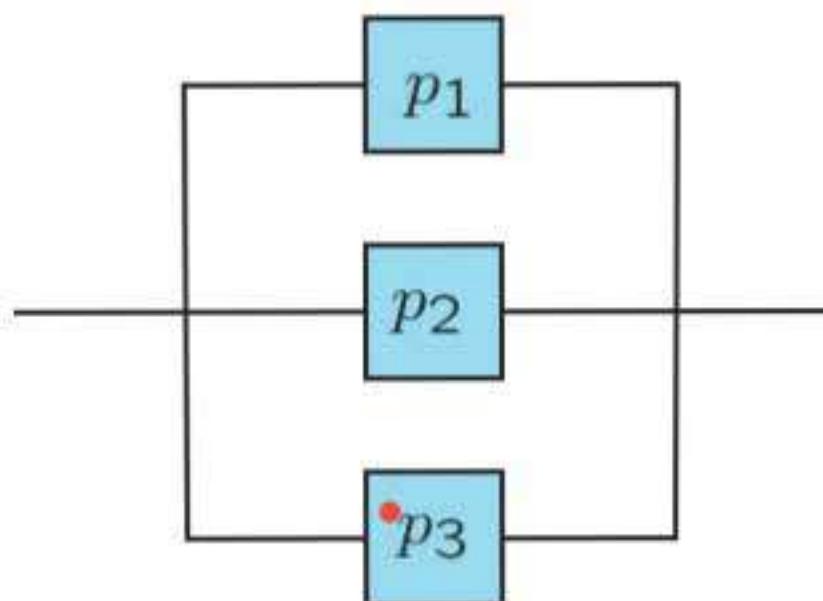


U_i : i th unit up
 U_1, U_2, \dots, U_m independent

F_i : i th unit down
 $\Rightarrow F_i$ independent

probability that system is "up"?

$$\begin{aligned} P(\text{system up}) &= P(U_1 \cap U_2 \cap U_3) \\ &= P(U_1) P(U_2) P(U_3) = p_1 p_2 p_3 \end{aligned}$$



$$\begin{aligned} P(\text{system is up}) &= P(U_1 \cup U_2 \cup U_3) \\ &= 1 - P(F_1 \cap F_2 \cap F_3) \\ &= 1 - P(F_1) P(F_2) P(F_3) \\ &= 1 - (1 - p_1)(1 - p_2)(1 - p_3) \end{aligned}$$

The king's sibling

- The king comes from a family of two children.
What is the probability that his sibling is female?

boy have precedence

$$\cancel{1/2} ?$$

$$P(\text{boy}) = P(\text{girl}) = 1/2$$

independent

BB $1/4$	BG $1/4$
GB $1/4$	GG $1/4$

2/3

- till 1 boy $\Rightarrow P(G) = 1$
- till 2 boys $\Rightarrow P(G) = 0$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 4: Counting

Discrete uniform law

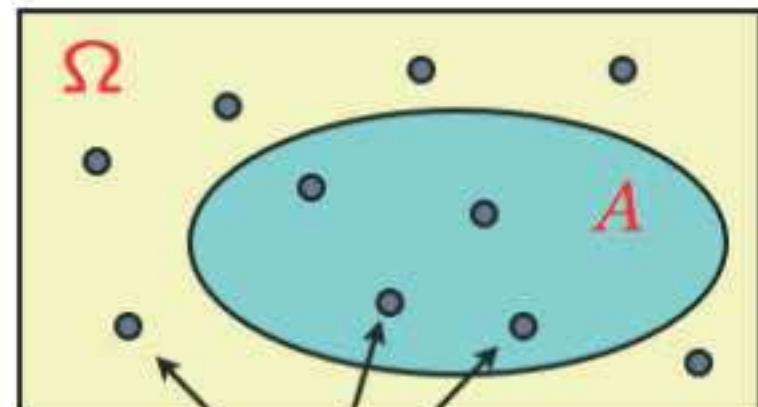
- Assume Ω consists of n equally likely elements
- Assume A consists of k elements

Then : $P(A) = \frac{\text{number of elements of } A}{\text{number of elements of } \Omega} = \frac{k}{n}$

- Basic counting principle
- Applications

permutations
combinations
partitions

number of subsets
binomial probabilities



$$\text{prob} = \frac{1}{n}$$

Basic counting principle

4 shirts

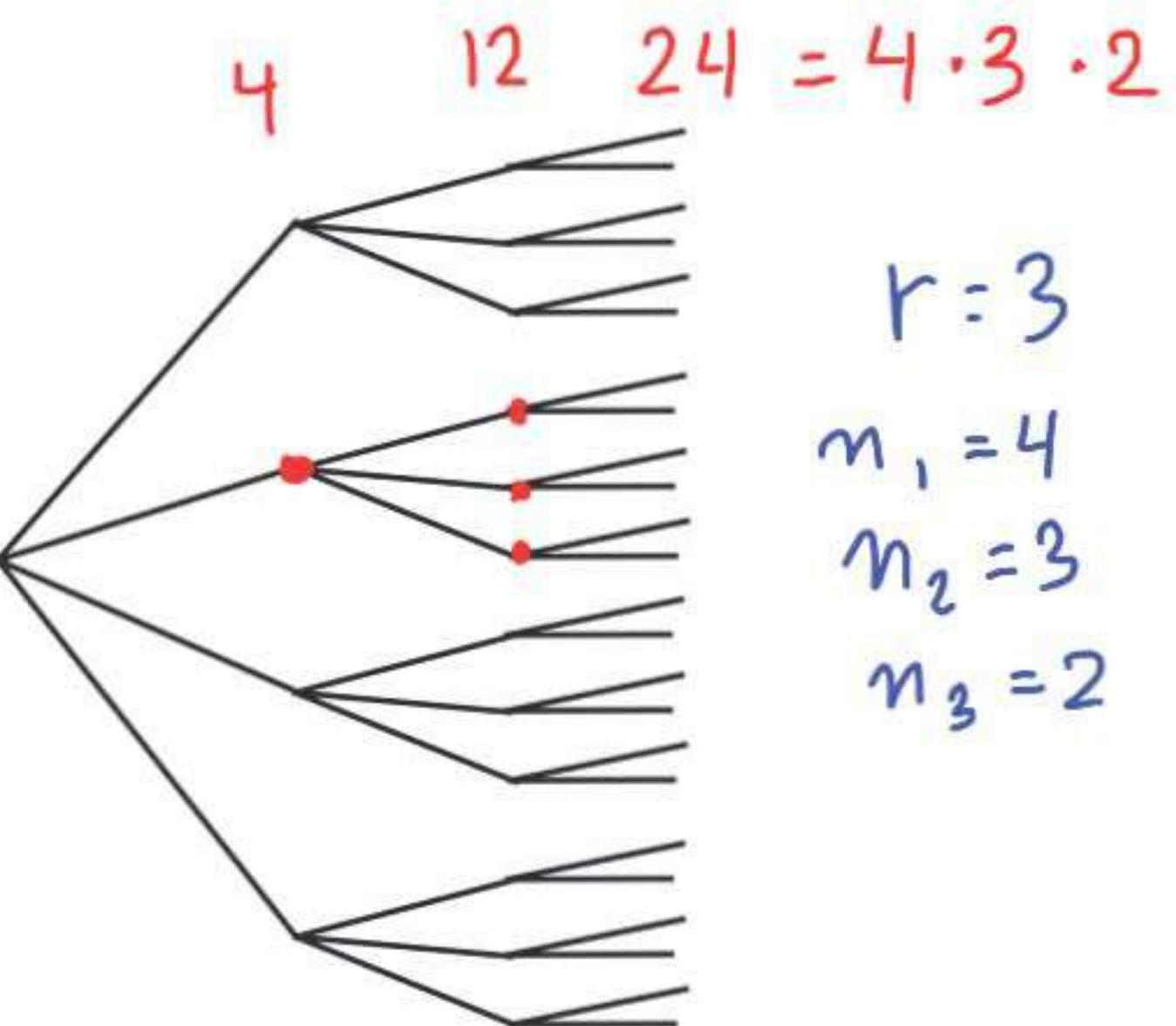
3 ties

2 jackets

Number of possible attires?

- r stages
- n_i choices at stage i

Number of choices is: $n_1 \cdot n_2 \cdots n_r$



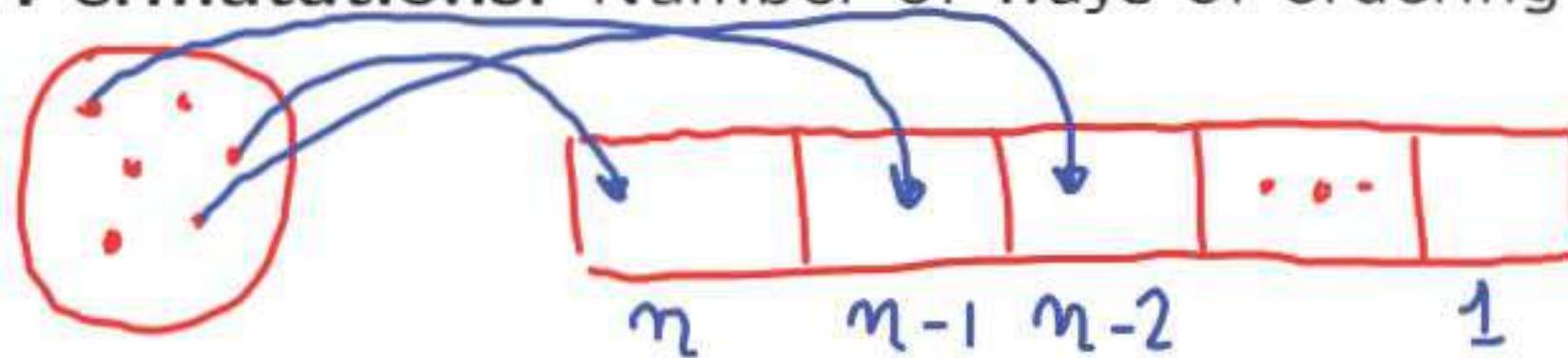
Basic counting principle examples

- Number of license plates with 2 letters followed by 3 digits:

$$26 \cdot 26 \cdot 10 \cdot 10 \cdot 10$$

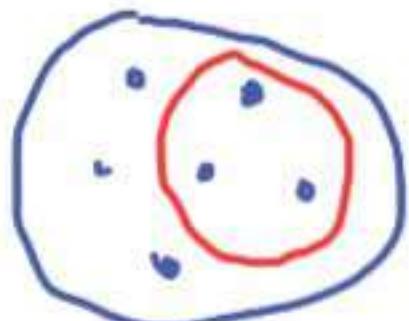
- ... if repetition is prohibited: $26 \cdot 25 \cdot 10 \cdot 9 \cdot 8$

- Permutations:** Number of ways of ordering n elements:



$$n \cdot (n-1) \cdot (n-2) \cdots \cdot 1 = n!$$

- Number of subsets of $\{1, \dots, n\}$:



$$2 \cdot 2 \cdots 2 = 2^n$$

$$\begin{array}{ccc} n=1 & \{\} & 2^1 = 2 \\ & \{\} & \emptyset \end{array}$$

Example

- Find the probability that:
six rolls of a (six-sided) die all give different numbers.

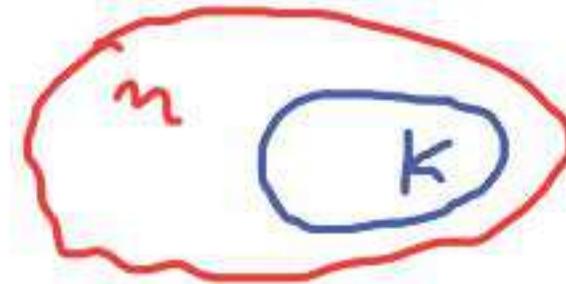
(Assume all outcomes equally likely.)

typical outcome $P(2,3,4,1,6,2) = 1/6^6$

" element of A : $(2,3,4,1,6,5) = 6!$

$$P(A) = \frac{\# \text{ in } A}{\# \text{ possible outcomes}} = \frac{6!}{6^6}$$

Combinations

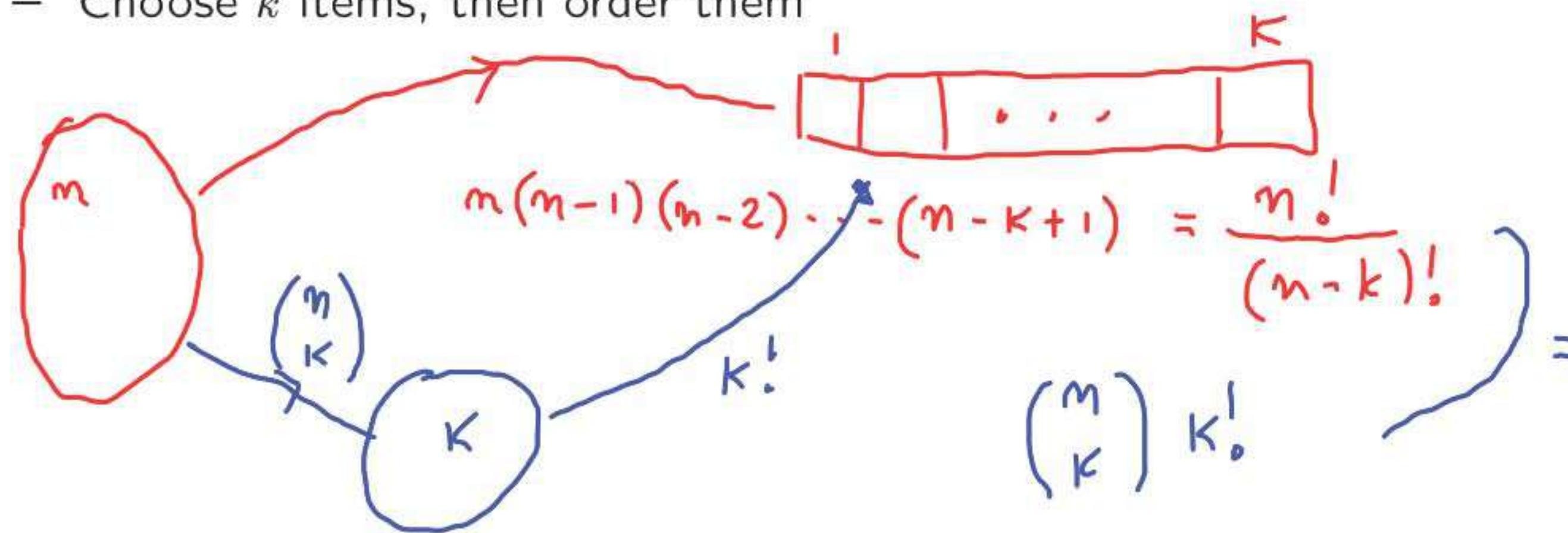


- Definition: $\binom{n}{k}$: number of k -element subsets
of a given n -element set

$$= \frac{n!}{k!(n-k)!}$$

$n = 0, 1, 2, \dots$

- Two ways of constructing an **ordered** sequence of k **distinct** items: $k = 0, 1, \dots, n$
 - Choose the k items one at a time
 - Choose k items, then order them



$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$\binom{n}{n} = 1 \quad \frac{n!}{n! 0!}$$

$$0! = 1 \quad \text{convention}$$

$$\binom{n}{0} = \frac{n!}{0! n!} = 1 \quad \emptyset$$

$$\sum_{k=0}^n \binom{n}{k} = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n} = \# \text{ all subsets} = 2^n$$

Binomial coefficient $\binom{n}{k}$ → Binomial probabilities

- $n \geq 1$ independent coin tosses; $P(H) = p$

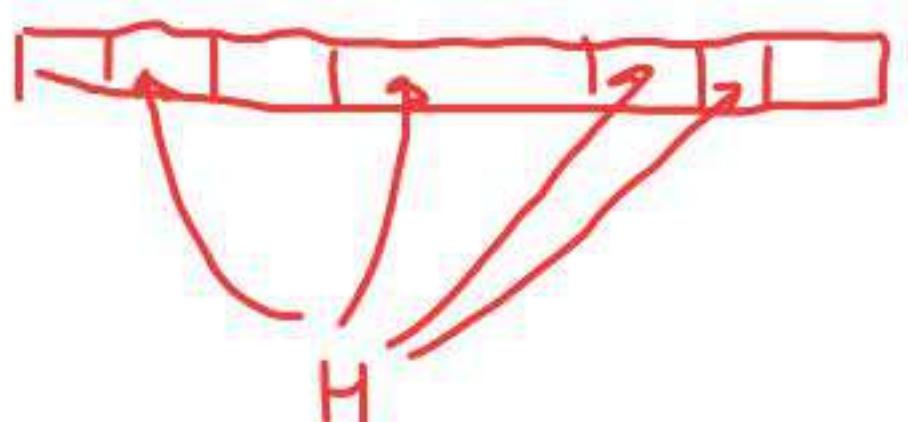
$n=6$

$$\bullet P(HTTHHH) = p(1-p)(1-p)p p p = p^4(1-p)^2$$

$$\bullet P(\text{particular sequence}) = p^{\#\text{heads}} (1-p)^{\#\text{tails}}$$

$$\bullet P(\text{particular } k\text{-head sequence}) = p^k (1-p)^{n-k}$$

$$P(k \text{ heads}) = p^k (1-p)^{n-k} \cdot (\# k\text{-head sequences})$$



$$\binom{n}{k}$$

A coin tossing problem

- Given that there were 3 heads in 10 tosses, what is the probability that the first two tosses were heads?
 - event A : the first 2 tosses were heads
 - event B : 3 out of 10 tosses were heads

Assumptions:

- independence
- $P(H) = p$

$$P(k \text{ heads}) = \binom{n}{k} p^k (1-p)^{n-k}$$

- First solution:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{\cancel{P(H_1 H_2 \text{ and one } H \text{ in tosses } 3, \dots, 10)}}{\cancel{P(B)}}$$

$$= \frac{p^2 \cdot \binom{8}{1} p^1 \cdot (1-p)^7}{\binom{10}{3} p^3 (1-p)^7} = \frac{\binom{8}{1}}{\binom{10}{3}} = \frac{8}{\binom{10}{3}}$$

A coin tossing problem

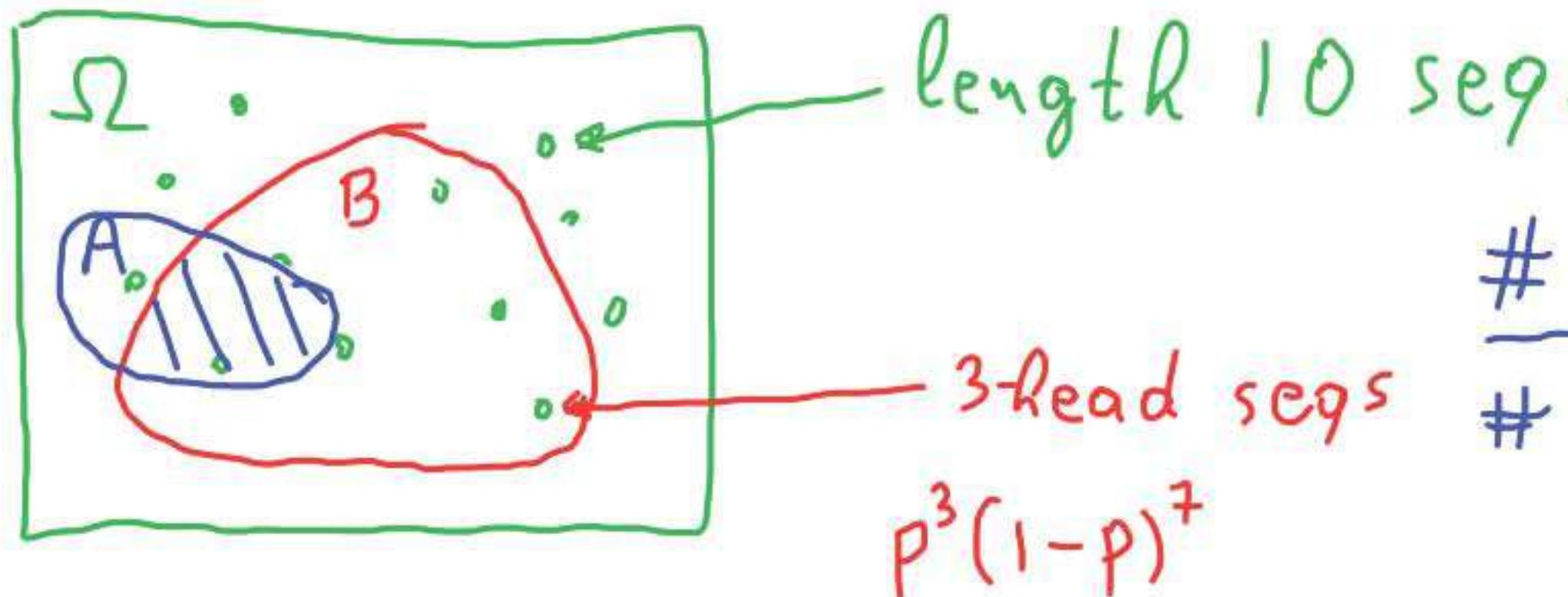
- Given that there were 3 heads in 10 tosses, what is the probability that the first two tosses were heads?
 - event A : the first 2 tosses were heads
 - event B : 3 out of 10 tosses were heads

Assumptions:

- independence
- $P(H) = p$

$$P(k \text{ heads}) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Second solution: Conditional probability law (on B) is uniform

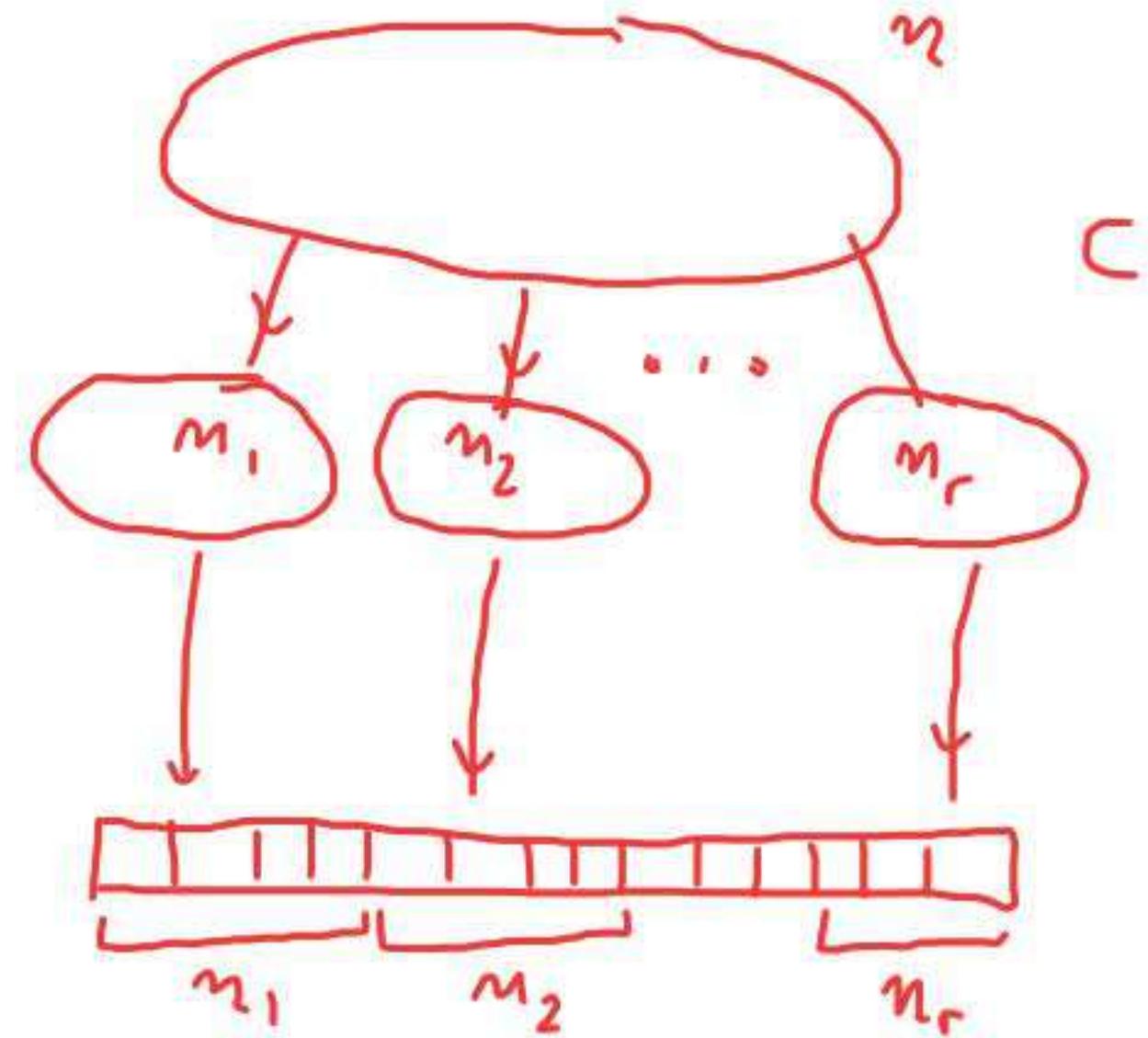


$$\frac{\# \text{ in } (A \cap B)}{\# \text{ in } B} = \frac{8}{\binom{10}{3}}$$

Partitions

- $n \geq 1$ distinct items; $r \geq 1$ persons give n_i items to person i
 - here n_1, \dots, n_r are given nonnegative integers
 - with $n_1 + \dots + n_r = n$
- Ordering n items: $n!$
 - Deal n_i to each person i , and then order

$$n_1! n_2! \dots n_r! = n!$$



$$r=2 \quad m_1 = k \quad m_2 = m - k$$

number of partitions = $\frac{n!}{n_1! n_2! \dots n_r!}$ (multinomial coefficient)

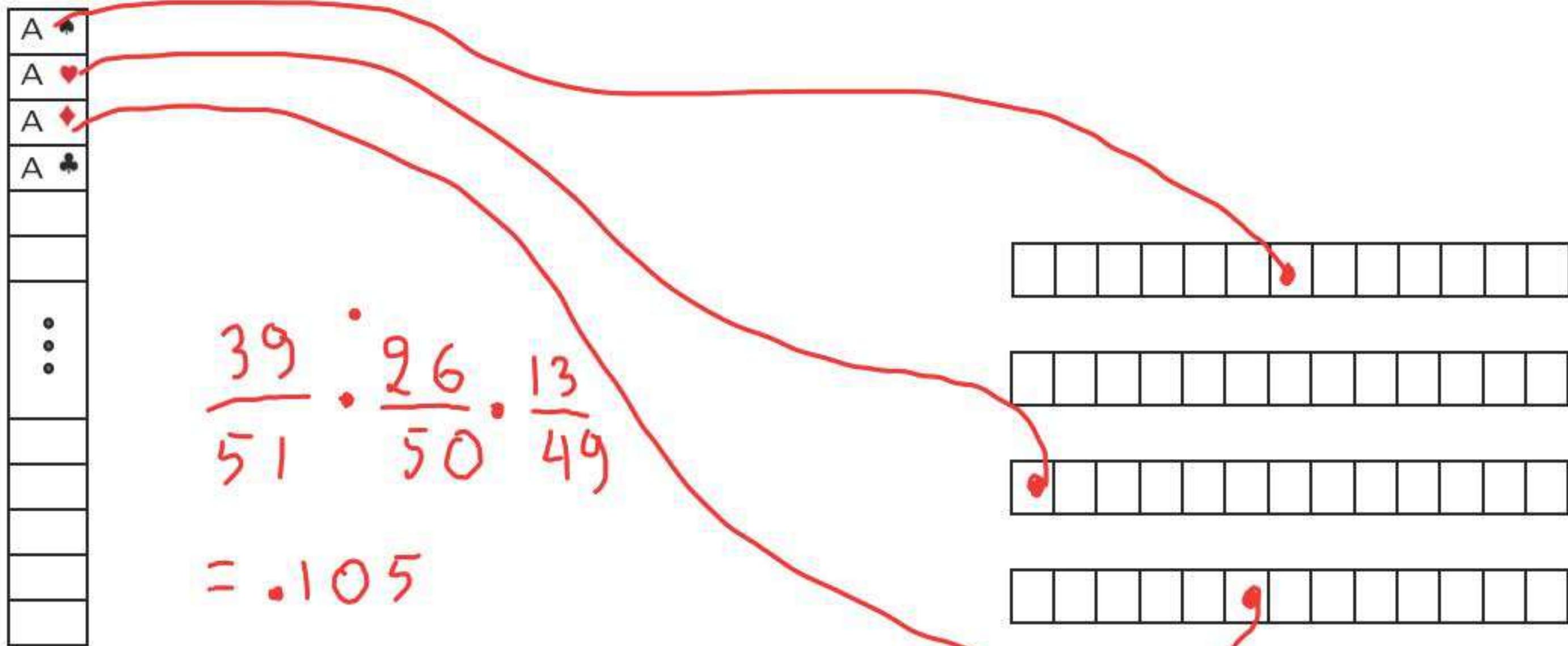
Example: 52-card deck, dealt (fairly) to four players.
Find $P(\text{each player gets an ace})$

- Outcomes are: **partition equally likely**
 - number of outcomes: $\frac{52!}{13! 13! 13! 13!}$
- Constructing an outcome with one ace for each person:
 - distribute the aces $4 \cdot 3 \cdot 2 \cdot 1$
 - distribute the remaining 48 cards $\frac{48!}{12! 12! 12! 12!}$
- Answer:
$$\frac{4 \cdot 3 \cdot 2 \cdot \frac{48!}{12! 12! 12! 12!}}{\frac{52!}{13! 13! 13! 13!}}$$

Example: 52-card deck, dealt (fairly) to four players.
Find $P(\text{each player gets an ace})$

A smart solution

Stack the deck, aces on top



MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

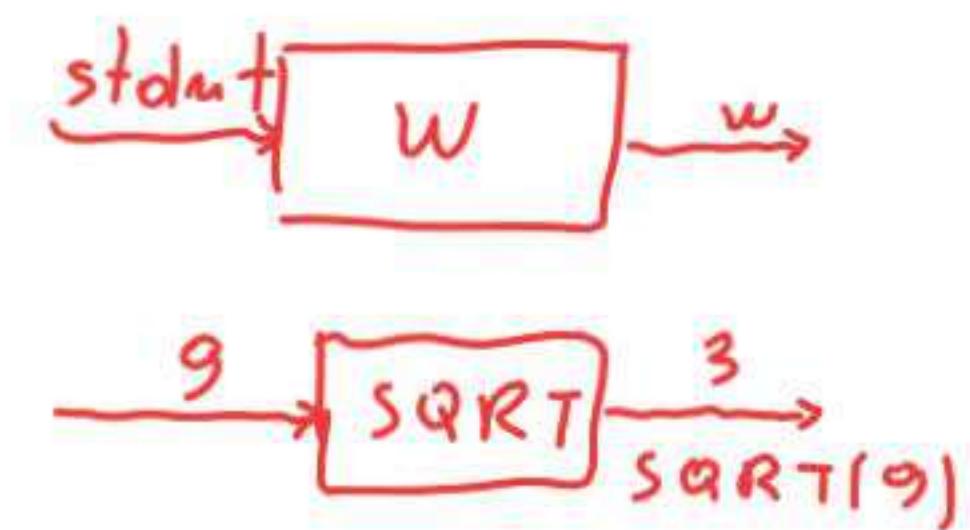
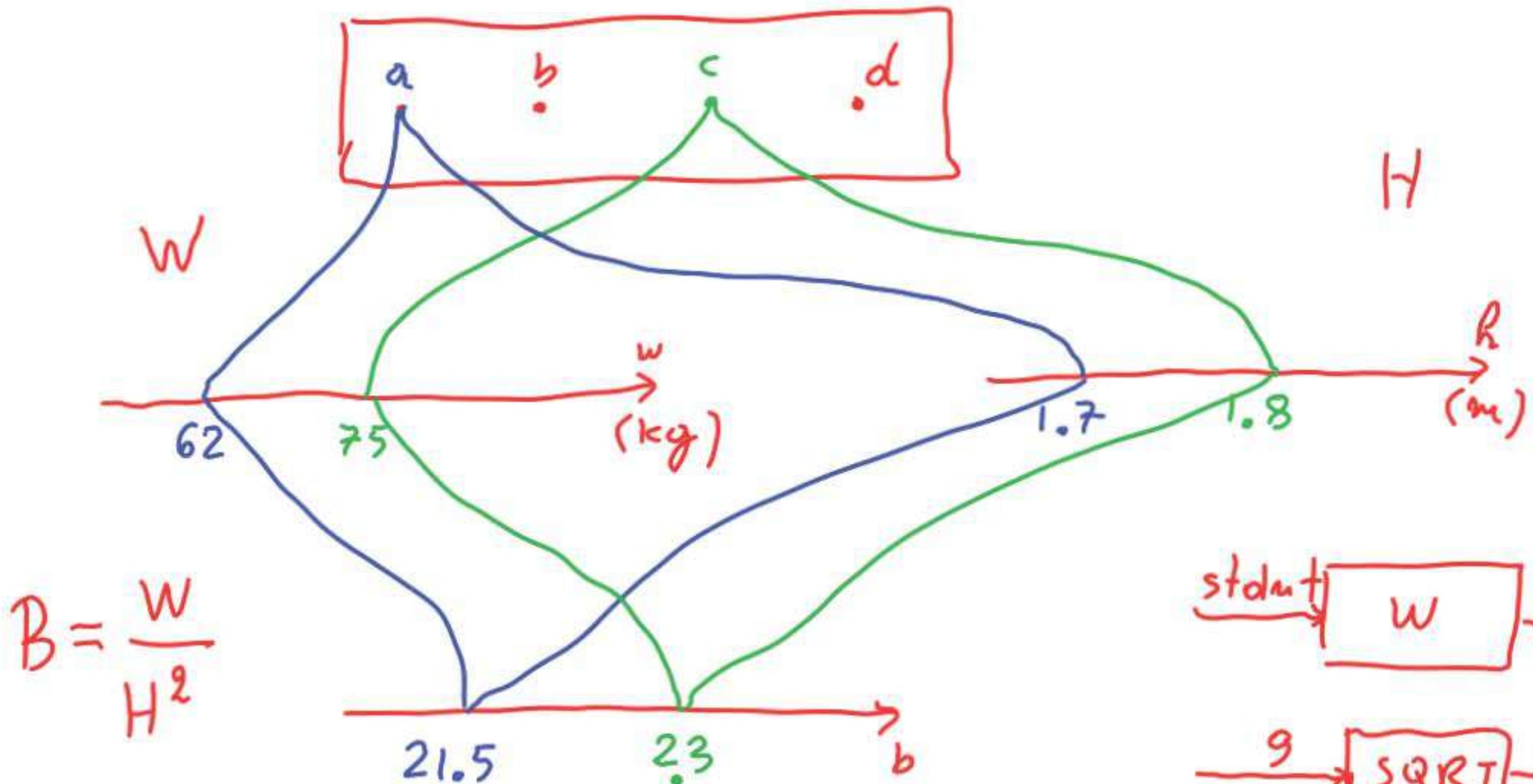
The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 5: Discrete random variables: probability mass functions and expectations

- Random variables: the idea and the definition
 - **Discrete:** take values in finite or countable set
- Probability mass function (PMF)
- Random variable examples
 - Bernoulli
 - Uniform
 - Binomial
 - Geometric
- Expectation (mean) and its properties
 - The expected value rule
 - Linearity

Random variables: the idea



Random variables: the formalism

- A random variable (“r.v.”) associates a value (a number) to every possible outcome
- Mathematically: A function from the sample space Ω to the real numbers
- It can take discrete or continuous values

Notation: random variable X numerical value x

- We can have several random variables defined on the same sample space
- A function of one or several random variables is also a random variable
 - meaning of $X + Y$:

r.v takes value $x+y$,
when X takes value x , Y takes value y

Probability mass function (PMF) of a discrete r.v. X

- It is the “probability law” or “probability distribution” of X
- If we fix some x , then “ $X = x$ ” is an event

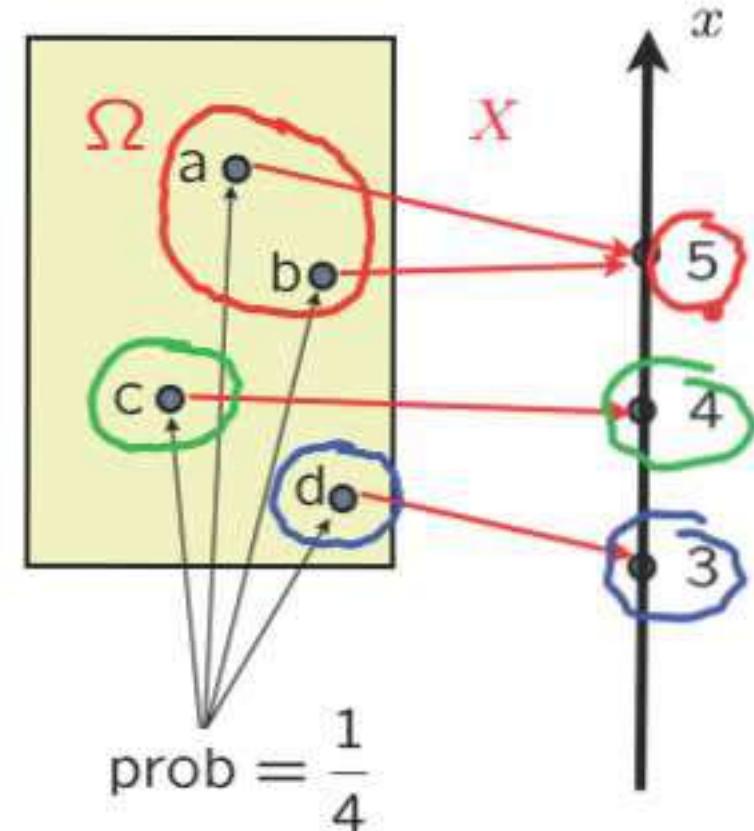
$$x=5 \quad X=5 \quad \{\omega : X(\omega) = 5\} = \{a, b\}$$

$$P_X(5) = 1/2$$

$$p_X(x) = P(X = x) = P(\{\omega \in \Omega \text{ s.t. } X(\omega) = x\})$$

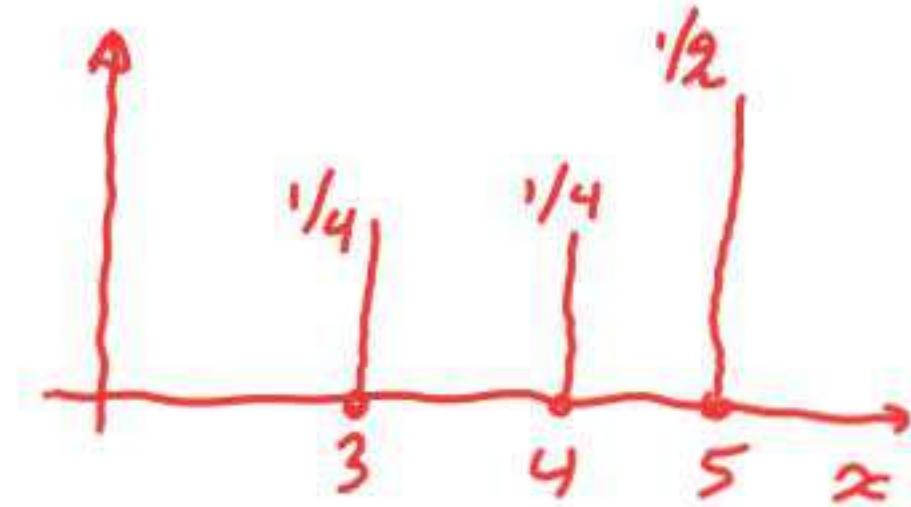
- Properties:** $p_X(x) \geq 0$

$$\sum_x p_X(x) = 1$$



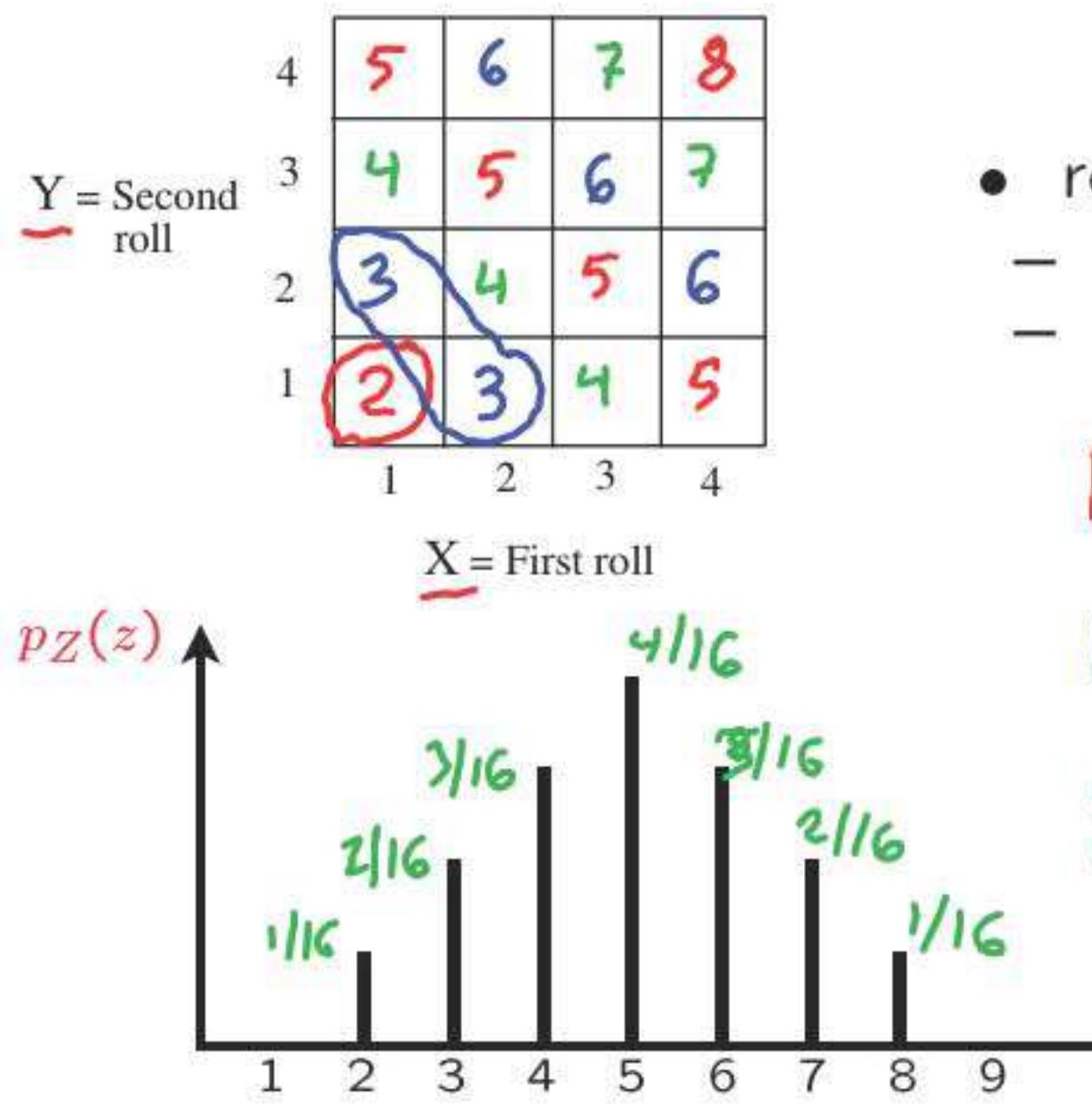
$$P_Y(y)$$

$$P_X$$



PMF calculation

- Two rolls of a tetrahedral die
- Let every possible outcome have probability $1/16$



- $Z = X + Y$
- Find $p_Z(z)$ for all z
- repeat for all z :
 - collect all possible outcomes for which Z is equal to z
 - add their probabilities

$$P_Z(2) = P(Z=2) = 1/16$$

$$P_Z(3) = P(Z=3) = 2/16$$

$$P_Z(4) = P(Z=4) = 3/16$$

•

•

•

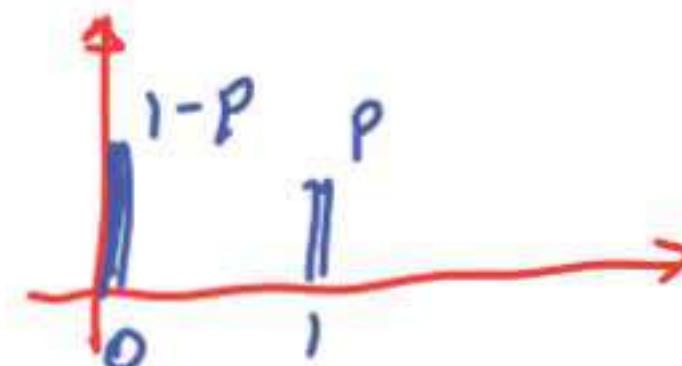
•

The simplest random variable: Bernoulli with parameter $p \in [0, 1]$

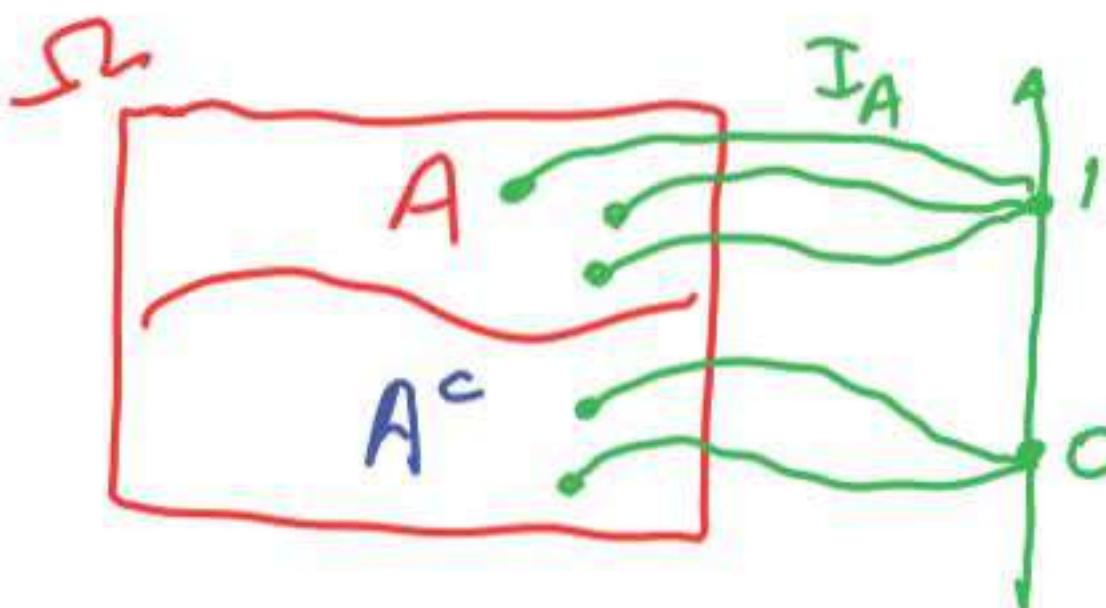
$$X = \begin{cases} 1, & \text{w.p. } p \\ 0, & \text{w.p. } 1 - p \end{cases}$$

$$P_X(0) = 1 - p$$

$$P_X(1) = p$$



- Models a trial that results in success/failure, Heads/Tails, etc.
- Indicator r.v. of an event A : $I_A = 1$ iff A occurs



$$P_{I_A}(1) = P(I_A = 1) = P(A)$$



Discrete uniform random variable; parameters a, b

- Parameters: integers a, b ; $a \leq b$

- Experiment: Pick one of $a, a+1, \dots, b$ at random; all equally likely

- Sample space: $\{a, a+1, \dots, b\}$

$b - a + 1$ possible values

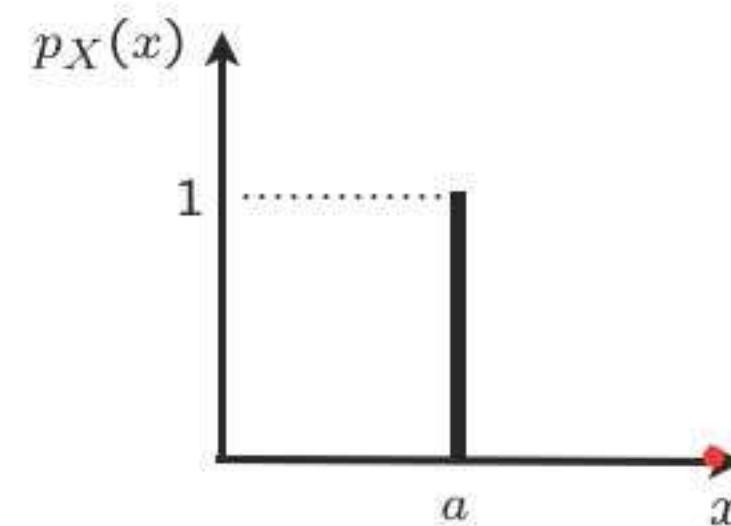
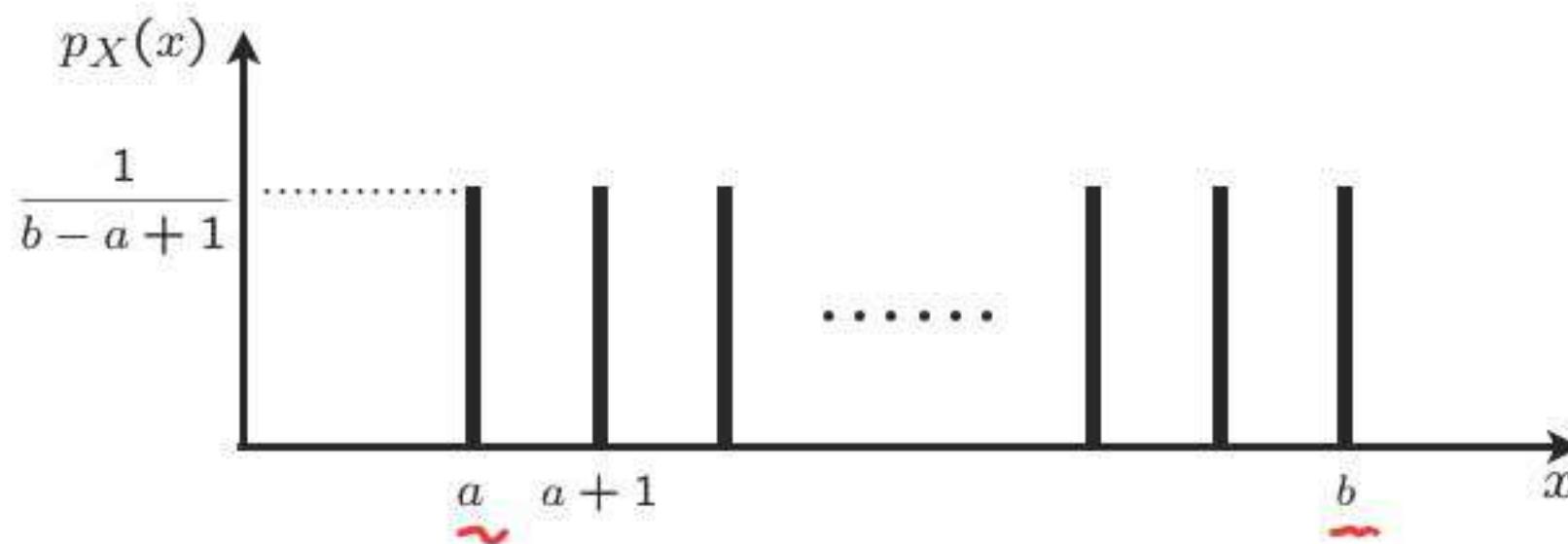
- Random variable X : $X(\omega) = \omega$

11:52:26 $\{0, 1, \dots, 59\}$

- Model of: complete ignorance

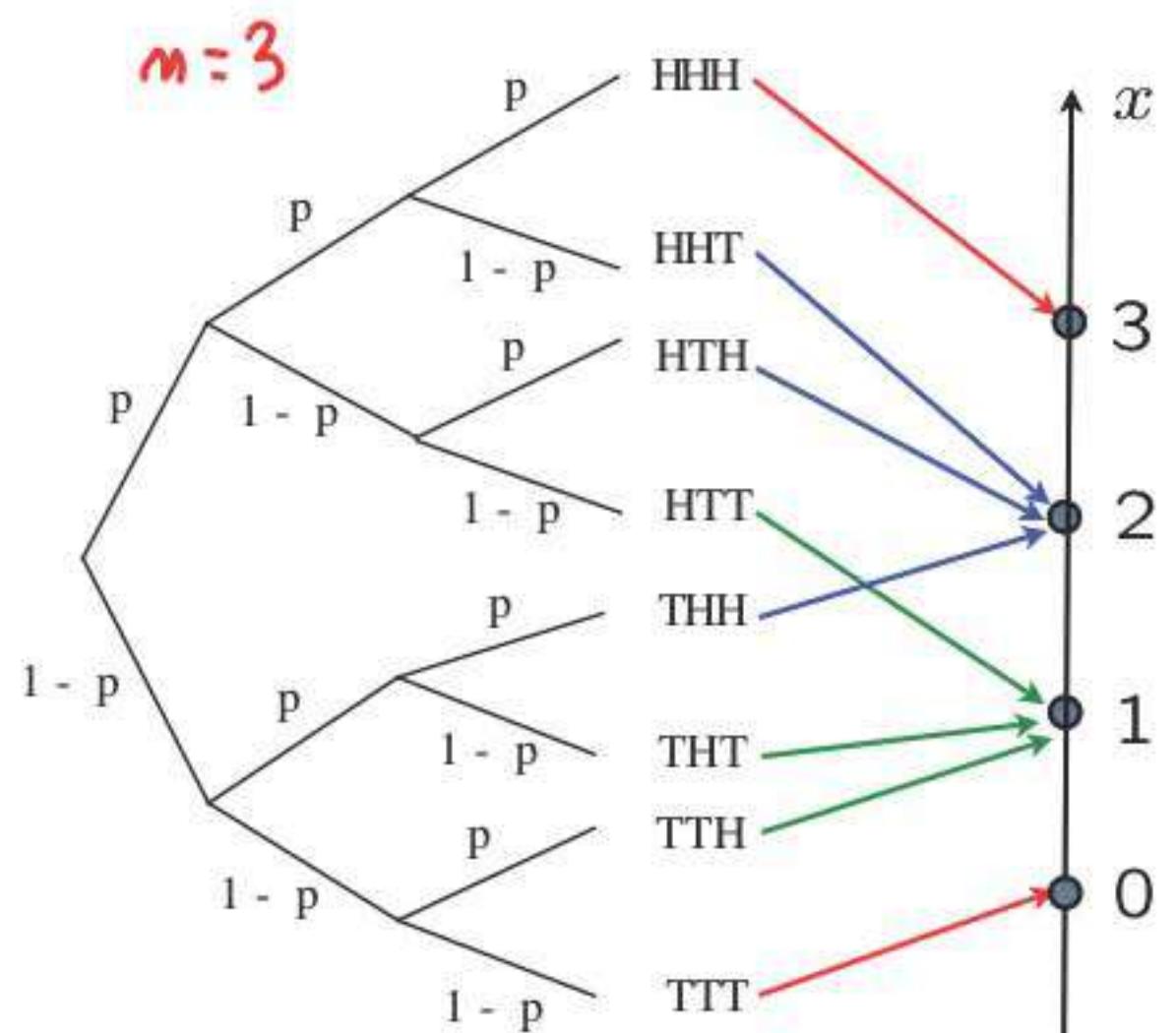
Special case: $a = b$

constant/deterministic r.v.



Binomial random variable; **parameters:** positive integer n ; $p \in [0, 1]$

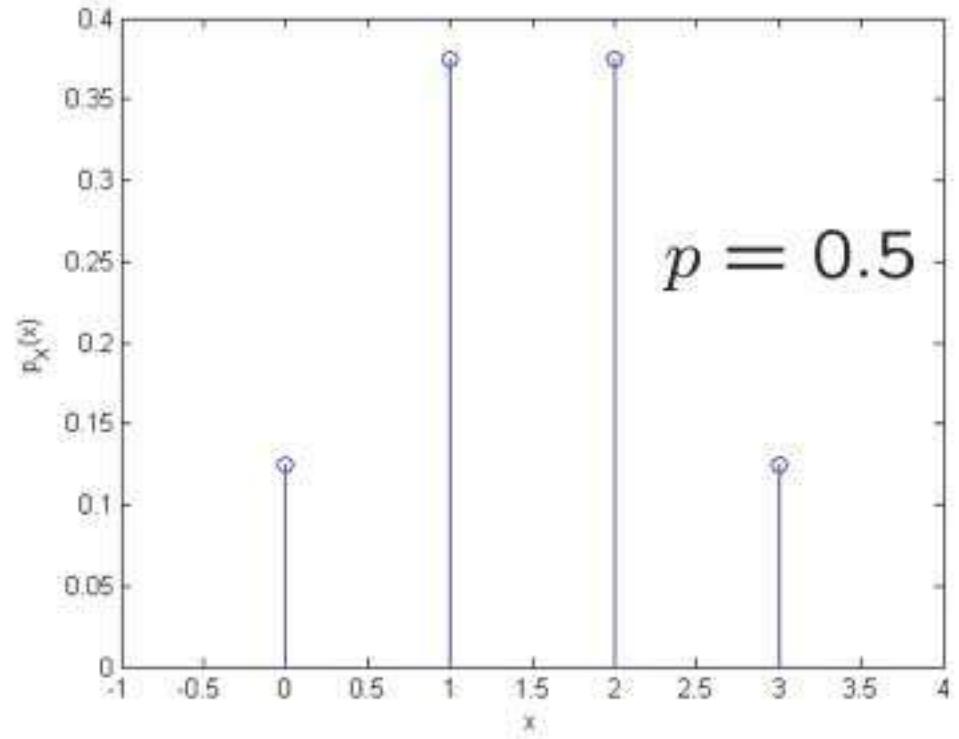
- **Experiment:** n independent tosses of a coin with $P(\text{Heads}) = p$
- **Sample space:** Set of sequences of H and T, of length n
- **Random variable X :** number of Heads observed
- **Model of:** number of successes in a given number of independent trials



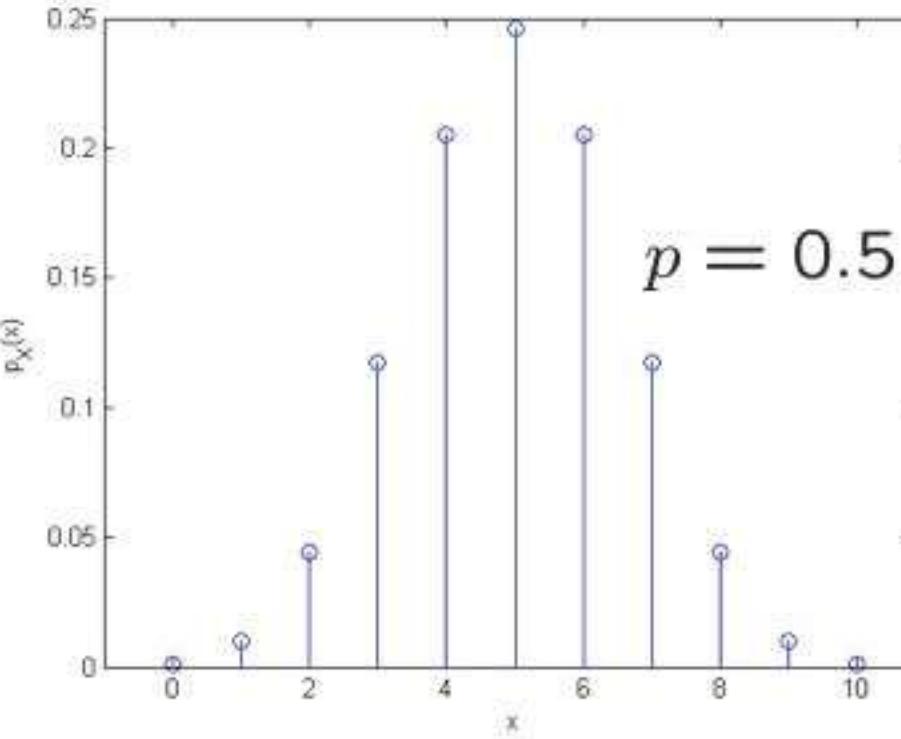
$$\begin{aligned}
 P_X(2) &= P(X=2) \\
 &= P(HHT) + P(HTH) + P(THH) \\
 &= 3p^2(1-p) \approx \binom{3}{2} p^2 (1-p)
 \end{aligned}$$

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

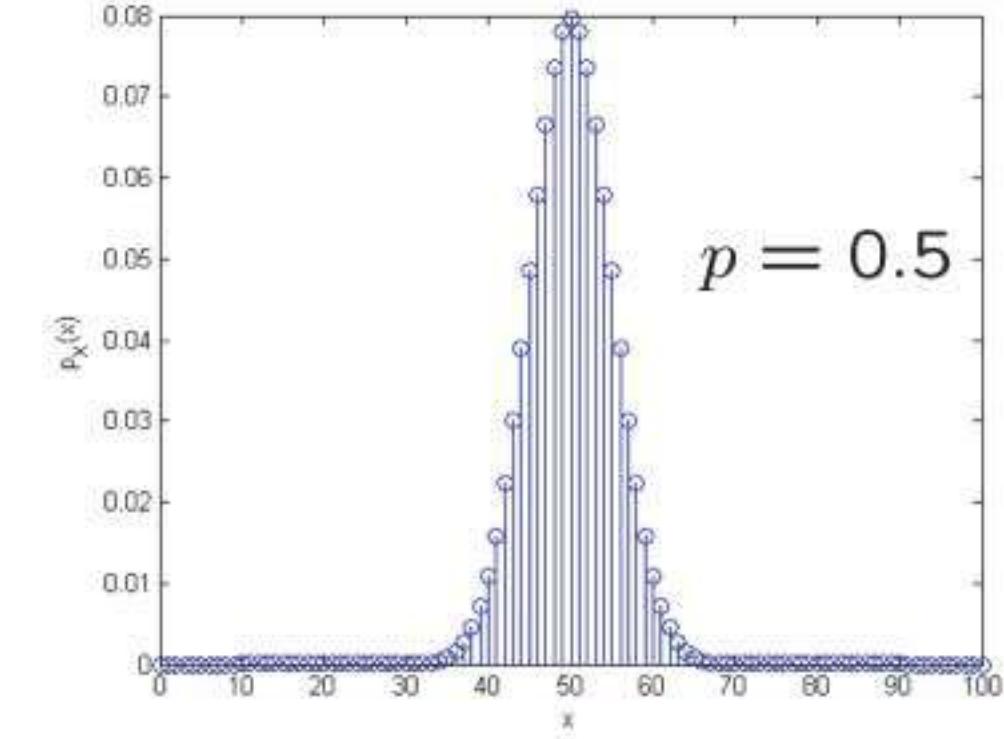
$n = 3$



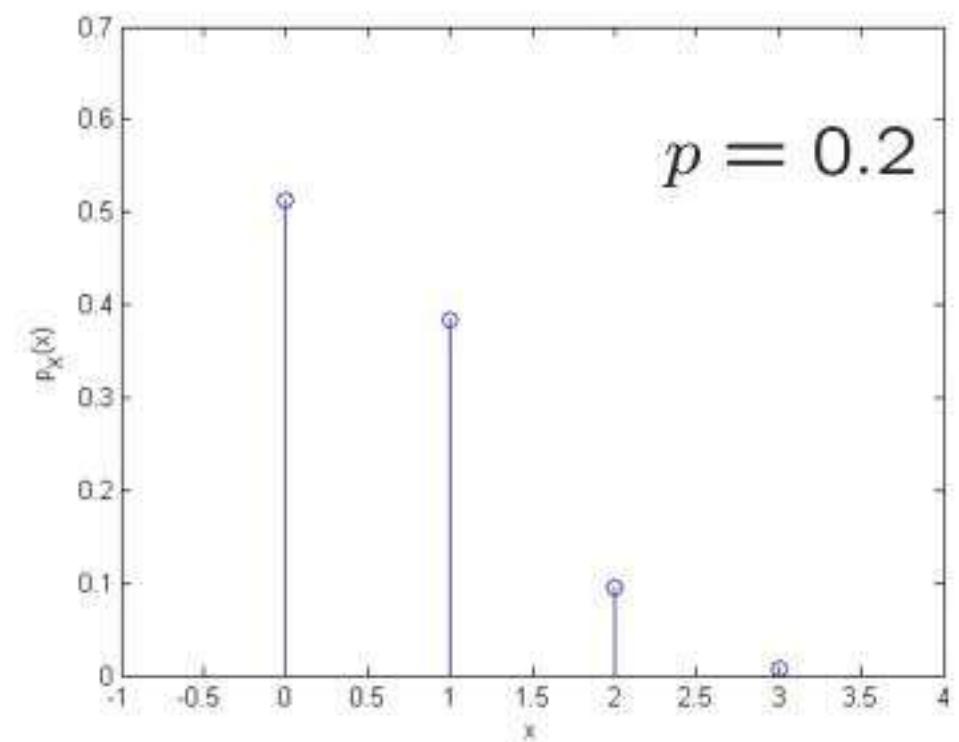
$n = 10$



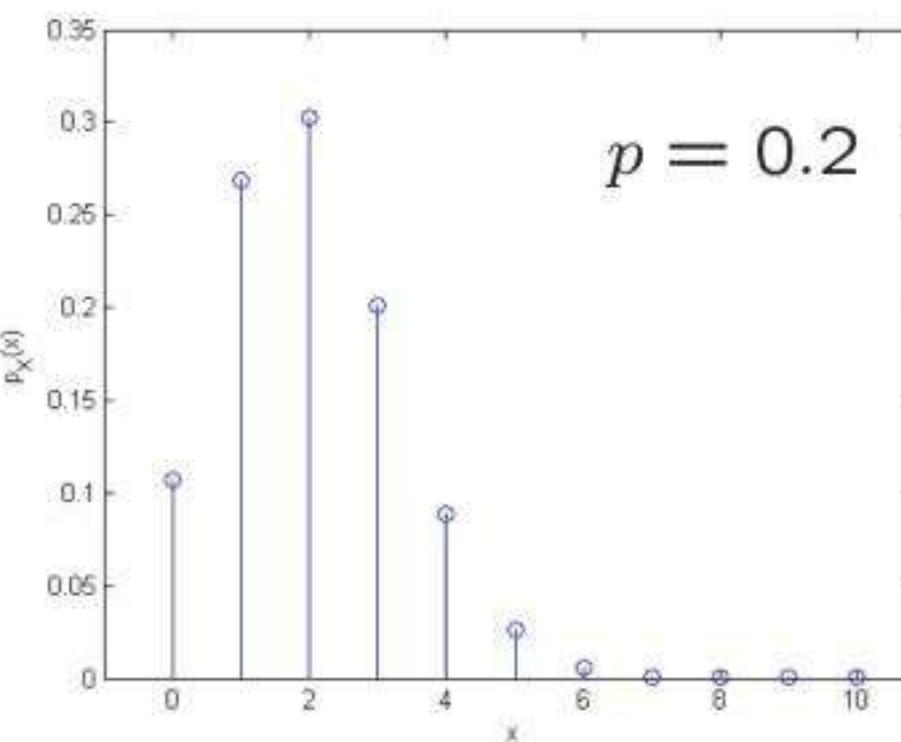
$n = 100$



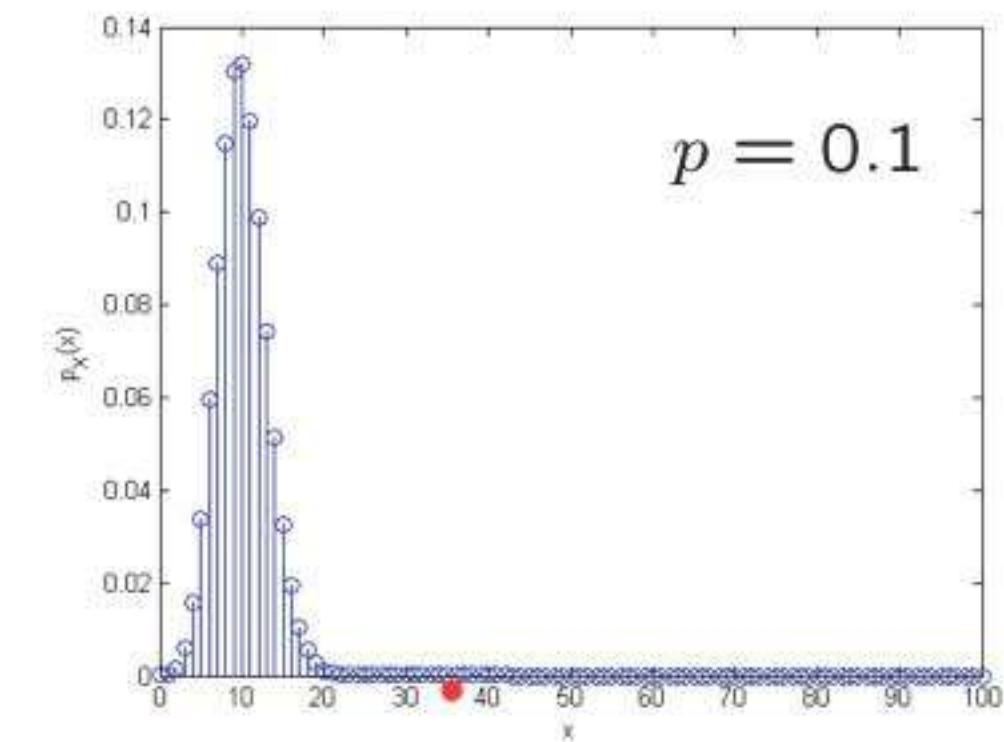
$p = 0.2$



$p = 0.2$



$p = 0.1$



Geometric random variable; parameter p : $0 < p \leq 1$

- **Experiment:** infinitely many independent tosses of a coin; $P(\text{Heads}) = p$

- **Sample space:** Set of infinite sequences of H and T

TTTTHHT...

- **Random variable X :** number of tosses until the first Heads $X = 5$

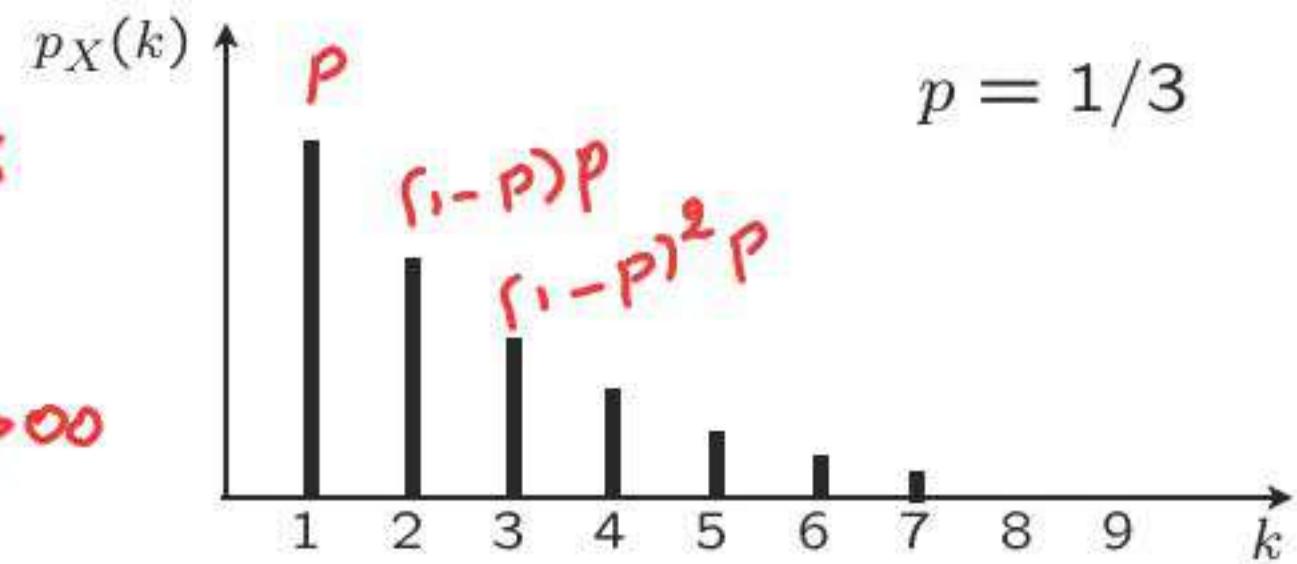
- **Model of:** waiting times; number of trials until a success

$$p_X(k) = P(X=k) = P(T \underbrace{\dots}_{k-1} T H) = (1-p)^{k-1} p \quad k=1, 2, 3, \dots$$

$$\left. \begin{aligned} P(\text{no Heads ever}) \\ (TTT\dots) \\ "X=\infty" \end{aligned} \right\} = 0$$

$$\leq P(T \underbrace{\dots}_{k} T) = (1-p)^k$$

↓
0 $\rightarrow \infty$



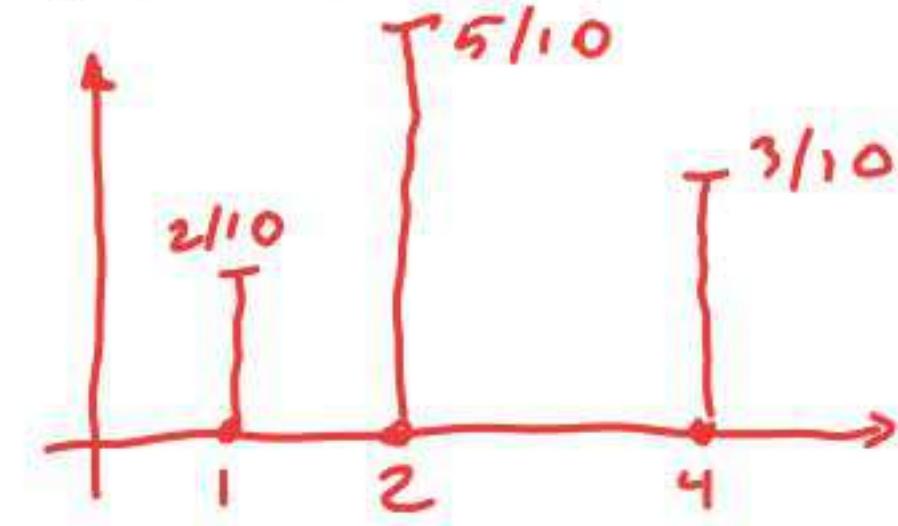
Expectation/mean of a random variable

- **Motivation:** Play a game 1000 times.
Random gain at each play described by:
- “Average” gain:

$$\begin{aligned} & \frac{1 \cdot 200 + 2 \cdot 500 + 4 \cdot 300}{1000} \\ &= 1 \cdot \frac{2}{10} + 2 \cdot \frac{5}{10} + 4 \cdot \frac{3}{10} \end{aligned}$$

$$X = \begin{cases} 1, & \text{w.p. } 2/10 \\ 2, & \text{w.p. } 5/10 \\ 4, & \text{w.p. } 3/10 \end{cases}$$

~ 200
 ~ 500
 ~ 300



- **Definition:** $E[X] = \sum_x x p_X(x)$

- **Caution:** If we have an infinite sum, it needs to be well-defined.
We assume $\sum_x |x| p_X(x) < \infty$

Expectation of a Bernoulli r.v.

$$X = \begin{cases} 1, & \text{w.p. } p \\ 0, & \text{w.p. } 1 - p \end{cases}$$

$$E[X] = 1 \cdot p + 0 \cdot (1 - p) = p$$

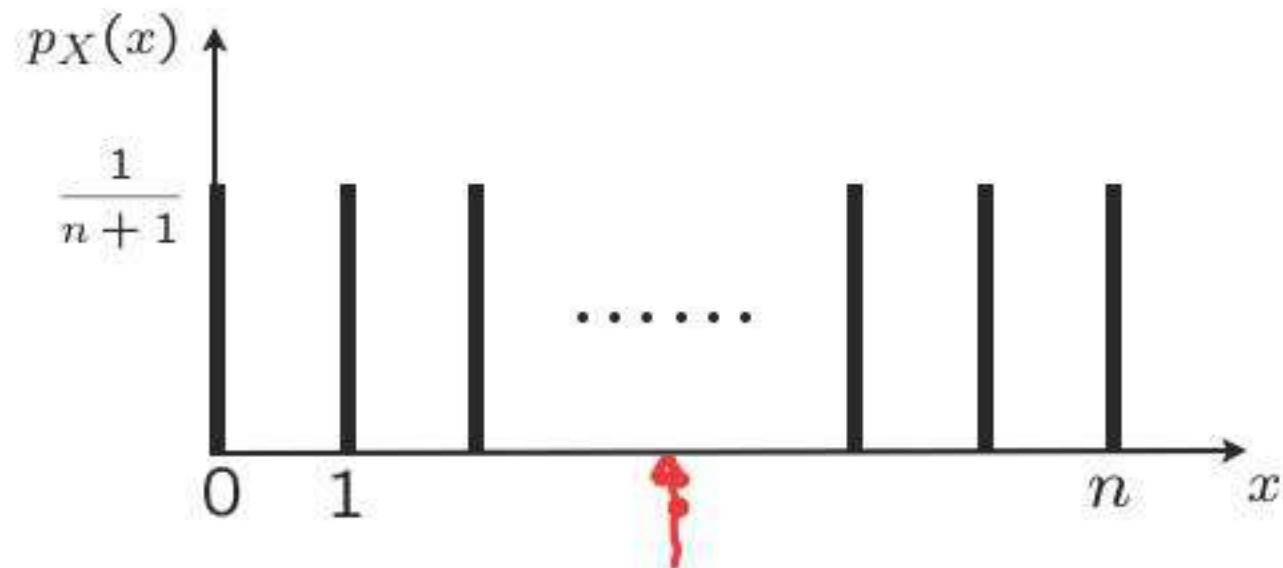
If X is the indicator of an event A , $X = I_A$:

$$X = 1 \quad \text{iff} \quad A \text{ occurs} \quad p = P(A)$$

$$E[I_A] = P(A)$$

Expectation of a uniform r.v.

- Uniform on $0, 1, \dots, n$



• **Definition:** $E[X] = \sum_x x p_X(x)$

$$E[X] = 0 \cdot \frac{1}{n+1} + 1 \cdot \frac{1}{n+1} + \dots + n \cdot \frac{1}{n+1}$$

$$= \frac{1}{n+1} (0 + 1 + \dots + n) = \frac{1}{n+1} \cdot \frac{n(n+1)}{2} = \frac{n}{2}$$

Expectation as a population average

- n students
- Weight of i th student: x_i
- Experiment: pick a student at random, all equally likely
- Random variable X : weight of selected student
 - assume the x_i are distinct

$$p_X(x_i) = \frac{1}{n}$$

$$E[X] = \sum_i x_i \cdot \frac{1}{n} = \frac{1}{n} \sum_i x_i$$

Elementary properties of expectations

- If $X \geq 0$, then $E[X] \geq 0$

for all $\omega: X(\omega) \geq 0$

- Definition: $E[X] = \sum_x xp_X(x)$

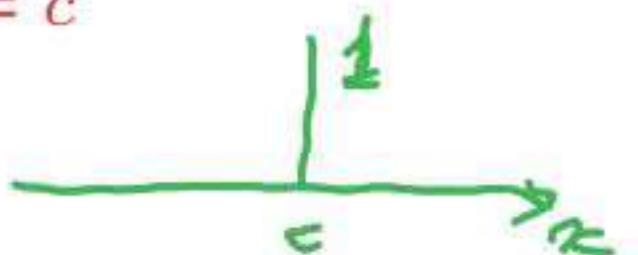
≥ 0 ≥ 0 ≥ 0

- If $a \leq X \leq b$, then $a \leq E[X] \leq b$

for all $\omega: a \leq X(\omega) \leq \bar{b}$

$$\begin{aligned} E[X] &= \sum_x x p_X(x) \geq \sum_x a p_X(x) \\ &= a \sum_x p_X(x) = a \cdot 1 = a \end{aligned}$$

- If c is a constant, $E[c] = c$

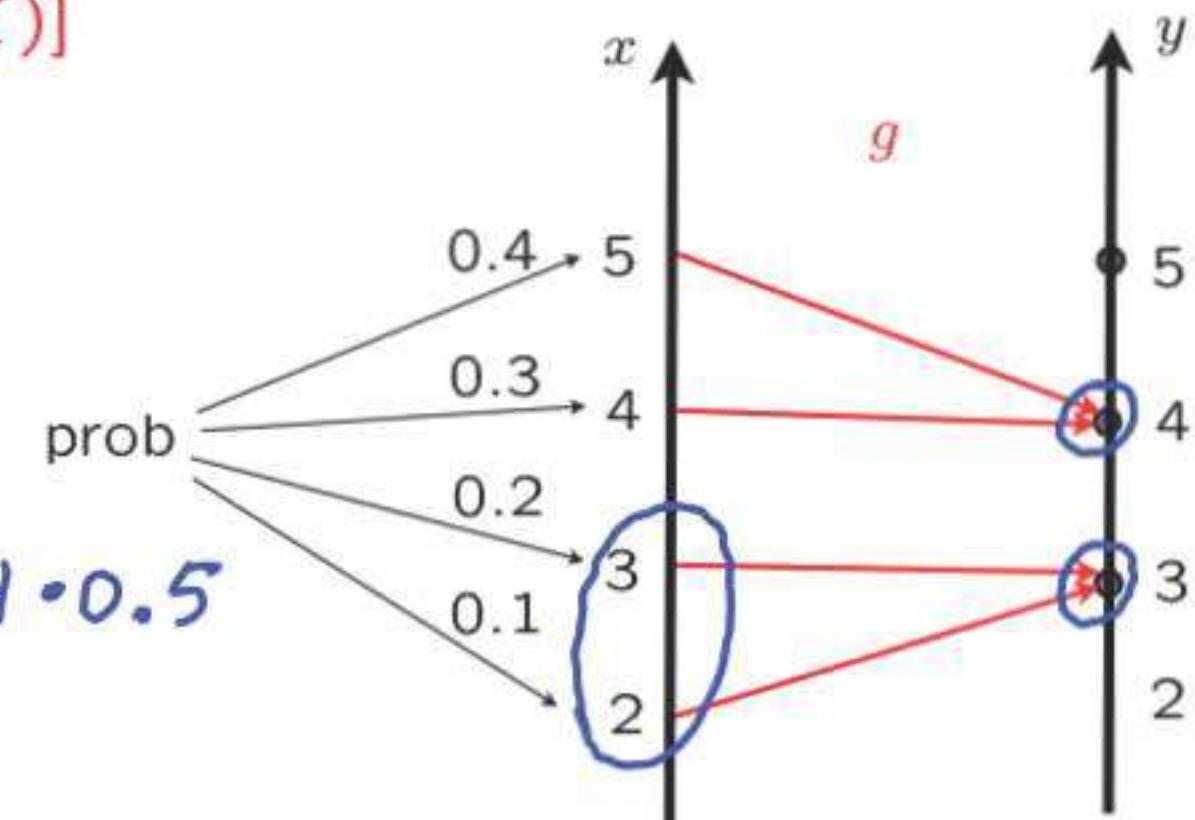


$$E[c] = c \cdot p(c) = c$$

The expected value rule, for calculating $E[g(X)]$

- Let X be a r.v. and let $Y = g(X)$
- Averaging over y : $E[Y] = \sum_y y p_Y(y)$
 $3 \cdot (0.1 + 0.2) + 4 \cdot (0.3 + 0.4)$
- Averaging over x : $3 \cdot 0.1 + 3 \cdot 0.2 + 4 \cdot 0.3 + 4 \cdot 0.5$

$$E[Y] = E[g(X)] = \sum_x g(x) p_X(x)$$



Proof:

$$\sum_y \sum_{x: g(x)=y} g(x) p_X(x)$$

$$= \sum_y \sum_{x: g(x)=y} y p_X(x) = \sum_y y \sum_{x: g(x)=y} p_X(x)$$

$$= \sum_y y p_Y(y) = E[Y]$$

- $E[X^2] = \sum_x x^2 p_X(x)$
 $g(x) = x^2$

- Caution: In general, $E[g(X)] \neq g(E[X])$

$$E[X^2] \neq (E[X])^2$$

Linearity of expectation: $E[aX + b] = aE[X] + b$

$X = \text{salary}$ $E[X] = \text{average salary}$

$Y = \text{new salary} = 2X + 100$ $E[Y] = E[2X + 100] = 2E[X] + 100$

- Intuitive
- **Derivation**, based on the expected value rule:

$$g(x) = ax + b$$
$$Y = g(X)$$

$$E[Y] = \sum_x g(x) p_x(x)$$

$$= \sum_x (ax + b) p_x(x) = a \sum_x x p_x(x) + b \underbrace{\sum_x p_x(x)}_1$$

$$E[g(x)] = g(E[x]) = aE[X] + b$$

exceptional g

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

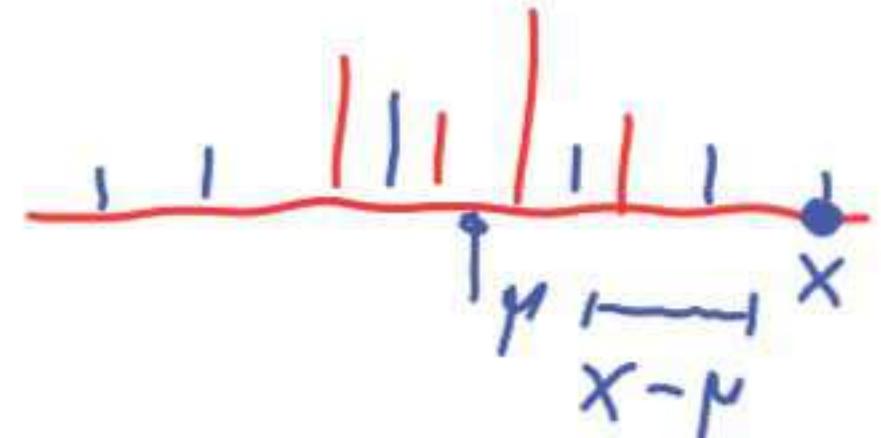
LECTURE 6: Variance; Conditioning on an event; Multiple random variables

- Variance and its properties
 - Variance of the Bernoulli and uniform PMFs
- Conditioning a r.v. on an event
 - Conditional PMF, mean, variance
 - Total expectation theorem
- Geometric PMF
 - Memorylessness
 - Mean value
- Multiple random variables
 - Joint and marginal PMFs
 - Expected value rule
 - Linearity of expectations
- The mean of the binomial PMF

Variance — a measure of the spread of a PMF

- Random variable X , with mean $\mu = E[X]$
- Distance from the mean: $X - \mu$
- Average distance from the mean?

$$E[X - \mu] = E[X] - \mu = \mu - \mu = 0$$



- **Definition of variance:** $\text{var}(X) = E[(X - \mu)^2] \geq 0$
- Calculation, using the expected value rule, $E[g(X)] = \sum_x g(x)p_X(x)$

$$g(x) = (x - \mu)^2 \quad \text{var}(X) = E[g(X)] = \sum_x (x - \mu)^2 p_X(x)$$

Standard deviation: $\sigma_X = \sqrt{\text{var}(X)}$

Properties of the variance

- Notation: $\mu = E[X]$

$$\text{var}(aX + b) = a^2 \text{var}(X)$$

$$\begin{aligned}\text{var}(3 - 4x) \\ &= (-4)^2 \text{var}(x) \\ &= 16 \text{var}(x)\end{aligned}$$

- Let $Y = X + b$

$$Y = E[Y] = \mu + b$$

$$\text{var}(Y) = E[(Y - Y)^2] = E[(X + b - (\mu + b))^2] = E[(X - \mu)^2] = \text{var}(X)$$

- Let $Y = aX$

$$Y = E[Y] = a\mu$$

$$\text{var}(Y) = E[(aX - a\mu)^2] = E[a^2(X - \mu)^2] = a^2 E[(X - \mu)^2] = a^2 \text{var}(X)$$

A useful formula:

$$\text{var}(X) = E[X^2] - (E[X])^2$$

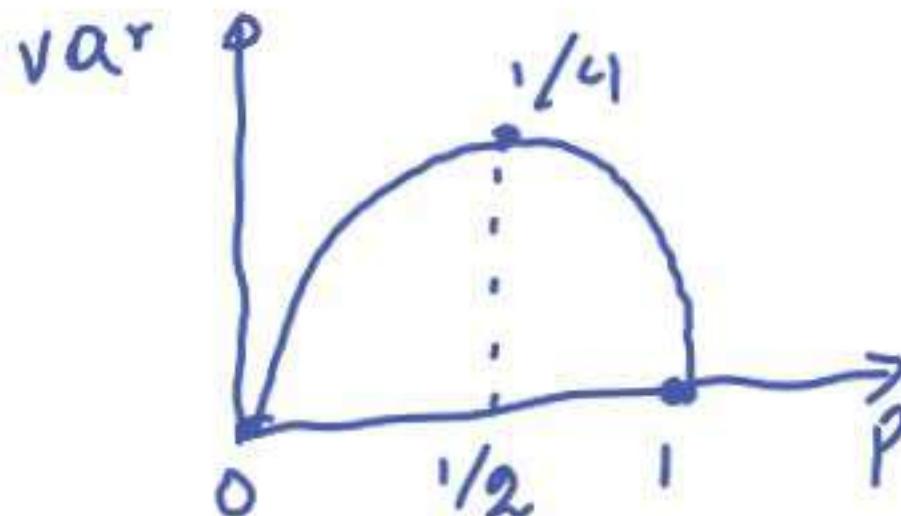
$$\text{var}(X) = E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2]$$

$$= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - (E[X])^2$$

Variance of the Bernoulli

$$E[X] = p$$

$$X = \begin{cases} 1, & \text{w.p. } p \\ 0, & \text{w.p. } 1-p \end{cases}$$



$$\begin{aligned} \text{var}(X) &= \sum_x (x - E[X])^2 p_X(x) = (1-p)^2 p + (0-p)^2 \cdot (1-p) \\ &= p - 2p^2 + p^2 + p^2 - p^3 = p - p^2 = p(1-p) \end{aligned}$$

$$\text{var}(X) = E[X^2] - (E[X])^2 = E[X] - (E[X])^2 = p - p^2 = \boxed{p(1-p)}$$

$$X^2 = X$$

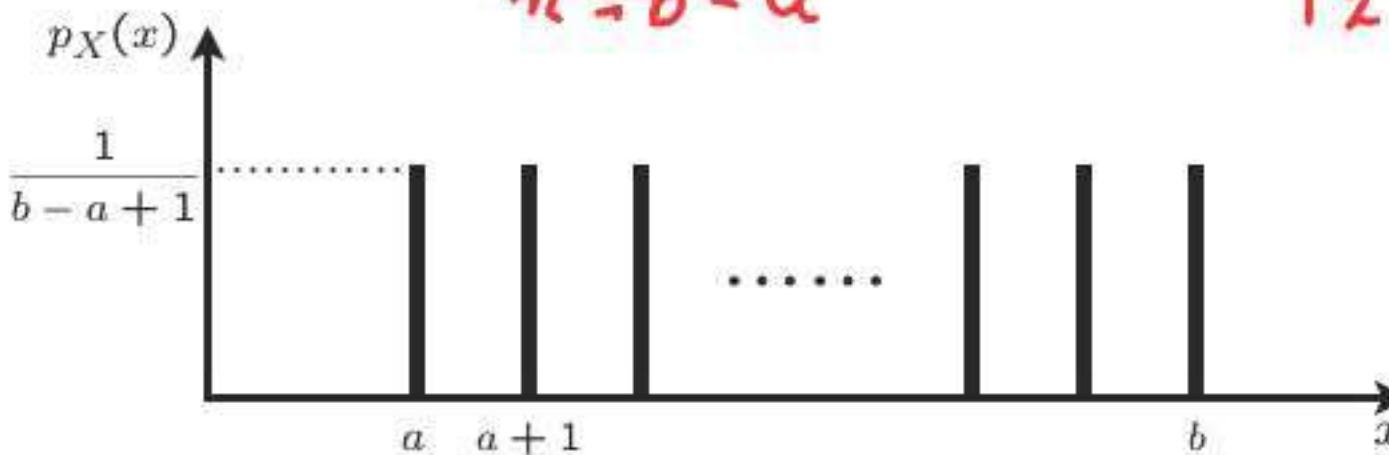
Variance of the uniform



$$\text{var}(x) = E[x^2] - (E[x])^2 = \frac{1}{n+1} (0^2 + 1^2 + 2^2 + \dots + n^2) - \left(\frac{n}{2}\right)^2$$

$$= \frac{1}{12} n(n+1)(2n+1)$$

$n = b - a$



$$\text{Var}(x) = \frac{1}{12} (b-a)(b-a+2)$$

Conditional PMF and expectation, given an event

- Condition on an event $A \Rightarrow$ use conditional probabilities

$$p_X(x) = \underline{\text{P}}(X = x)$$

$$\underline{p_{X|A}(x)} = \underline{\text{P}}(X = x | A)$$

assume
 $\underline{\text{P}}(A) > 0$

$$\sum_x p_X(x) = 1$$

$$\sum_x p_{X|A}(x) = 1$$

$$\mathbf{E}[X] = \sum_x x p_X(x)$$

$$\mathbf{E}[X | A] = \sum_x x p_{X|A}(x)$$

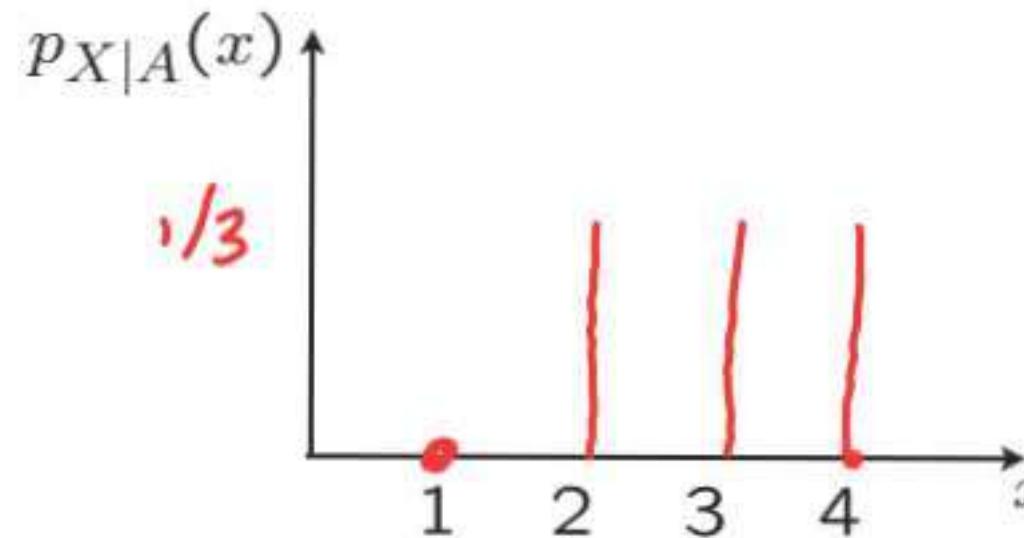
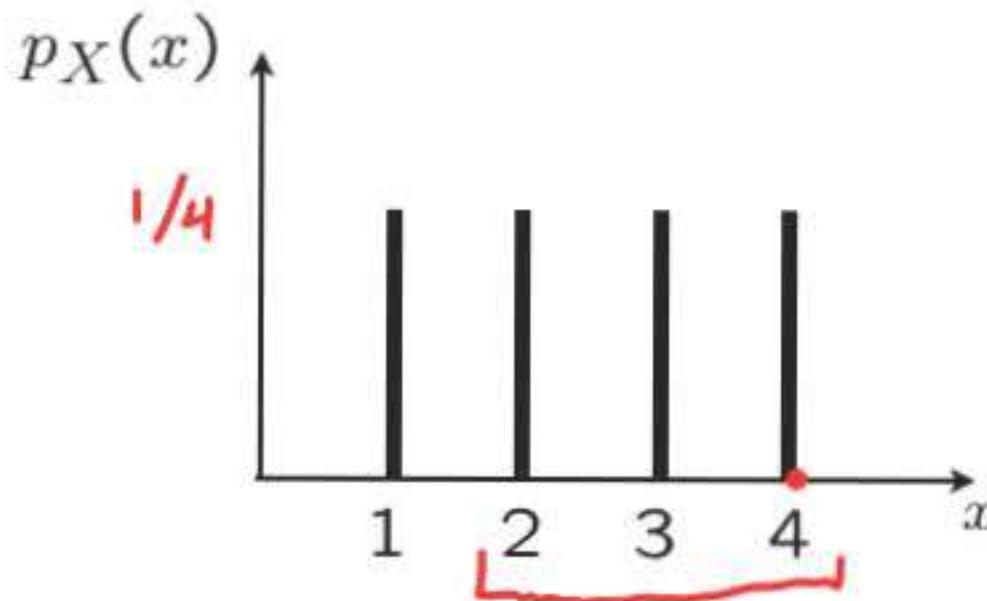
$$\mathbf{E}[g(X)] = \sum_x g(x) p_X(x)$$

$$\mathbf{E}[g(X) | A] = \sum_x g(x) p_{X|A}(x)$$

•

Example of conditioning

- Let $A = \{X \geq 2\}$



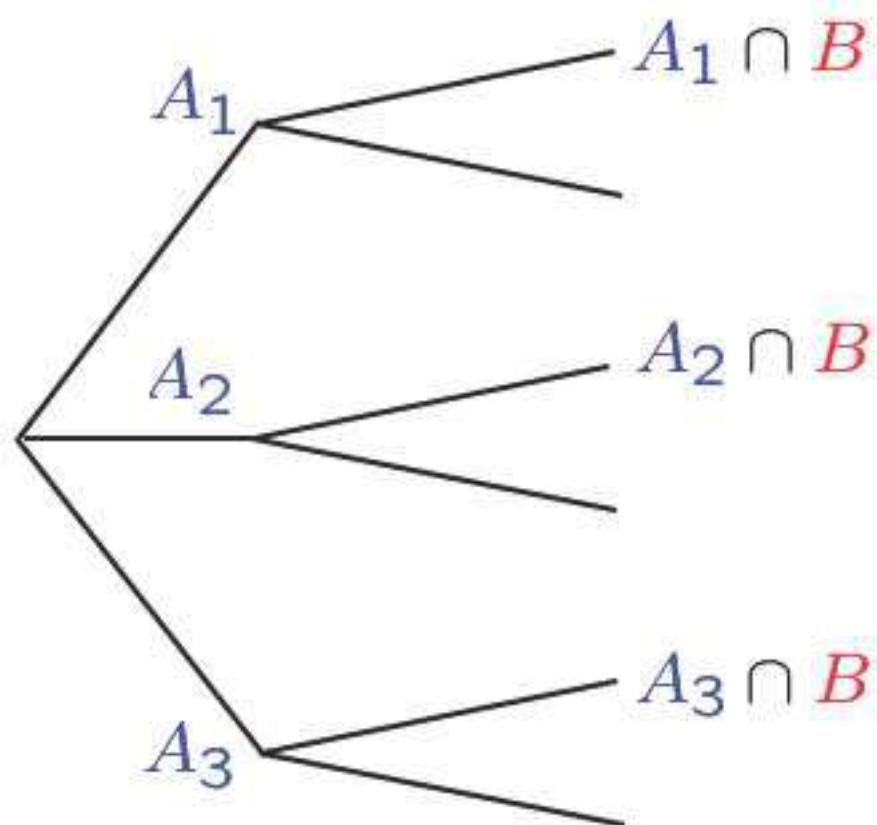
$$E[X] = 2.5$$

$$E[X | A] = 3$$

$$\begin{aligned}\text{var}(X) &= \frac{1}{12}(b-a)(b-a+2) \\ &= \frac{1}{12} 3 \cdot 5 = \frac{5}{4}\end{aligned}$$

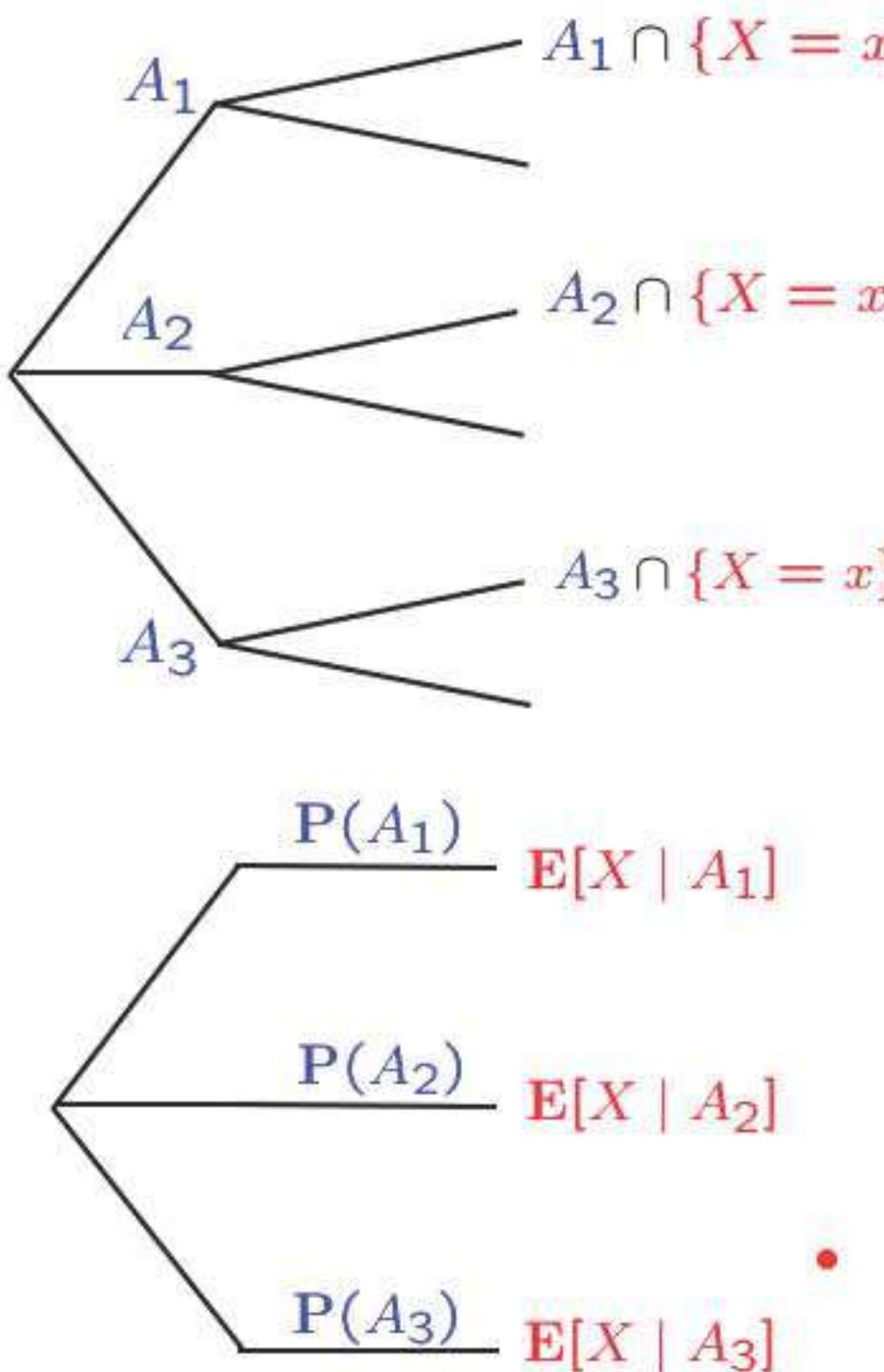
$$\begin{aligned}\text{var}(X | A) &= \frac{1}{3} (4-3)^2 + \frac{1}{3} (3-3)^2 \\ &\quad + \frac{1}{3} (2-3)^2 = \frac{2}{3}\end{aligned}$$

Total expectation theorem



$$P(B) = P(A_1) P(B | A_1) + \cdots + P(A_n) P(B | A_n)$$
$$B = \{x = \alpha\}$$

Total expectation theorem



$$P(B) = P(A_1) P(B | A_1) + \cdots + P(A_n) P(B | A_n)$$

$$B = \{x = z\}$$

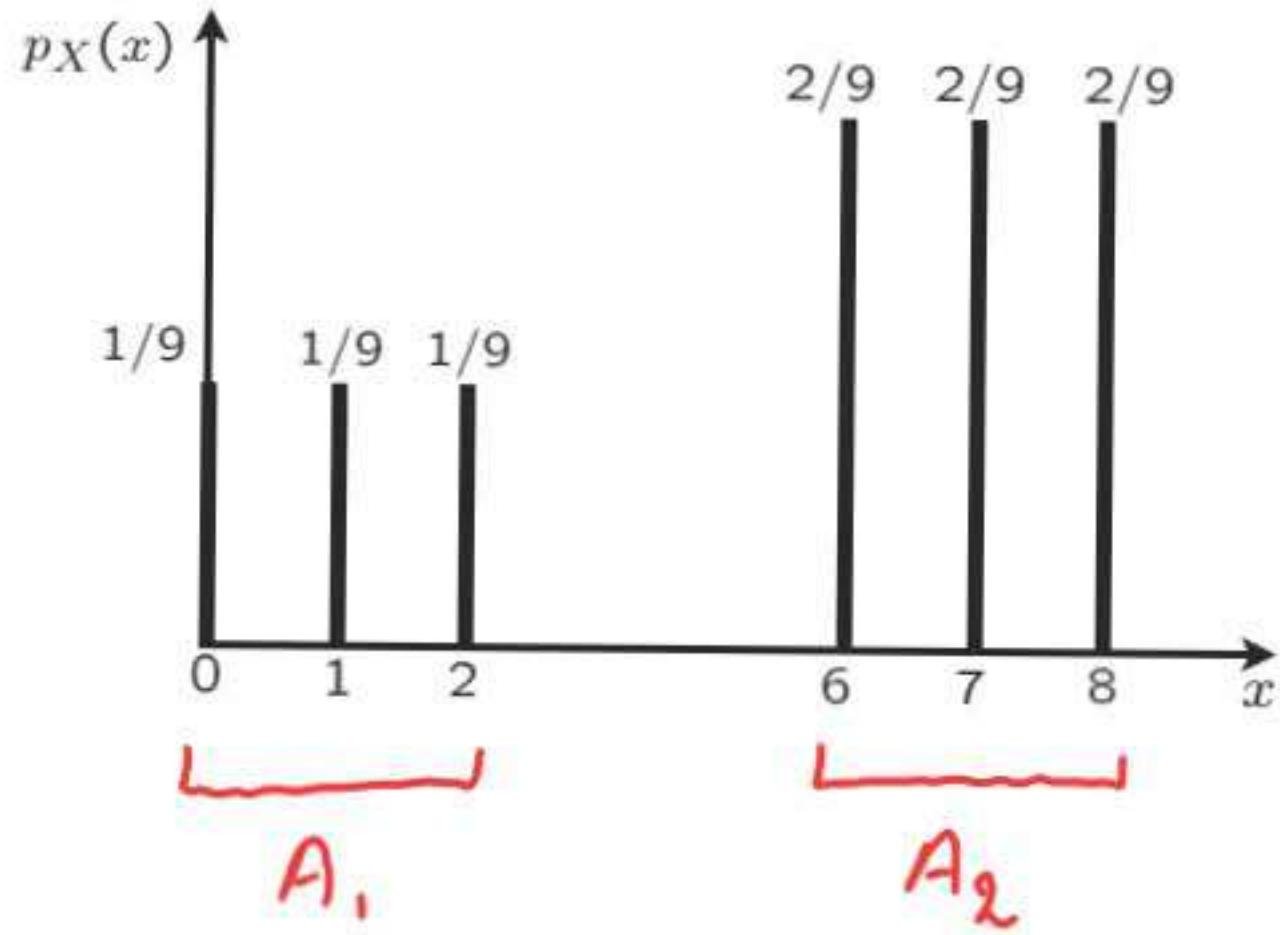
$$p_X(x) = P(A_1) p_{X|A_1}(x) + \cdots + P(A_n) p_{X|A_n}(x)$$

for all x

$$\sum_x x p_X(x) = P(A_1) \underbrace{\sum_x x p_{X|A_1}(x)}_{E[X | A_1]} + \cdots$$

$$E[X] = P(A_1) E[X | A_1] + \cdots + P(A_n) E[X | A_n]$$

Total expectation example



$$P(A_1) = \frac{1}{3}$$

$$P(A_2) = \frac{2}{3}$$

$$E[X|A_1] = 1$$

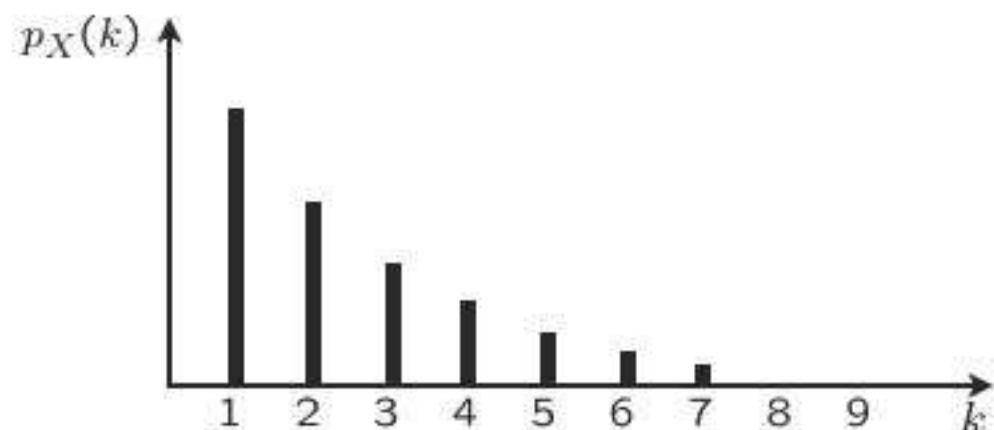
$$E[X|A_2] = 7$$

$$E[X] = \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot 7 .$$

Conditioning a geometric random variable

- X : number of independent coin tosses until first head; $P(H) = p$

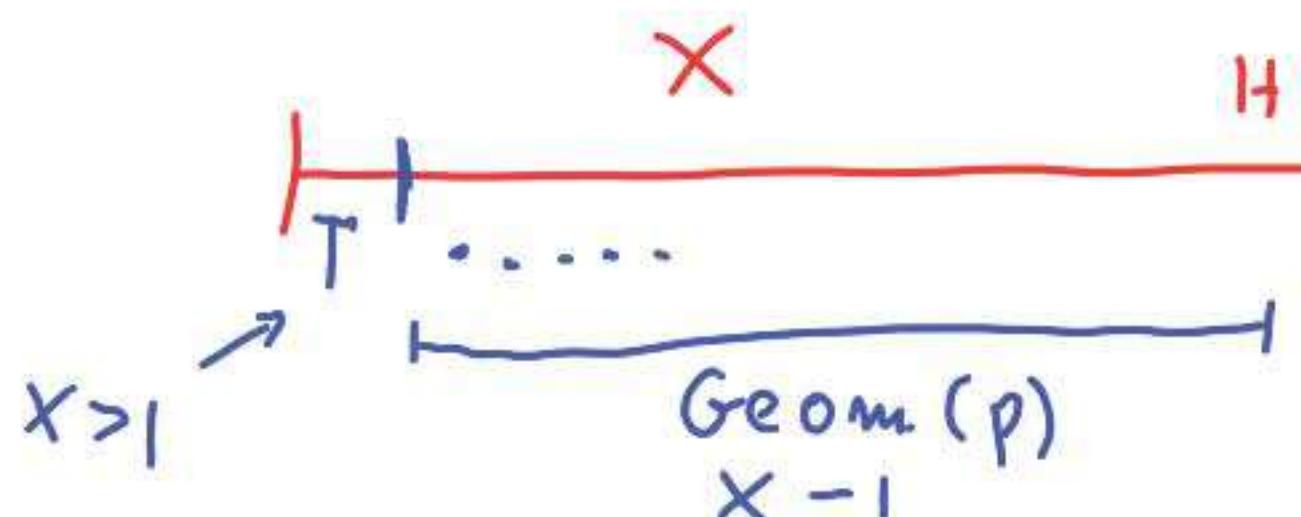
$$p_X(k) = (1 - p)^{k-1} p, \quad k = 1, 2, \dots$$



Memorylessness:

Number of **remaining** coin tosses, conditioned on Tails in the first toss, is **Geometric**, with parameter p

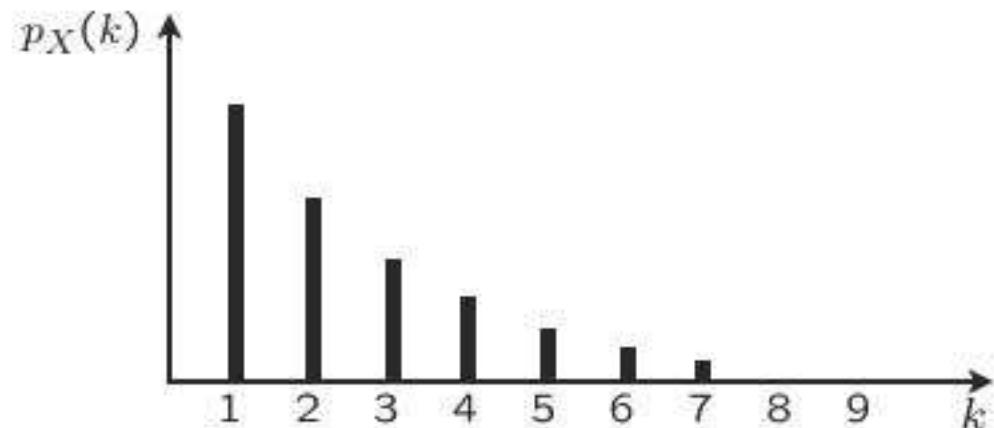
Conditioned on $X > 1$, $X - 1$ is geometric with parameter p



Conditioning a geometric random variable

- X : number of independent coin tosses until first head; $P(H) = p$

$$p_X(k) = (1-p)^{k-1}p, \quad k = 1, 2, \dots$$



Memorylessness:

Number of **remaining** coin tosses, conditioned on Tails in the first toss, is **Geometric**, with parameter p

Conditioned on $X > 1$, $X - 1$ is geometric with parameter p

$$P_{X-1|X>1}(3) = P(X-1=3 | X>1) = P(T_2 T_3 H_4 | \bar{T}_1) = P(T_2 T_3 H_4)$$

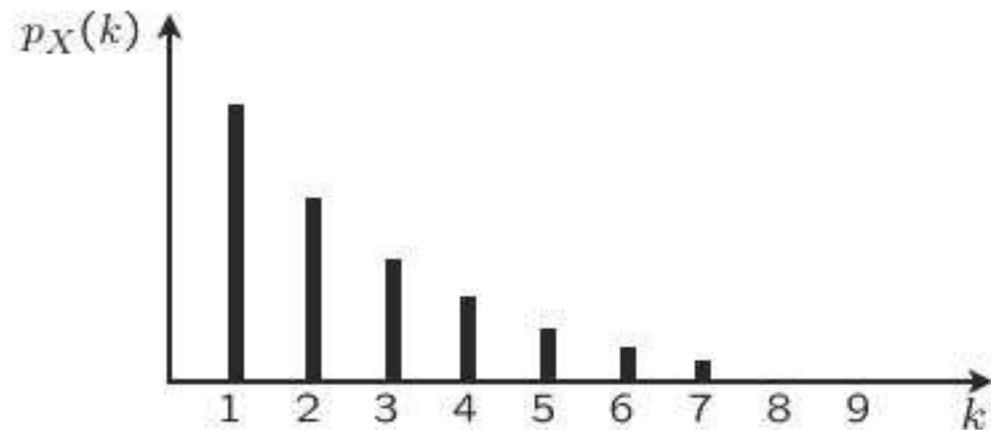
$$= (1-p)^2 p = p_X(3)$$

$$P_{X-1|X>1}(k) = p_X(k)$$

Conditioning a geometric random variable

- X : number of independent coin tosses until first head; $P(H) = p$

$$p_X(k) = (1-p)^{k-1}p, \quad k = 1, 2, \dots$$



Memorylessness:

Number of **remaining** coin tosses, conditioned on Tails in the first toss, is **Geometric**, with parameter p

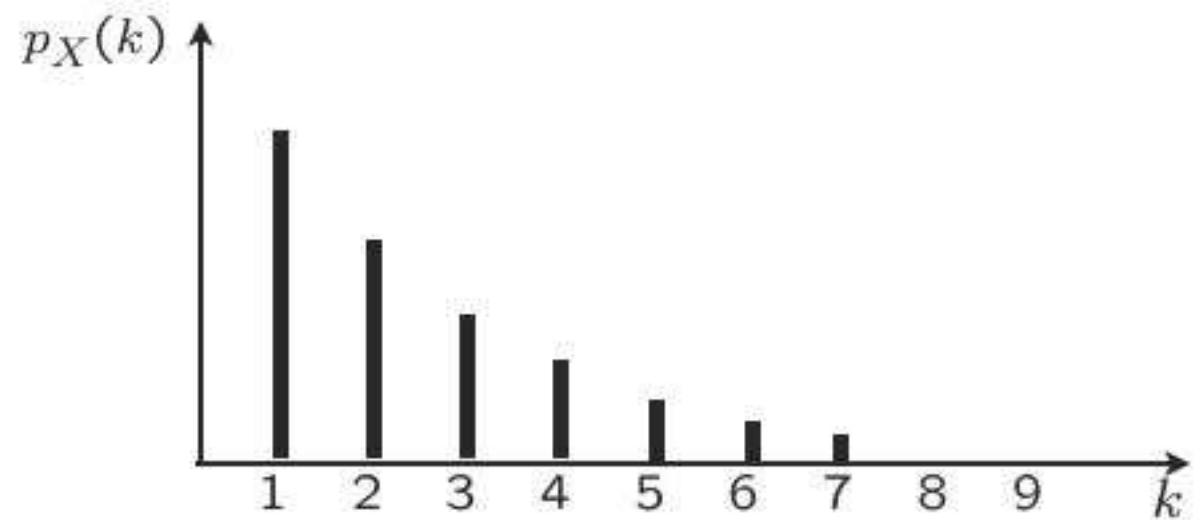
Conditioned on $X > n$, $X - n$ is geometric with parameter p

$$P_{X-1|X>1}(3) = P(X-1=3 | X>1) = P(T_2 T_3 H_4 | \bar{T}_1) = P(T_2 T_3 H_4)$$

$$= (1-p)^2 p = p_X(3)$$

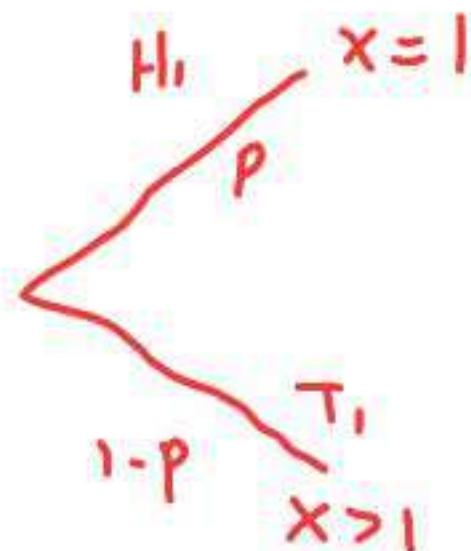
$$P_{X-1|X>1}(k) = P_X(k) = P_{X-n|X>n}(k)$$

The mean of the geometric



$$E[X] = \sum_{k=1}^{\infty} kp_X(k) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p$$

$$E[X] = \frac{1}{p}$$



$$\begin{aligned} E[x] &= 1 + E[x-1] \\ &= 1 + p \cdot E[x-1 | x=1] + (1-p)E[x-1 | x>1] \\ &= 1 + 0 + (1-p)E[x] \end{aligned}$$

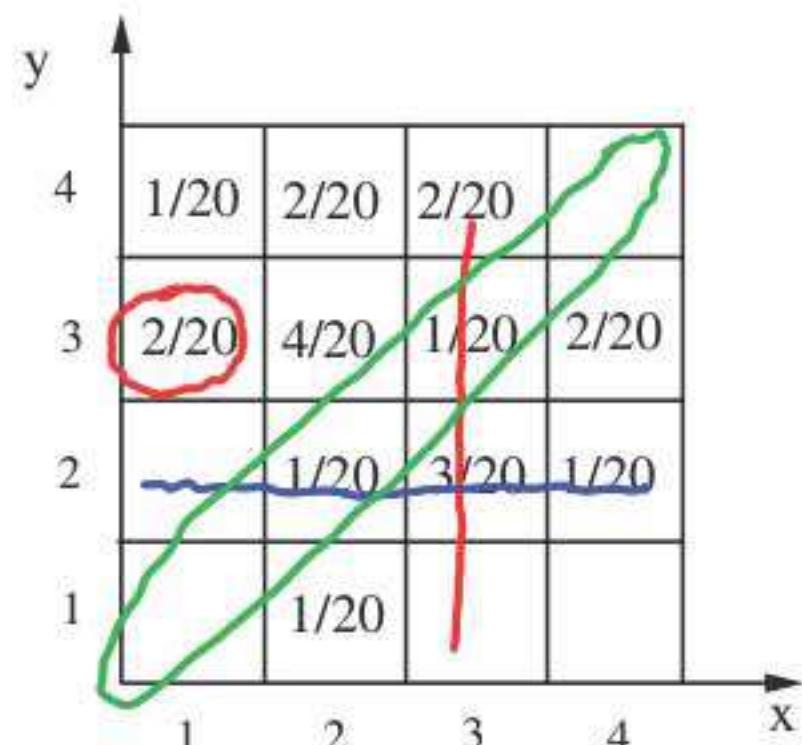
Multiple random variables and joint PMFs

marginal pmfs

$$X : p_X$$

$$Y : p_Y \quad P(X = Y) = \frac{2}{20}$$

Joint PMF: $p_{X,Y}(x,y) = P(X = x \text{ and } Y = y)$



$$P_{X,Y}(1,3) = \frac{2}{20}$$

$$p_X(4) = \frac{1}{20} + \frac{2}{20}$$

$$p_Y(2) = \frac{1}{20} + \frac{3}{20} + \frac{1}{20}$$

$$\sum_x \sum_y p_{X,Y}(x,y) = 1$$

$$p_X(x) = \sum_y p_{X,Y}(x,y)$$

$$p_Y(y) = \sum_x p_{X,Y}(x,y)$$

More than two random variables

$$p_{X,Y,Z}(x, y, z) = \mathbf{P}(X = x \text{ and } Y = y \text{ and } Z = z)$$

$$\sum_x \sum_y \sum_z p_{X,Y,Z}(x, y, z) = 1$$

$$p_X(x) = \sum_y \sum_z p_{X,Y,Z}(x, \underline{y}, \underline{z})$$

$$p_{X,Y}(x, y) = \sum_z p_{X,Y,Z}(x, y, \underline{z})$$

Functions of multiple random variables

$$Z = g(X, Y)$$

PMF: $p_Z(z) = \mathbf{P}(Z = z) = \mathbf{P}(g(X, Y) = z) = \sum_{(x, y) : g(x, y) = z} p_{X,Y}(x, y)$

Expected value rule: $\mathbf{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y)$

$$\mathbf{E}[g(x)]$$

Linearity of expectations

$$E[aX + b] = aE[X] + b$$

$$E[X + Y] = E[X] + E[Y]$$

$$E[X + Y] = E[g(X, Y)]$$

$$(g(x, y) = x + y)$$

$$= \sum_x \sum_y (x + y) P_{X,Y}(x, y)$$

$$= \underbrace{\sum_x \sum_y x P_{X,Y}(x, y)} + \underbrace{\sum_x \sum_y y P_{X,Y}(x, y)}$$

$$= \underbrace{\sum_x x \sum_y P_{X,Y}(x, y)} + \underbrace{\dots}$$

$$= \sum_x x P_X(x) + \sum_y y P_Y(y) = E[X] + E[Y]$$

Linearity of expectations

$$\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$$

$$\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$$

$$\mathbf{E}[X_1 + \dots + X_n] = \mathbf{E}[X_1] + \dots + \mathbf{E}[X_n]$$

$$\mathbf{E}[2X + 3Y - Z] = E[2x] + E[3y] - E[z] = 2E[x] + 3E[y] - E[z]$$

The mean of the binomial

- X : binomial with parameters n, p
 - number of successes in n independent trials

$X_i = 1$ if i th trial is a success;
 $X_i = 0$ otherwise

(indicator variable)

$$X = X_1 + \cdots + X_n$$

$$E[X] = \underbrace{E[X_1]}_p + \cdots + \underbrace{E[X_n]}_p = np$$

$$E[X] = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k}$$

$$\boxed{P_X(k)}$$
$$E[X] = np$$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 7: Conditioning on a random variable; Independence of r.v.'s

- Conditional PMFs
 - Conditional expectations
 - Total expectation theorem
- Independence of r.v.'s
 - Expectation properties
 - Variance properties
- The variance of the binomial
- The hat problem: mean and variance

Conditional PMFs

$$A = \{Y = y\}$$

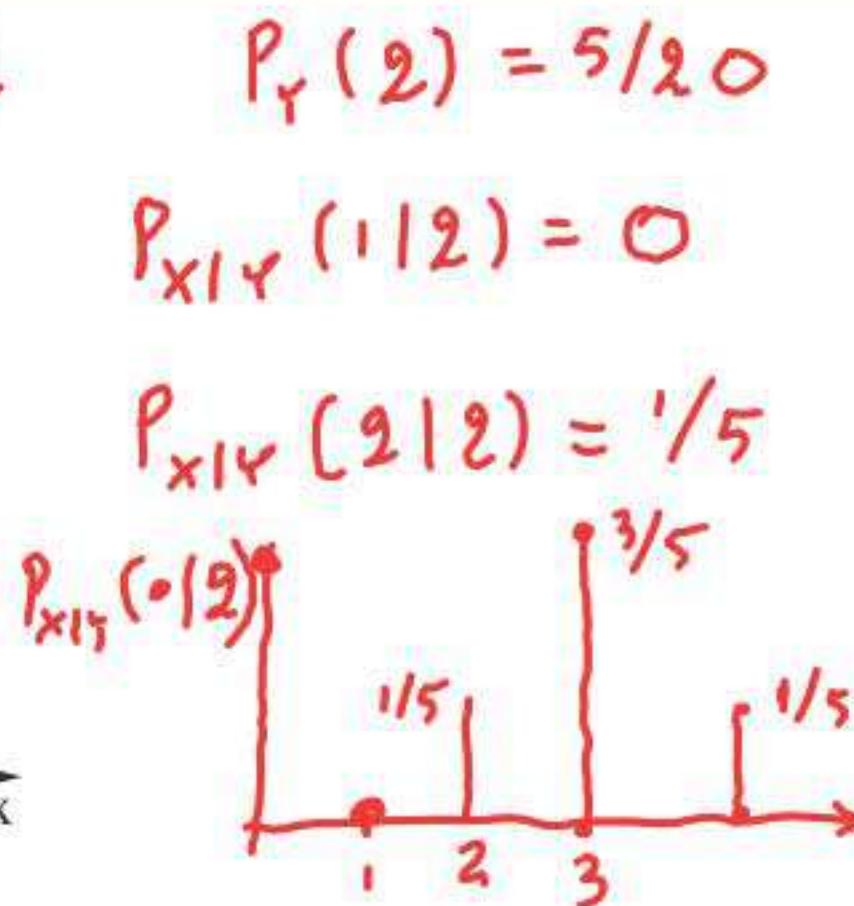
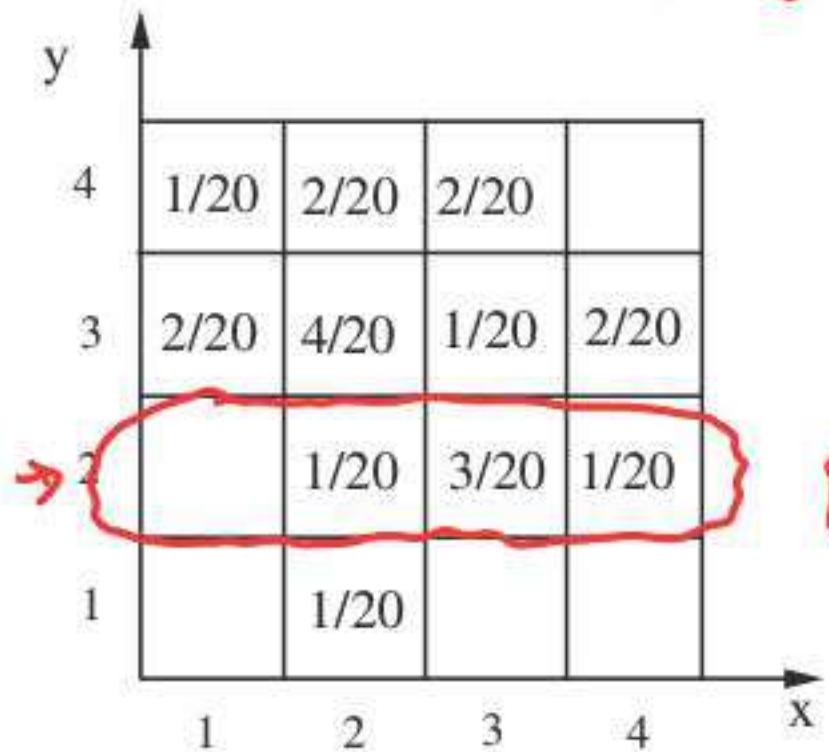
$$p_{X|A}(x | A) = P(X = x | A)$$

$$\underline{p_{X|Y}(x | y)} = P(X = x | Y = y) = \frac{P(x = x, Y = y)}{P(Y = y)}$$

$$p_{X|Y}(x | y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

defined for y such that $p_Y(y) > 0$

$$\sum_x p_{X|Y}(x | y) = 1$$



$$p_{X,Y}(x, y) = p_Y(y) p_{X|Y}(x | y)$$

$$p_{X,Y}(x, y) = p_X(x) p_{Y|X}(y | x)$$

Conditional PMFs involving more than two r.v.'s

- Self-explanatory notation

$$p_{X|Y,Z}(x | y, z) = \frac{P(X=x | Y=y, Z=z)}{P(Y=y, Z=z)} = \frac{P_{x,y,z}(x, y, z)}{P_{y,z}(y, z)}$$

$$p_{X,Y|Z}(x, y | z) = P(X=x, Y=y | Z=z)$$

- Multiplication rule

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

$$A = \{X=x\} \quad B = \{Y=y\} \quad C = \{Z=z\}$$

$$p_{X,Y,Z}(x, y, z) = p_X(x) p_{Y|X}(y | x) p_{Z|X,Y}(z | x, y)$$

Conditional expectation

$$A = \{Y = y\}$$

$$\mathbb{E}[X] = \sum_x x p_X(x)$$

$$\mathbb{E}[X | A] = \sum_x x p_{X|A}(x)$$

$$\mathbb{E}[X | Y = y] = \sum_x x p_{X|Y}(x | y)$$

- Expected value rule

$$\mathbb{E}[g(X)] = \sum_x g(x) p_X(x) \quad \mathbb{E}[g(X) | A] = \sum_x g(x) p_{X|A}(x)$$

$$\mathbb{E}[g(X) | Y = y] = \sum_x g(x) p_{X|Y}(x | y)$$

Total probability and expectation theorems

- A_1, \dots, A_n : partition of Ω

$$Y = \{y_1, \dots, y_n\} \quad A_i = \{Y = y_i\}$$

- $p_X(x) = P(A_1)p_{X|A_1}(x) + \dots + P(A_n)p_{X|A_n}(x)$

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x | y)$$

- $E[X] = P(A_1)E[X | A_1] + \dots + P(A_n)E[X | A_n]$

$$E[X] = \sum_y p_Y(y) E[X | Y = y]$$

•

- Fine print:

Also valid when Y is a discrete r.v. that ranges over an infinite set,
as long as $E[|X|] < \infty$

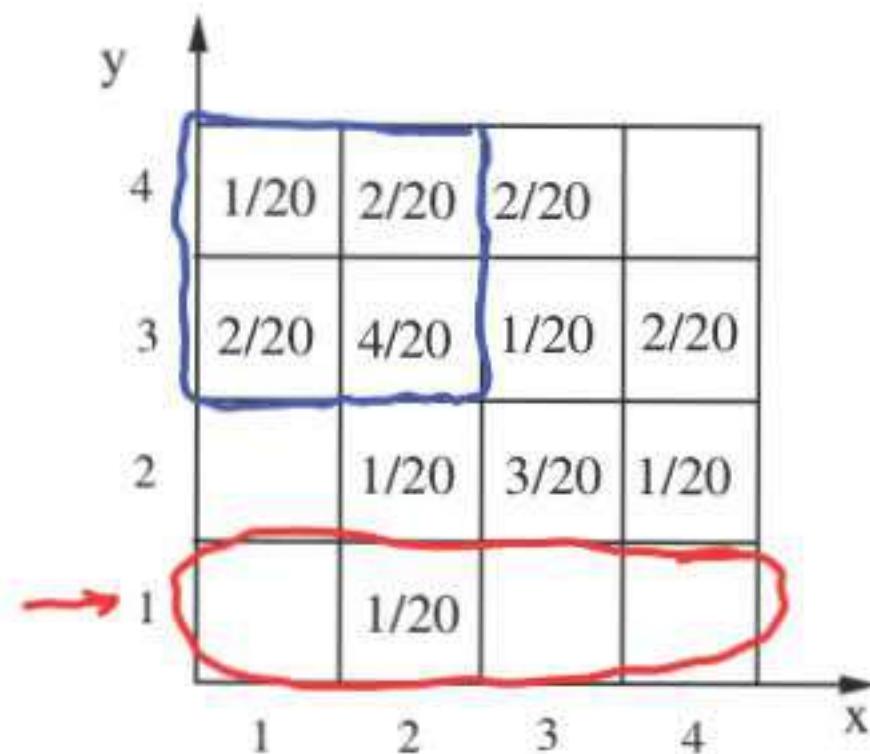
Independence

- of two events: $P(A \cap B) = P(A) \cdot P(B)$ $P(A | B) = P(A)$
- of a r.v. and an event: $P(\underline{X = x} \text{ and } \underline{A}) = P(X = x) \cdot P(A), \text{ for all } \underline{\underline{x}}$
 $p_{x|A}(x) = p_x(x), \text{ for all } x$ $P(A | X = x) = P(A), \text{ for all } x$
- of two r.v.'s: $P(\underline{X = x} \text{ and } \underline{Y = y}) = P(X = x) \cdot P(Y = y), \text{ for all } \underline{\underline{x, y}}$
 $p_{x|y}(x|y) = p_x(x)$ $p_{X,Y}(x, y) = p_X(x)p_Y(y), \text{ for all } x, y$
 $p_{y|x}(y|x) = p_y(y)$

X, Y, Z are **independent** if:

$$p_{X,Y,Z}(x, y, z) = p_X(x)p_Y(y)p_Z(z), \text{ for all } x, y, z^*$$

Example: independence and conditional independence



- Independent? *No*

$$P_X(1) = 3/20$$

$$P_{X|Y}(1|1) = 0$$

- What if we condition on $X \leq 2$ and $Y \geq 3$?

1/9	2/9
2/9	4/9

Yes

.

Independence and expectations

- In general: $E[g(X, Y)] \neq g(E[X], E[Y])$

always true

- Exceptions: $E[aX + b] = aE[X] + b$

$$E[X + Y + Z] = E[X] + E[Y] + E[Z]$$

If X, Y are independent: $E[XY] = E[X]E[Y]$

$g(X)$ and $h(Y)$ are also independent: $E[g(X)h(Y)] = E[g(X)] \cdot E[h(Y)]$

$$E[g(x, y)] \quad g(x, y) = xy$$

$$= \sum_x \sum_y xy P_{x,y}(x, y) = \sum_x \sum_y \underbrace{xy}_{\text{factored}} P_x(x) P_y(y)$$

$$= \sum_x x P_x(x) \underbrace{\sum_y y P_y(y)}_{\text{constant}} = E[x] E[y]$$

Independence and variances

- Always true: $\text{var}(aX) = a^2 \text{var}(X)$ $\text{var}(X + a) = \text{var}(X)$
- In general: $\text{var}(X + Y) \neq \text{var}(X) + \text{var}(Y)$

If X, Y are independent: $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$

assume
 $E[X] = E[Y] = 0$

$$\begin{aligned}\text{var}(X+Y) &= E[(X+Y)^2] = E[X^2 + 2XY + Y^2] \\ &= E[X^2] + 2E[XY] + E[Y^2] = \text{var}(X) + \text{var}(Y)\end{aligned}$$

- Examples:

- If $X = Y$: $\text{var}(X + Y) = \text{var}(2X) = 4\text{var}(X)$

- If $X = -Y$: $\text{var}(X + Y) = \text{var}(0) = 0$

- If X, Y independent: $\text{var}(X - 3Y) = \text{var}(X) + \text{var}(-3Y) = \text{var}(X) + 9\text{var}(Y)$

Variance of the binomial

- X : binomial with parameters n, p
 - number of successes in n independent trials

$$\begin{aligned} X_i &= 1 \text{ if } i\text{th trial is a success;} \\ X_i &= 0 \text{ otherwise} \end{aligned} \quad (\text{indicator variable})$$

independent

$$X = X_1 + \cdots + X_n$$

$$\text{var}(x) = \text{var}(x_1) + \dots + \text{var}(x_n)$$

$$= n \cdot \text{var}(X_1) = \boxed{np(1-p)}$$

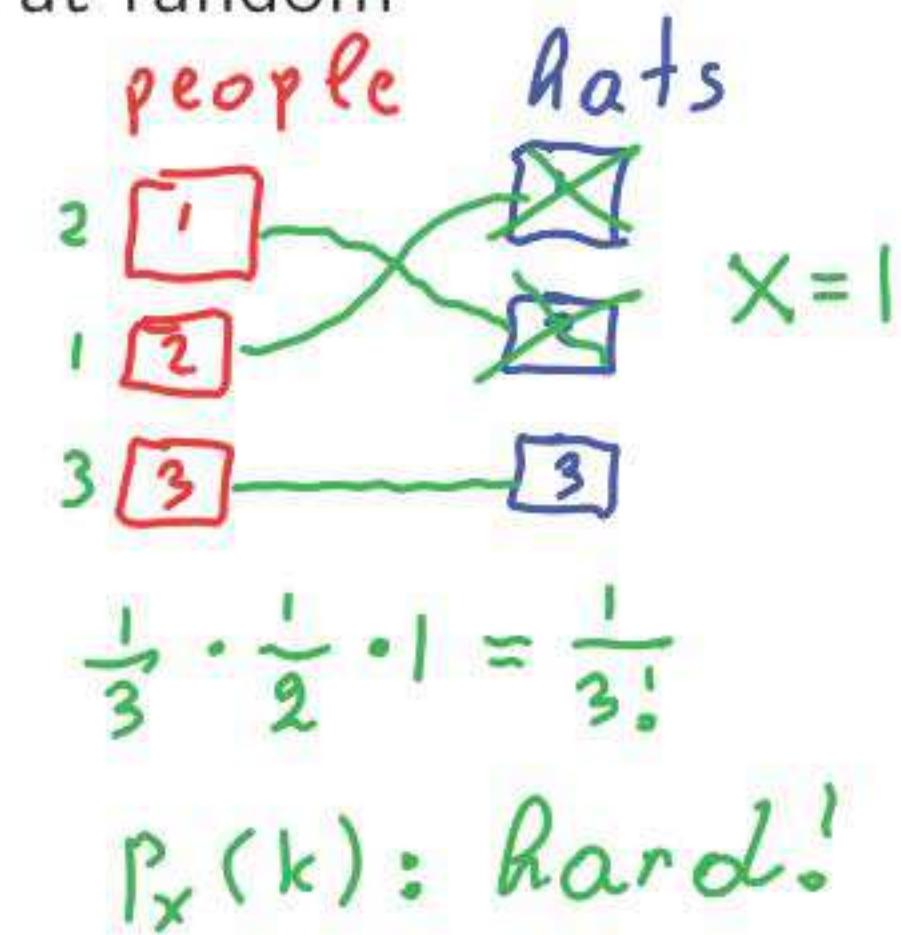
The hat problem

- n people throw their hats in a box and then pick one at random
 - All permutations equally likely $1/n!$
 - Equivalent to picking one hat at a time
- X : number of people who get their own hat
 - Find $E[X] = E[X_1] + \dots + E[X_n] = n \cdot \frac{1}{n} = 1$

$$X_i = \begin{cases} 1, & \text{if } i \text{ selects own hat} \\ 0, & \text{otherwise.} \end{cases}$$

$$X = X_1 + X_2 + \dots + X_n$$

$$\bullet E[X_i] = E[X_1] = P(X_1 = 1) = \frac{1}{n}$$



$$\sum_k k p_X(k)$$

The variance in the hat problem

- X : number of people who get their own hat
 - Find $\text{var}(X)$

$$X_i = \begin{cases} 1, & \text{if } i \text{ selects own hat} \\ 0, & \text{otherwise.} \end{cases}$$

- $\boxed{\text{var}(X)} = E[X^2] - (E[X])^2 = 2 - 1 = \boxed{1}$

- $E[X_i^2] = E[X_1^2] = E[X_1] = 1/n$

- For $i \neq j$: $E[X_i X_j] = E[X_1 X_2] = P(X_1, X_2 = 1) = P(X_1 = 1, X_2 = 1)$

$$= P(X_1 = 1) P(X_2 = 1 | X_1 = 1) = \frac{1}{n} \cdot \frac{1}{n-1}$$

$$n = 2$$

$$X_1 = 1 \Rightarrow X_2 = 1$$

$$X_1 = 0 \Rightarrow X_2 = 0$$

$$X = X_1 + X_2 + \cdots + X_n$$

$$n(n-1)$$

$$n^2 - n$$

$$X^2 = \underbrace{\sum_i X_i^2}_{n} + \underbrace{\sum_{i,j: i \neq j} X_i X_j}_{n^2 - n}.$$

$$E[X^2] = n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n} \cdot \frac{1}{n-1}$$

$$E[X^2] = n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n} \cdot \frac{1}{n-1}$$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

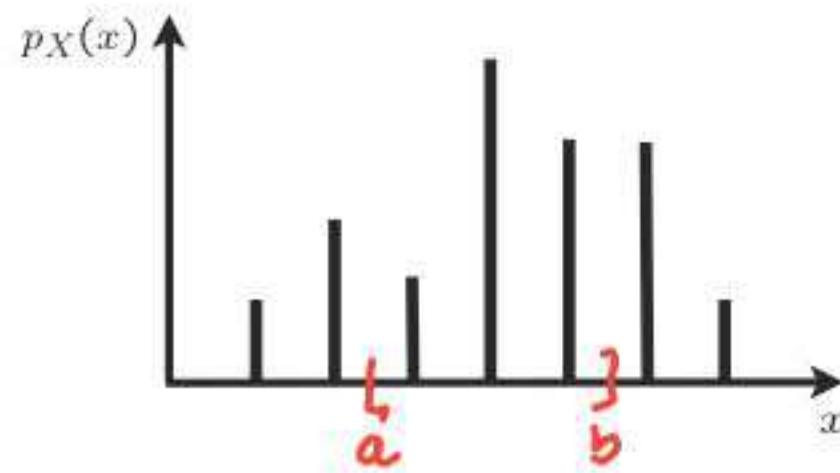
The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

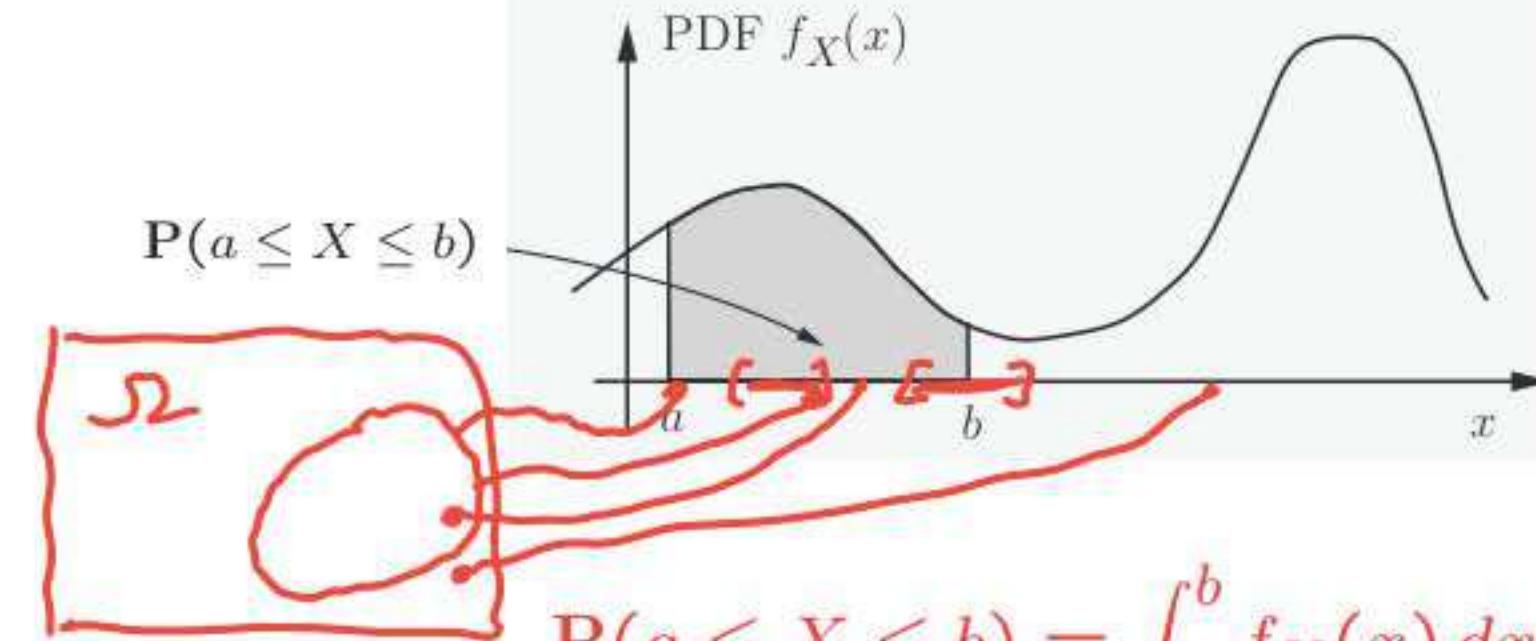
LECTURE 8: Continuous random variables and probability density functions

- Probability density functions
 - Properties
 - Examples
- Expectation and its properties
 - The expected value rule
 - Linearity
- Variance and its properties
- Uniform and exponential random variables
- Cumulative distribution functions
- Normal random variables
 - Expectation and variance
 - Linearity properties
 - Using tables to calculate probabilities

Probability density functions (PDFs)



$$P(a \leq X \leq b) = \sum_{x: a \leq x \leq b} p_X(x)$$



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

$$p_X(x) \geq 0$$

$$\sum_x p_X(x) = 1$$

$$f_X(x) \geq 0$$

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

Definition: A random variable is **continuous if it can be described by a PDF**

$$P(1 \leq X \leq 3 \text{ or } 4 \leq X \leq 5) = P(1 \leq X \leq 3) + P(4 \leq X \leq 5)$$

Probability density functions (PDFs)

$\delta > 0$, small

$$P(a \leq X \leq a + \delta)$$

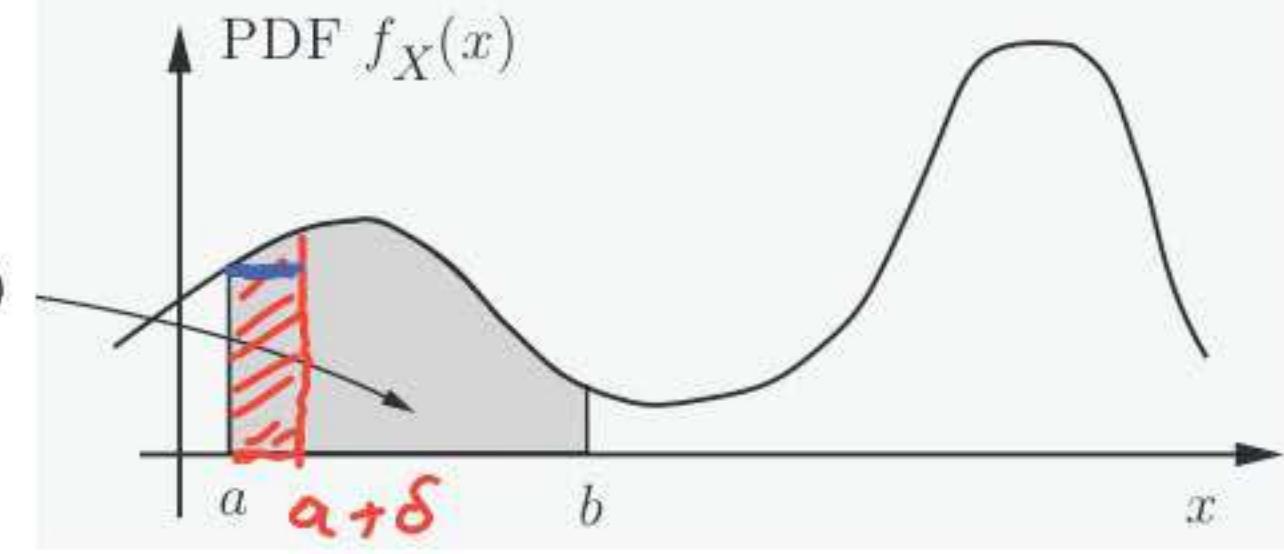
$$\approx f_X(a) \cdot \delta$$

$$P(a \leq X \leq a + \delta) \approx f_X(a) \cdot \delta$$

$$P(X = a) = 0$$

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

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

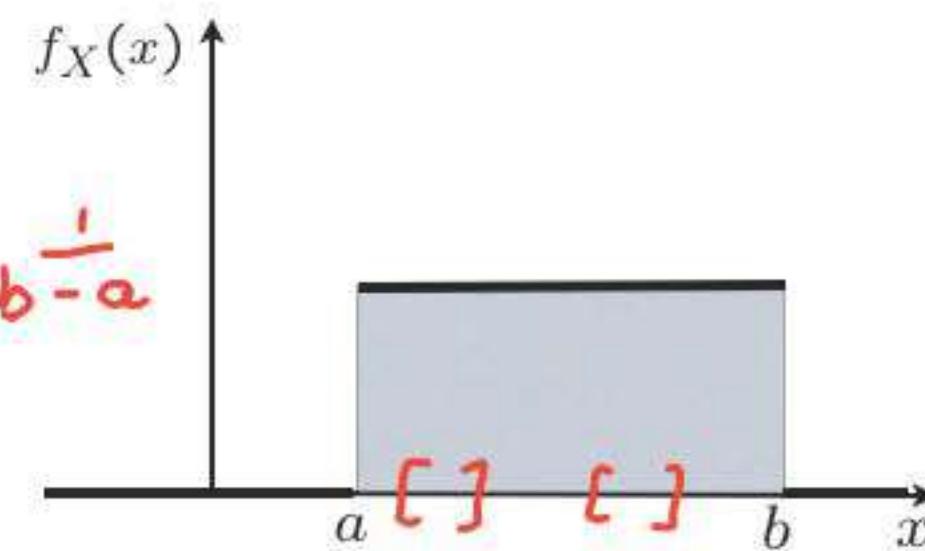
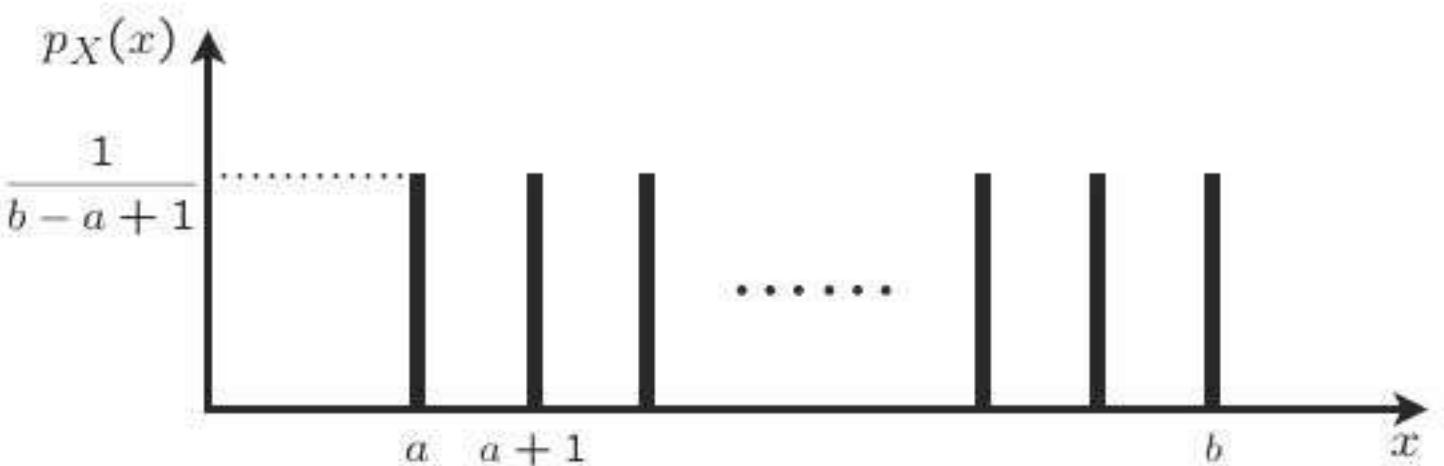


$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

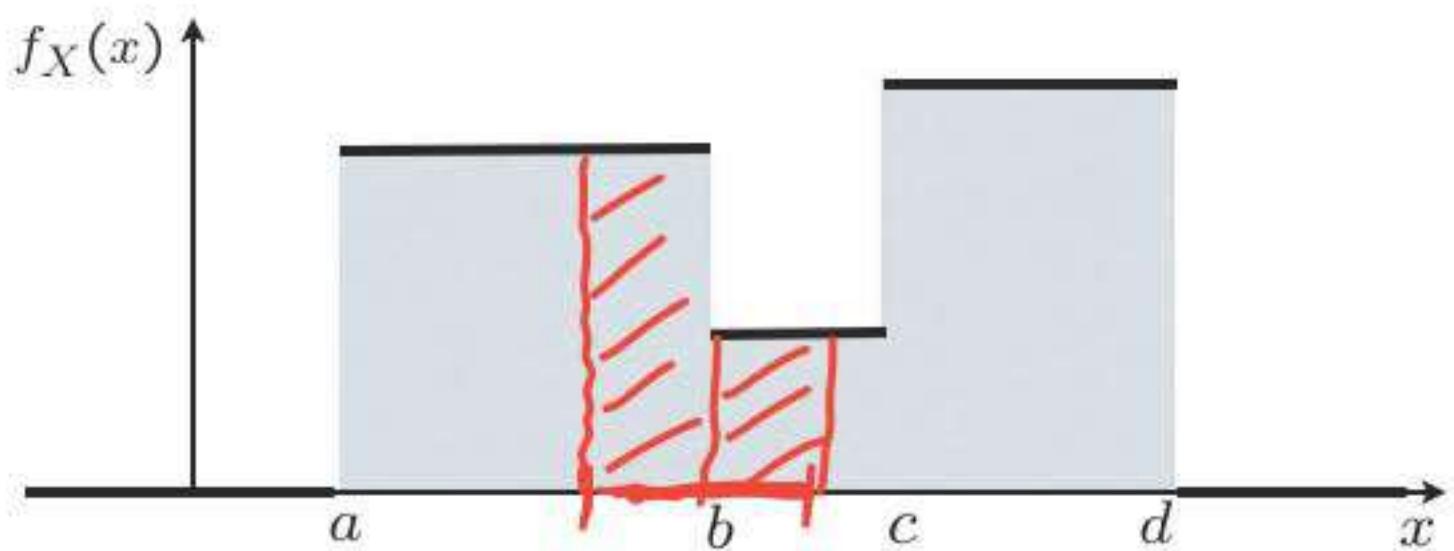
$$f_X(x) \geq 0$$

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

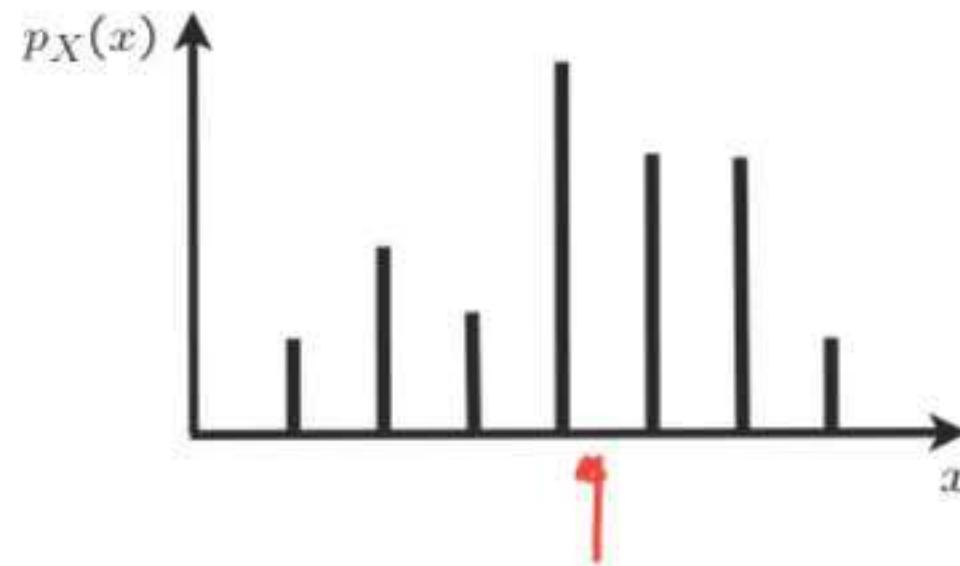
Example: continuous uniform PDF



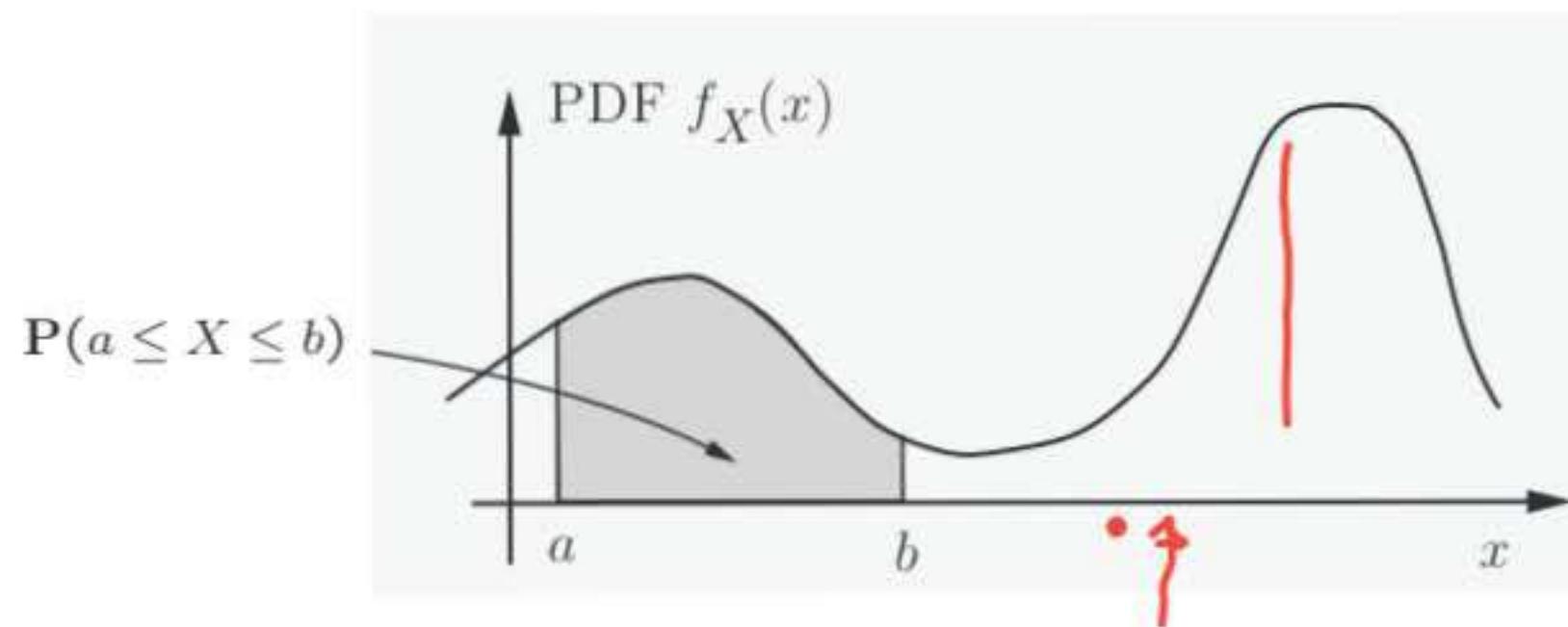
- Generalization: piecewise constant PDF



Expectation/mean of a continuous random variable



$$E[X] = \sum_x x p_X(x)$$



$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

- **Interpretation:** Average in large number of independent repetitions of the experiment

Fine print:
Assume $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$

Properties of expectations

- If $X \geq 0$, then $E[X] \geq 0$
- If $a \leq X \leq b$, then $a \leq E[X] \leq b$
- Expected value rule:

$$E[g(X)] = \sum_x g(x)p_X(x)$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x) dx$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx$$

- Linearity

$$E[aX + b] = aE[X] + b$$

Variance and its properties

- Definition of variance: $\text{var}(X) = E[(X - \mu)^2]$

$$\mu = E[X]$$

- Calculation using the expected value rule, $E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x) dx$

$$\text{var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx$$

$$g(x) = (x - \mu)^2$$

Standard deviation: $\sigma_X = \sqrt{\text{var}(X)}$

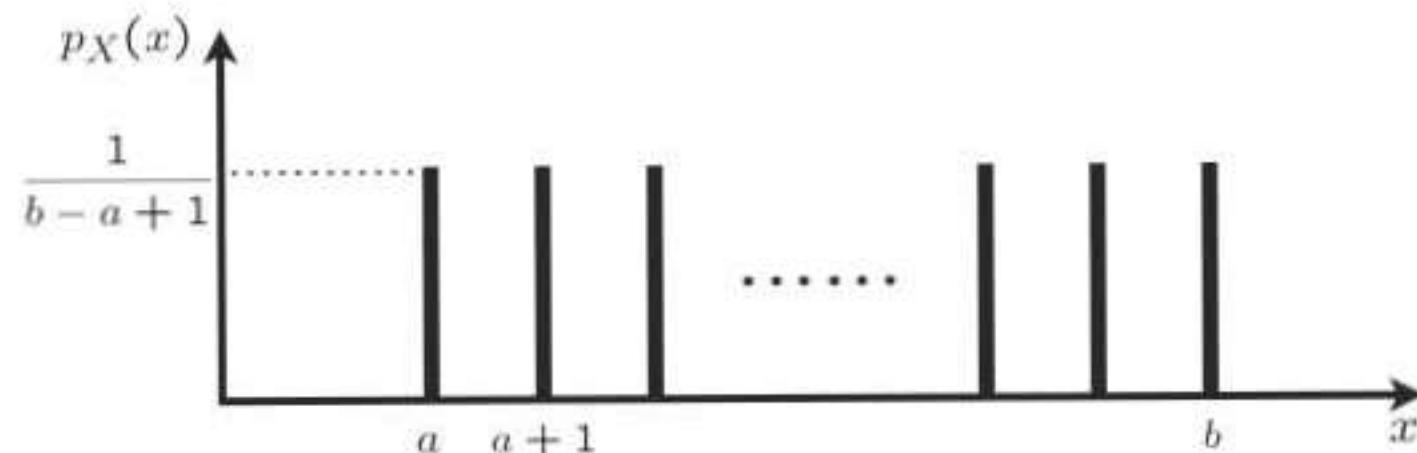
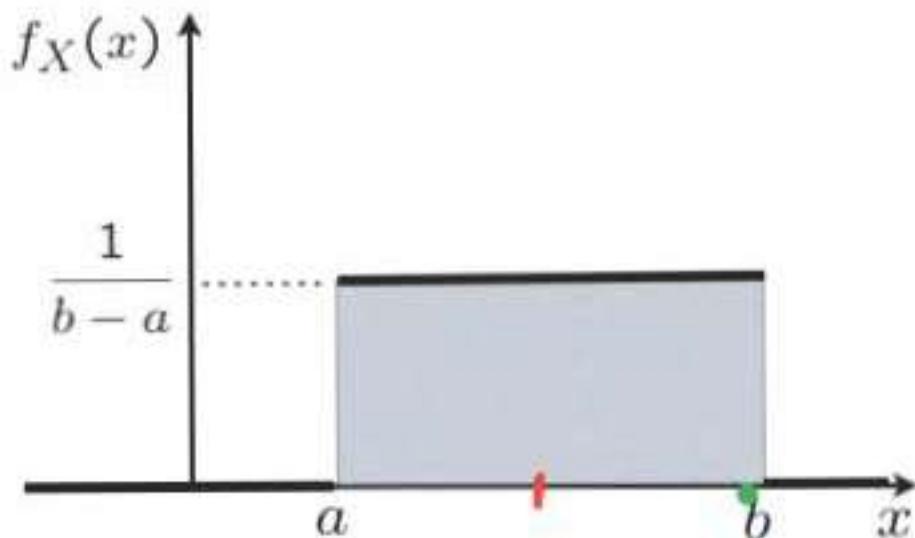


$$\text{var}(aX + b) = a^2\text{var}(X)$$



$$\text{var}(X) = E[X^2] - (E[X])^2$$

Continuous uniform random variable; parameters a, b



$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_a^b x \cdot \frac{1}{b-a} dx = \frac{a+b}{2} \end{aligned}$$

$$E[X^2] = \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \left(\frac{b^3}{3} - \frac{a^3}{3} \right)$$

$$\text{var}(X) = E[X^2] - (E[X])^2 = \boxed{\frac{(b-a)^2}{12}} \quad \sigma = \frac{b-a}{\sqrt{12}}$$

$$E[X] = \frac{a+b}{2}$$

$$\text{var}(X) = \frac{1}{12}(b-a)(b-a+2)$$

$$\sigma = \frac{b-a}{\sqrt{12}}$$

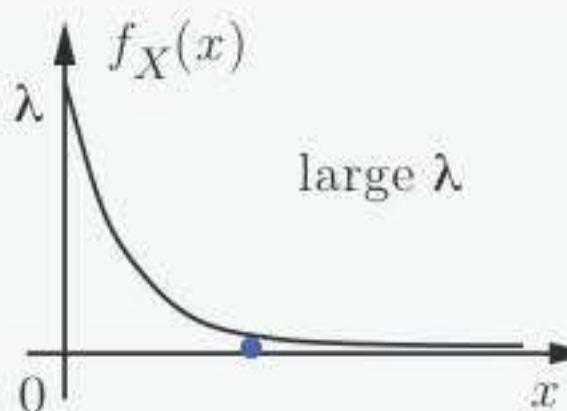
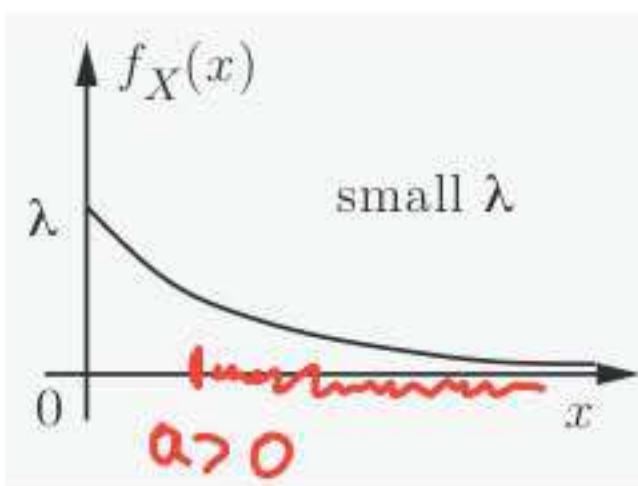
Exponential random variable; parameter $\lambda > 0$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$\int f_X(x) dx = 1$$

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

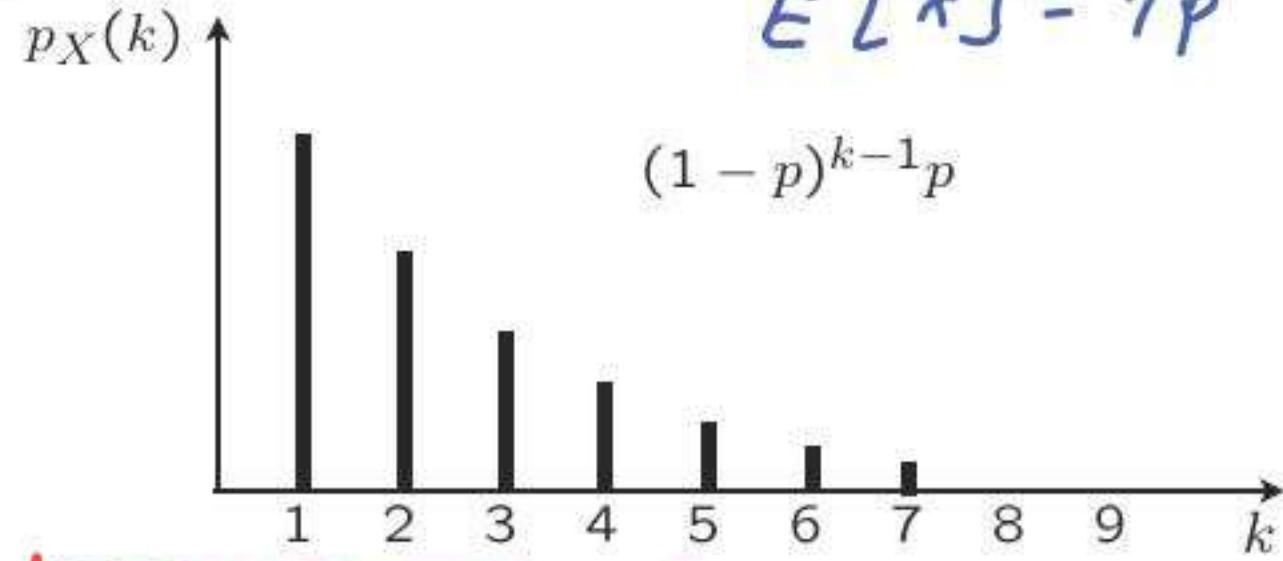
$$(1 - p)^{k-1} p$$



$$E[X] = \int_0^\infty x \cdot \lambda e^{-\lambda x} dx = 1/\lambda$$

$$E[X^2] = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = 2/\lambda^2$$

$$\text{var}(X) = E[X^2] - (E[X])^2 = 1/\lambda^2$$



$$\boxed{P(X \geq a)} = \int_a^\infty \lambda e^{-\lambda x} dx$$

$$\left[\int e^{ax} dx = \frac{1}{a} e^{ax} \quad a \leftrightarrow -\lambda \right]$$

$$= \lambda \cdot \left(-\frac{1}{\lambda} \right) e^{-\lambda x} \Big|_a^\infty$$

$$= -e^{-\lambda \cdot 0} + e^{-\lambda a} = \boxed{e^{-\lambda a}}$$

Cumulative distribution function (CDF)

CDF definition: $F_X(x) = P(X \leq x)$

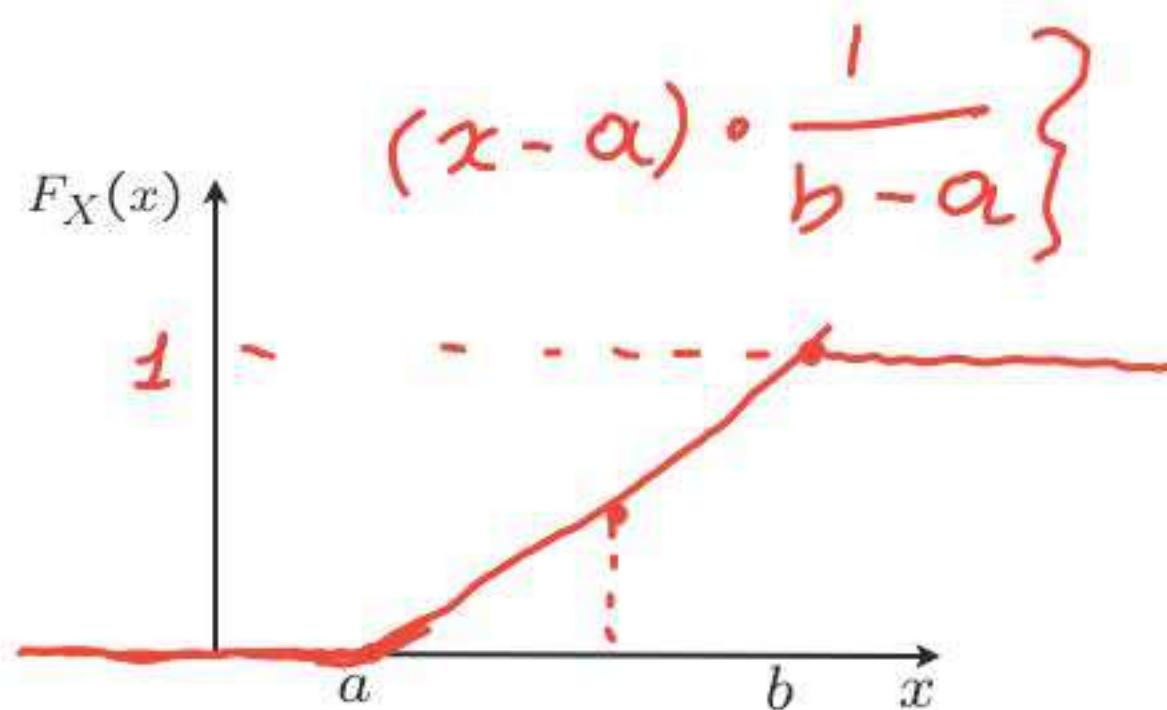
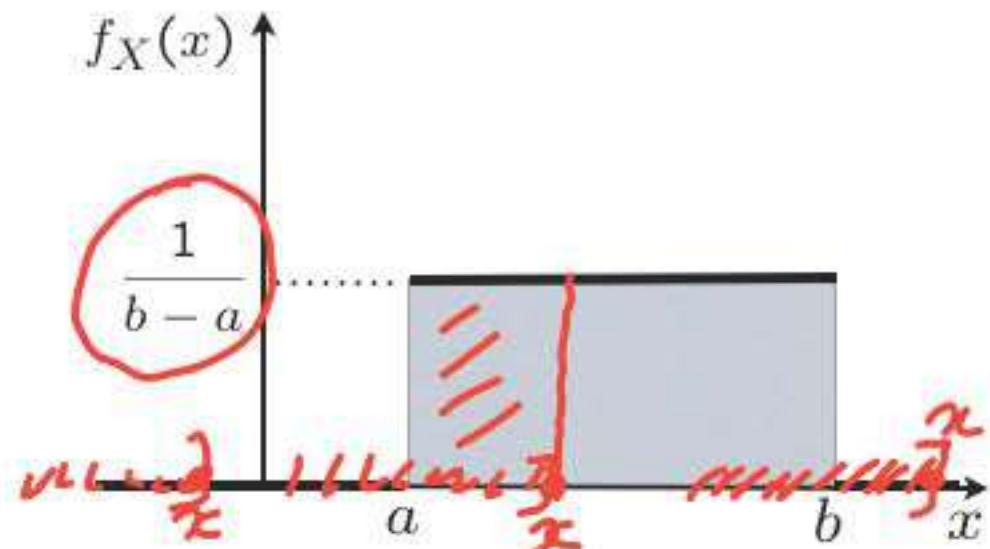
- Continuous random variables:

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



$$P(X \leq 4) = P(X \leq 3) + P(3 < X \leq 4)$$

$$\boxed{\frac{dF_X}{dx}(x) = f_X(x)}$$

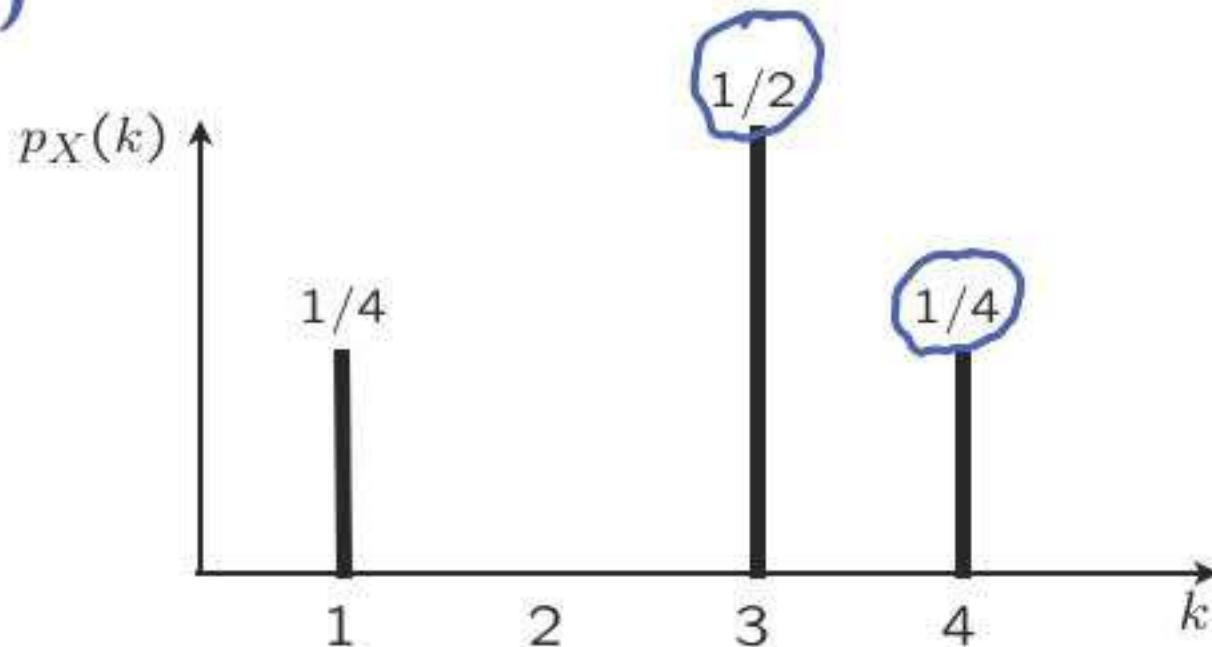


Cumulative distribution function (CDF)

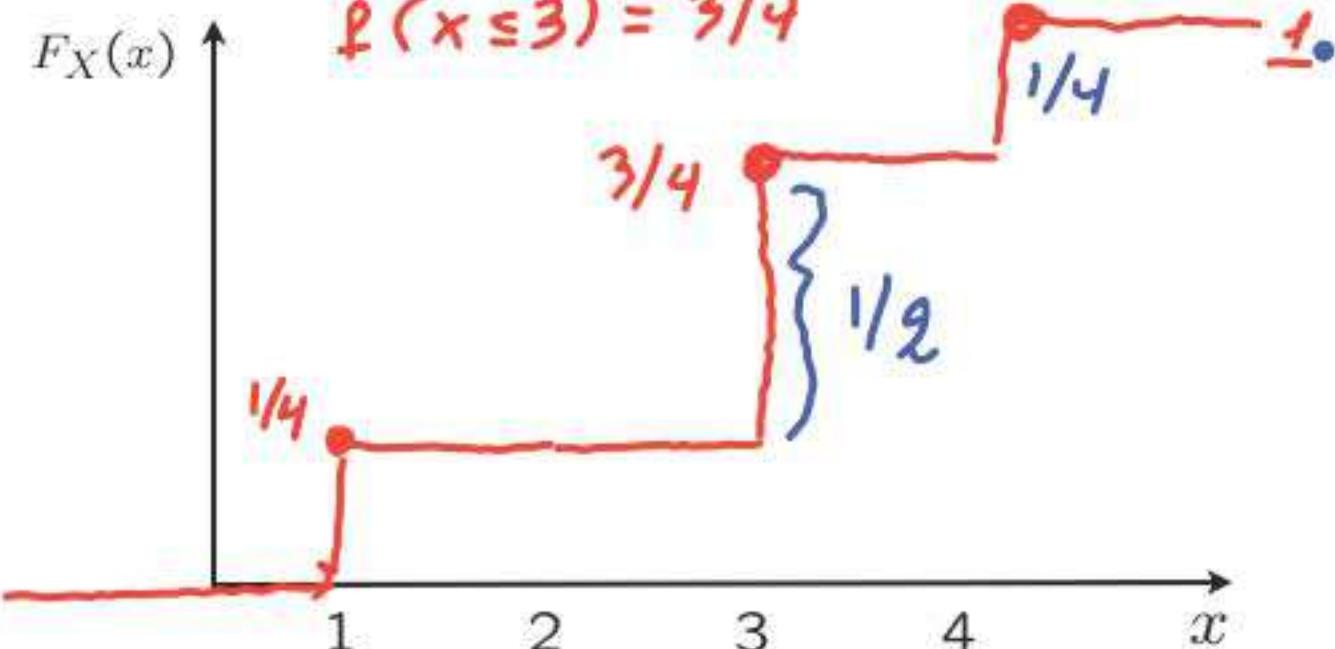
CDF definition: $F_X(x) = P(X \leq x)$

- Discrete random variables:

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



$$\begin{aligned}P(X \leq 1) &= 1/4 \\P(X \leq 3) &= 3/4 \\P(X \leq 4) &= 1\end{aligned}$$



General CDF properties

$$F_X(x) = P(X \leq x)$$



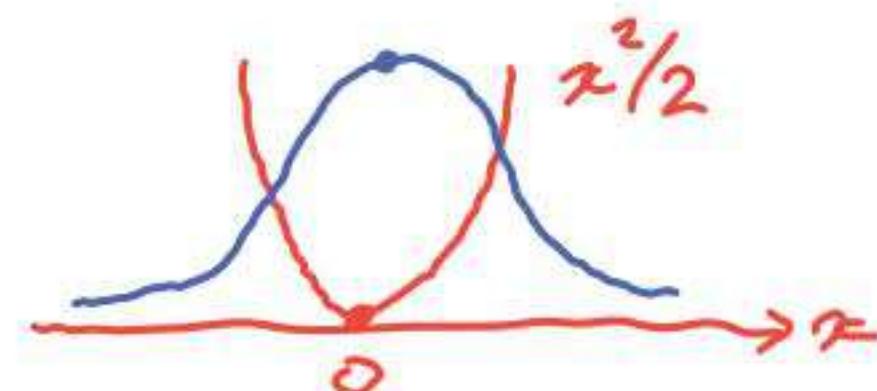
- Non-decreasing If $y \geq x \Rightarrow F_X(y) \geq F_X(x)$
- $F_X(x)$ tends to 1, as $x \rightarrow \infty$
- $F_X(x)$ tends to 0, as $x \rightarrow -\infty$

Normal (Gaussian) random variables

- Important in the theory of probability
 - Central limit theorem
- Prevalent in applications
 - Convenient analytical properties
 - Model of noise consisting of many, small independent noise terms

Standard normal (Gaussian) random variables

- Standard normal $N(0, 1)$: $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$



calculus:

$$\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}$$

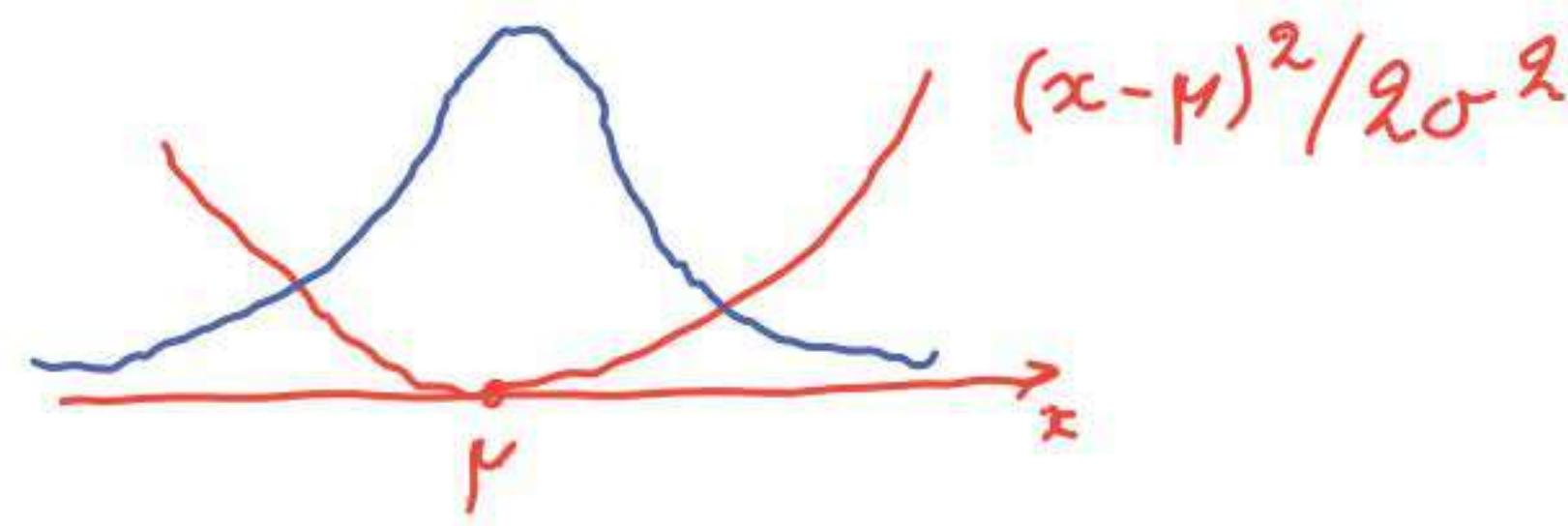
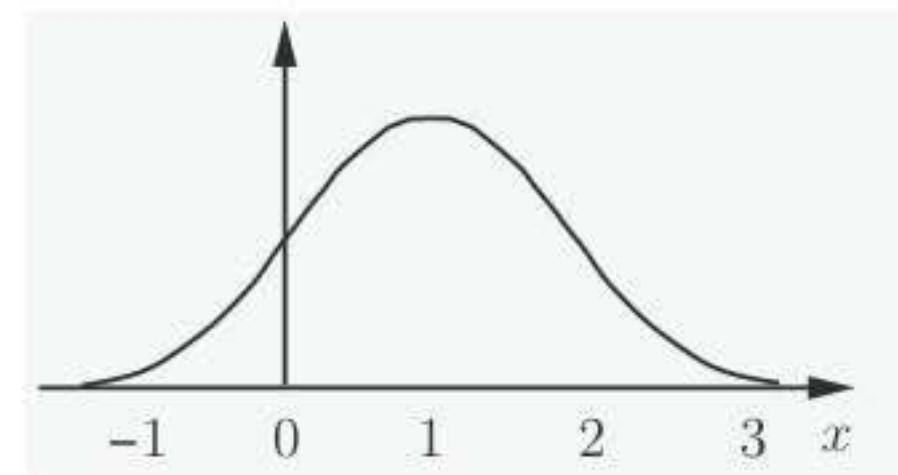
- $E[X] = 0$

- $\text{var}(X) = 1$

integrate by parts

General normal (Gaussian) random variables

- General normal $N(\mu, \sigma^2)$: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$
 $\sigma > 0$



- $E[X] = \mu$
- $\text{var}(X) = \sigma^2$

Linear functions of a normal random variable

- Let $Y = aX + b \quad X \sim N(\mu, \sigma^2)$

$$E[Y] = a\mu + b$$

$$\text{Var}(Y) = a^2 \sigma^2$$

- Fact (will prove later in this course):

$$Y \sim N(a\mu + b, a^2\sigma^2)$$

- Special case: $a = 0$?

$$\begin{array}{l} Y = b \quad \text{discrete} \\ \nearrow \\ n(b, 0) \end{array}$$

Standard normal tables

- No closed form available for CDF

but have tables, for the standard normal

$$Y \sim N(0, 1)$$

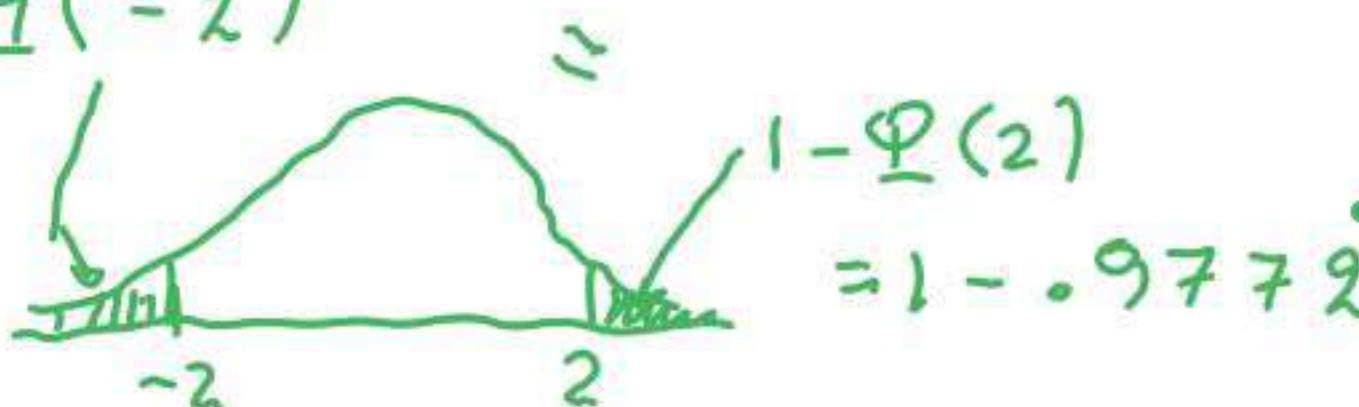
$$\Phi(y) = F_Y(y) = P(Y \leq y)$$



$$\Phi(0) = P(Y \leq 0) = 0.5$$

$$\Phi(1.16) = 0.8770 \quad \Phi(2.9) = 0.9981$$

$$\Phi(-2)$$



	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0.1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0.2	5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0.3	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0.4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0.5	6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0.6	7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0.7	7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0.8	7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0.9	8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1.0	8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1.1	8643	8665	8686	8708	8729	8749	8770	8790	8810	8830
1.2	8849	8869	8888	8907	8925	8944	8962	8980	8997	9015
1.3	9032	9049	9066	9082	9099	9115	9131	9147	9162	9177
1.4	9192	9207	9222	9236	9251	9265	9279	9292	9306	9319
1.5	9332	9345	9357	9370	9382	9394	9406	9418	9429	9441
1.6	9452	9463	9474	9484	9495	9505	9515	9525	9535	9545
1.7	9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
1.8	9641	9649	9656	9664	9671	9678	9686	9693	9699	9706
1.9	9713	9719	9726	9732	9738	9744	9750	9756	9761	9767
2.0	9772	9778	9783	9788	9793	9798	9803	9808	9812	9817
2.1	9821	9826	9830	9834	9838	9842	9846	9850	9854	9857
2.2	9861	9864	9868	9871	9875	9878	9881	9884	9887	9890
2.3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
2.4	9918	9920	9922	9925	9927	9929	9931	9932	9934	9936
2.5	9938	9940	9941	9943	9945	9946	9948	9949	9951	9952
2.6	9953	9955	9956	9957	9959	9960	9961	9962	9963	9964
2.7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974
2.8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981
2.9	9981	9982	9982	9983	9984	9984	9985	9985	9986	9986

Standardizing a random variable

- Let X have mean μ and variance $\sigma^2 > 0$

- Let $Y = \frac{X - \mu}{\sigma}$ $E[Y] = 0$ $\text{Var}(Y) = \frac{1}{\sigma^2} \text{Var}(x) = 1$

$$X = \mu + \sigma Y$$

- If also X is normal, then: $Y \sim N(0, 1)$

Calculating normal probabilities

- Express an event of interest in terms of standard normal

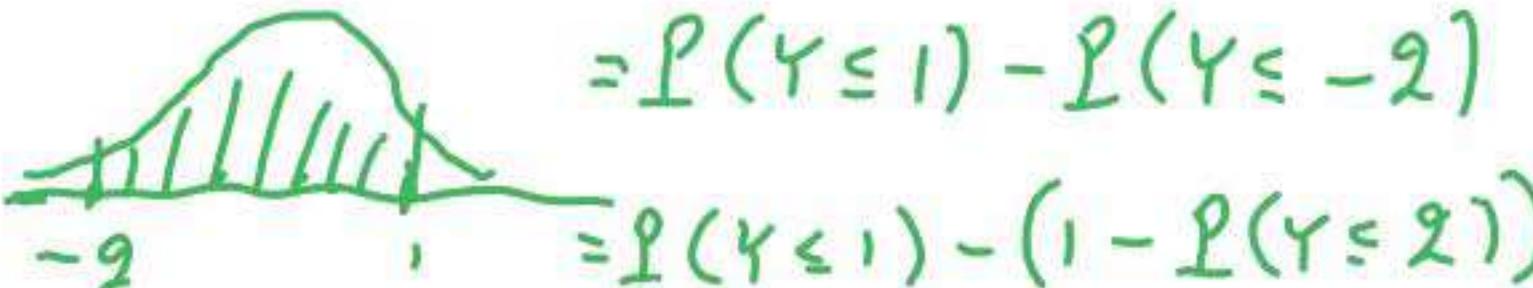
$$X \sim N(6, 4) \quad \sigma = 2$$

st. normal

$$\frac{2 - 6}{2} \leq \frac{X - 6}{2} \leq \frac{8 - 6}{2}$$

$$P(2 \leq X \leq 8) = P(-2 \leq Y \leq 1)$$

$$= P(Y \leq 1) - P(Y \leq -2)$$



$$= P(Y \leq 1) - (1 - P(Y \leq -2))$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 9: Conditioning on an event; Multiple continuous r.v.'s

- Conditioning a r.v. on an event
 - Conditional PDF
 - Conditional expectation and the expected value rule
 - Exponential PDF: memorylessness
 - Total probability and expectation theorems
 - Mixed distributions
- Jointly continuous r.v.'s and joint PDFs
 - From the joints to the marginals
 - Uniform joint PDF example
 - The expected value rule and linearity of expectations
 - The joint CDF

Conditional PDF, given an event

$$P(A) > 0$$

$$p_X(x) = P(X = x)$$

$$f_X(x) \cdot \delta \approx P(x \leq X \leq x + \delta)$$

$$p_{X|A}(x) = P(X = x | A)$$

$$\underline{f_{X|A}(x)} \cdot \delta \approx P(x \leq X \leq x + \delta | A)$$

$$P(X \in B) = \sum_{x \in B} p_X(x)$$

$$P(X \in B) = \int_B f_X(x) dx$$

$$P(X \in B | A) = \sum_{x \in B} p_{X|A}(x)$$

$$P(X \in B | A) = \int_B f_{X|A}(x) dx$$

Def

$$\sum_x p_{X|A}(x) = 1$$

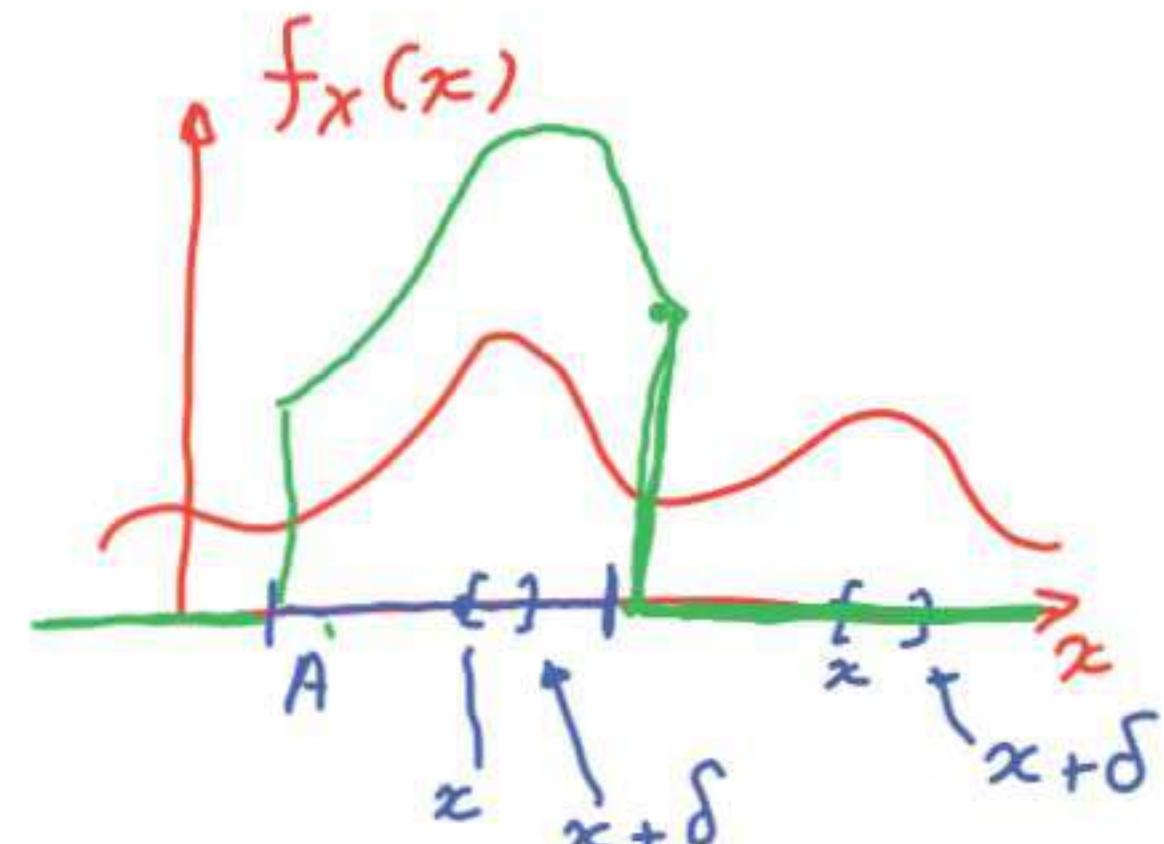
$$\int_{\bullet} f_{X|A}(x) dx = 1$$

Conditional PDF of X , given that $\underline{X \in A}$

$$\text{P}(x \leq X \leq x + \delta | X \in A) \approx f_{X|X \in A}(x) \cdot \cancel{\delta}$$

$$= \frac{\text{P}(x \leq X \leq x + \delta, X \in A)}{\text{P}(A)}$$

$$= \frac{\text{P}(x \leq X \leq x + \delta)}{\text{P}(A)} \approx \frac{f_X(x)}{\text{P}(A)} \cancel{\delta}$$



$$f_{X|X \in A}(x) = \begin{cases} 0, & \text{if } x \notin A \\ \frac{f_X(x)}{\text{P}(A)}, & \text{if } x \in A \end{cases}$$

Conditional expectation of X , given an event

$$\mathbb{E}[X] = \sum_x x p_X(x)$$

$$\mathbb{E}[X] = \int x f_X(x) dx$$

$$\mathbb{E}[X | A] = \sum_x x p_{X|A}(x)$$

$$\mathbb{E}[X | A] = \int x f_{X|A}(x) dx \quad \text{Def}$$

Expected value rule:

$$\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$$

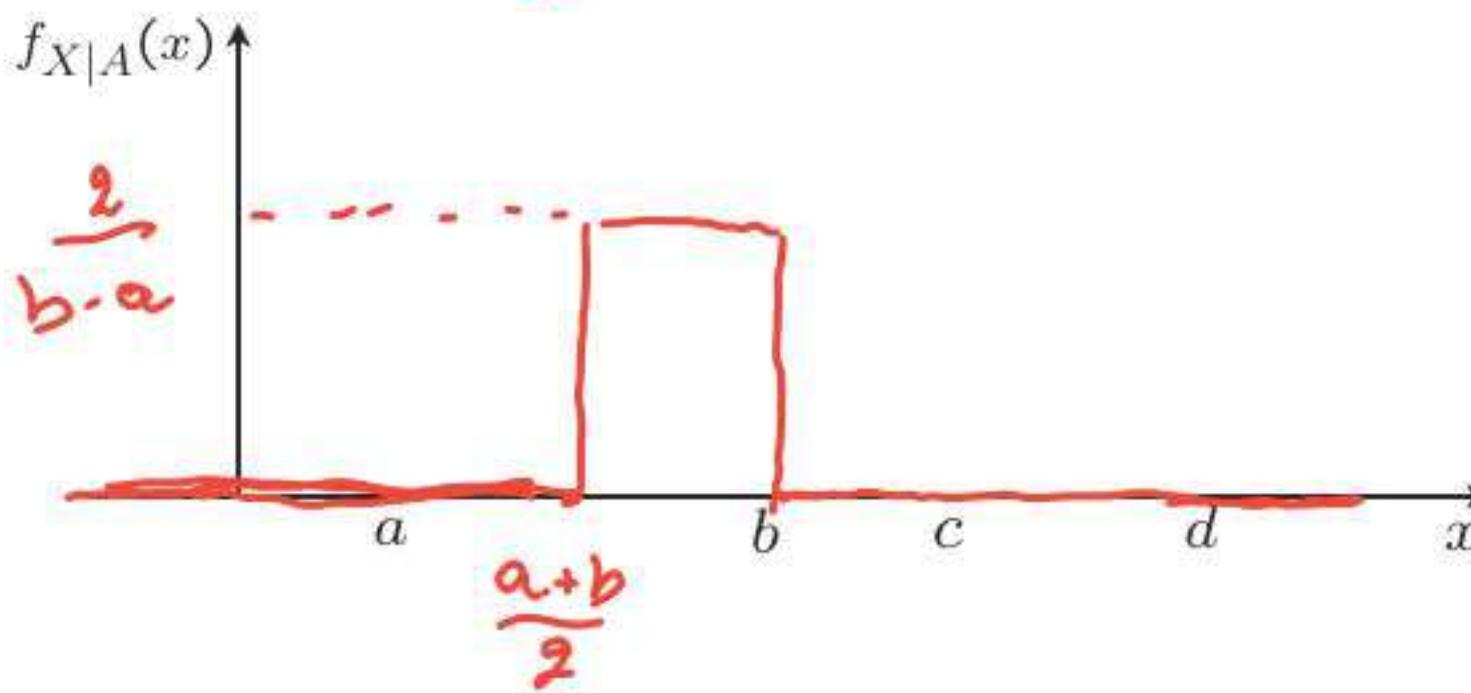
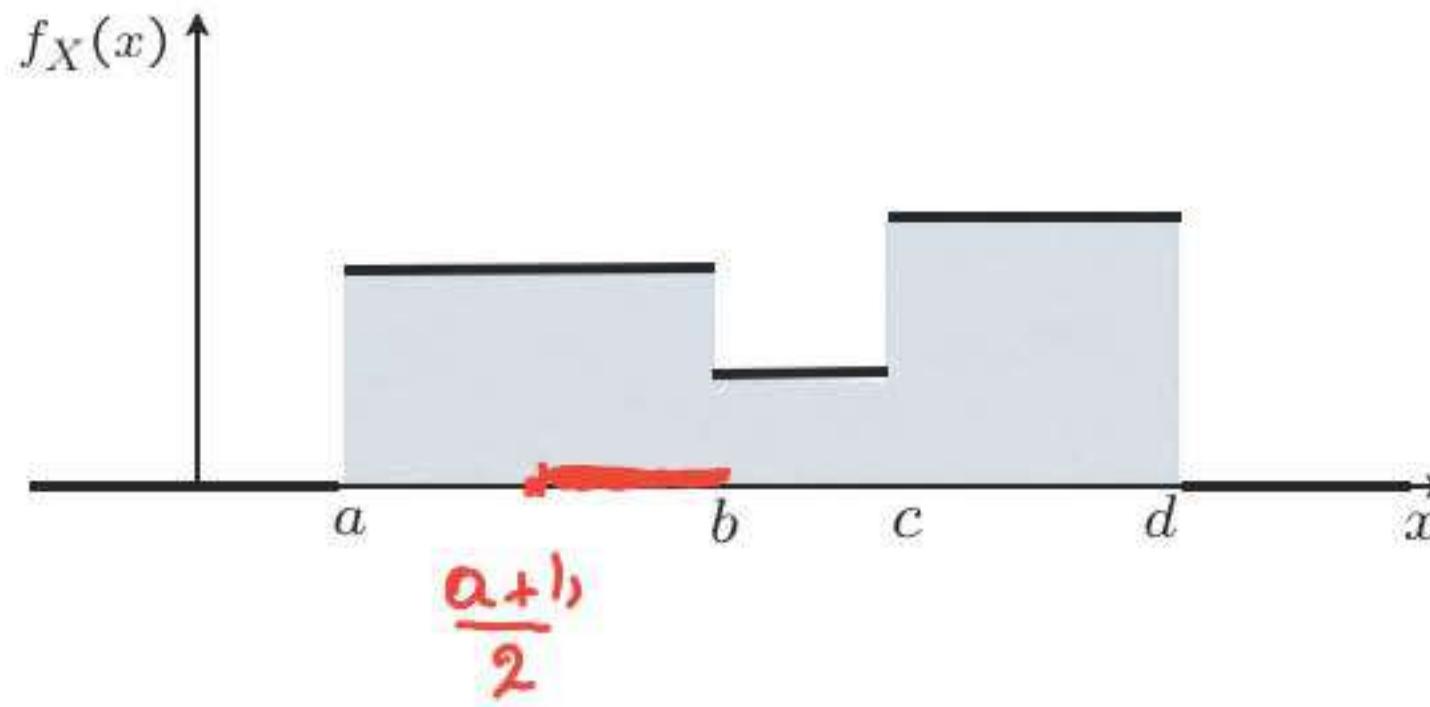
$$\mathbb{E}[g(X)] = \int g(x) f_X(x) dx$$

$$\mathbb{E}[g(X) | A] = \sum_x g(x) p_{X|A}(x)$$

$$\mathbb{E}[g(X) | A] = \int g(x) f_{X|A}(x) dx$$

Example

$$A : \frac{a+b}{2} \leq X \leq b$$



$$\mathbb{E}[X | A] = \frac{1}{2} \cdot \frac{a+b}{2} + \frac{1}{2} b$$

$$= \frac{1}{4} a + \frac{3}{4} b$$

$$\mathbb{E}[X^2 | A] =$$

$$\int_{\frac{a+b}{2}}^b \frac{2}{b-a} \cdot x^2 dx$$

Memorylessness of the exponential PDF

- Do you prefer a used or a new “exponential” light bulb? **Probabilistically identical!**

- Bulb lifetime T : $\text{exponential}(\lambda)$

$$P(T > x) = e^{-\lambda x}, \text{ for } x \geq 0$$

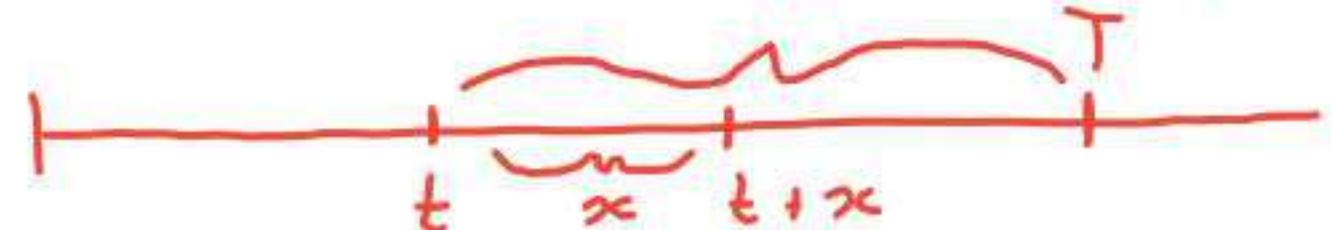
- we are told that $T > t$

- r.v. X : remaining lifetime $= T - t$

$$P(X > x | T > t) = e^{-\lambda x}, \text{ for } x \geq 0$$

$$= \frac{P(T-t > x, T > t)}{P(T > t)} = \frac{P(T > t+x, T > t)}{P(T > t)} = \frac{P(T > t+x)}{P(T > t)}$$

$$= \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x}$$

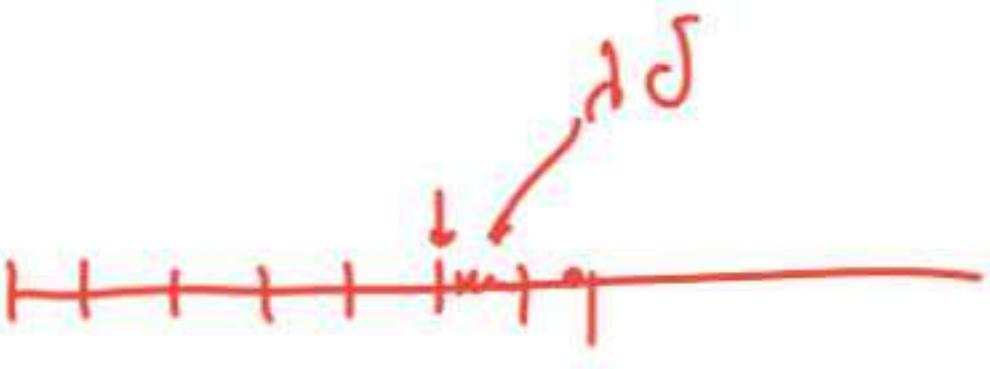


Memorylessness of the exponential PDF

$$f_T(x) = \lambda e^{-\lambda x}, \quad \text{for } x \geq 0$$

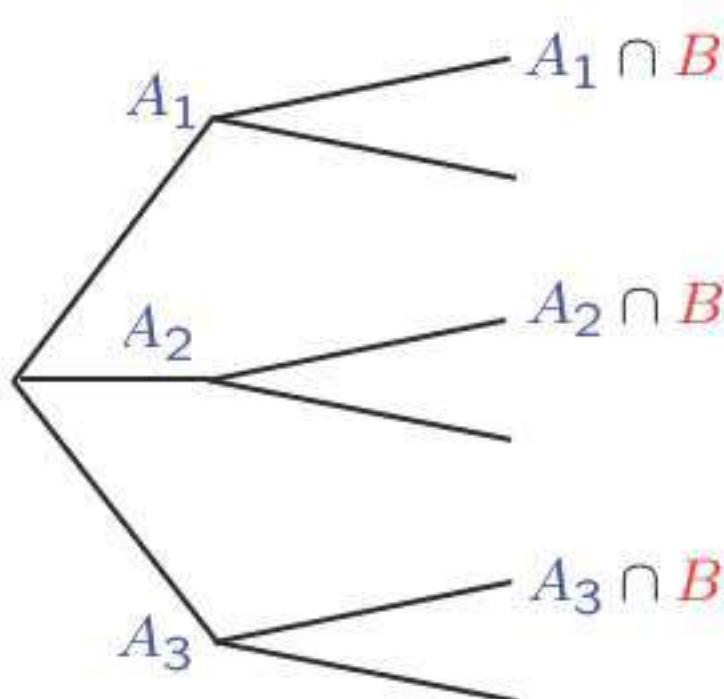
$$\mathbf{P}(0 \leq T \leq \delta) \approx f_T(0) \cdot \delta = \lambda \delta$$

$$\mathbf{P}(t \leq T \leq t + \delta \mid T > t) = \approx \lambda \delta$$



similar to an independent coin flip,
every δ time steps,
with $\mathbf{P}(\text{success}) \approx \lambda \delta$

Total probability and expectation theorems

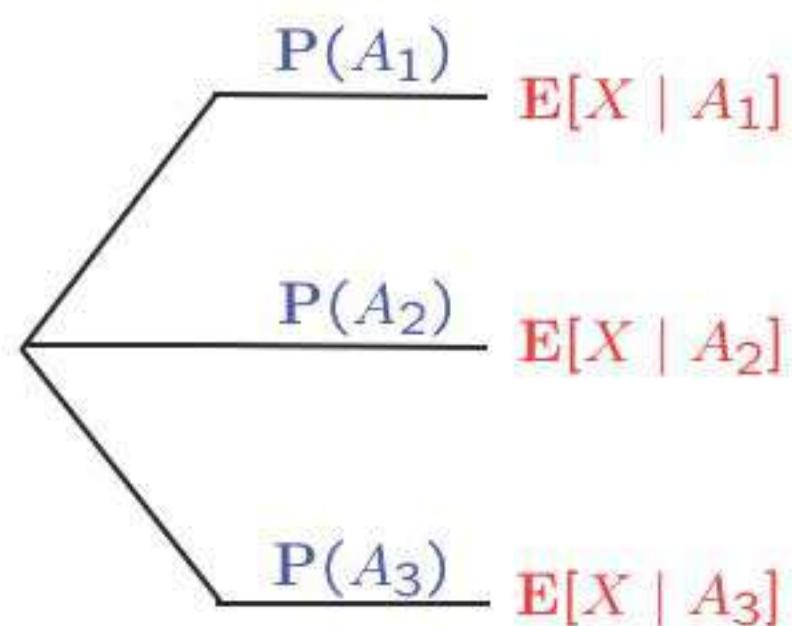


$$P(B) = P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)$$

$$p_X(x) = P(A_1)p_{X|A_1}(x) + \cdots + P(A_n)p_{X|A_n}(x)$$

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(A_1)P(X \leq x | A_1) + \cdots \\ &= P(A_1)F_{X|A_1}(x) + \cdots \end{aligned}$$

$$f_X(x) = P(A_1)f_{X|A_1}(x) + \cdots + P(A_n)f_{X|A_n}(x)$$



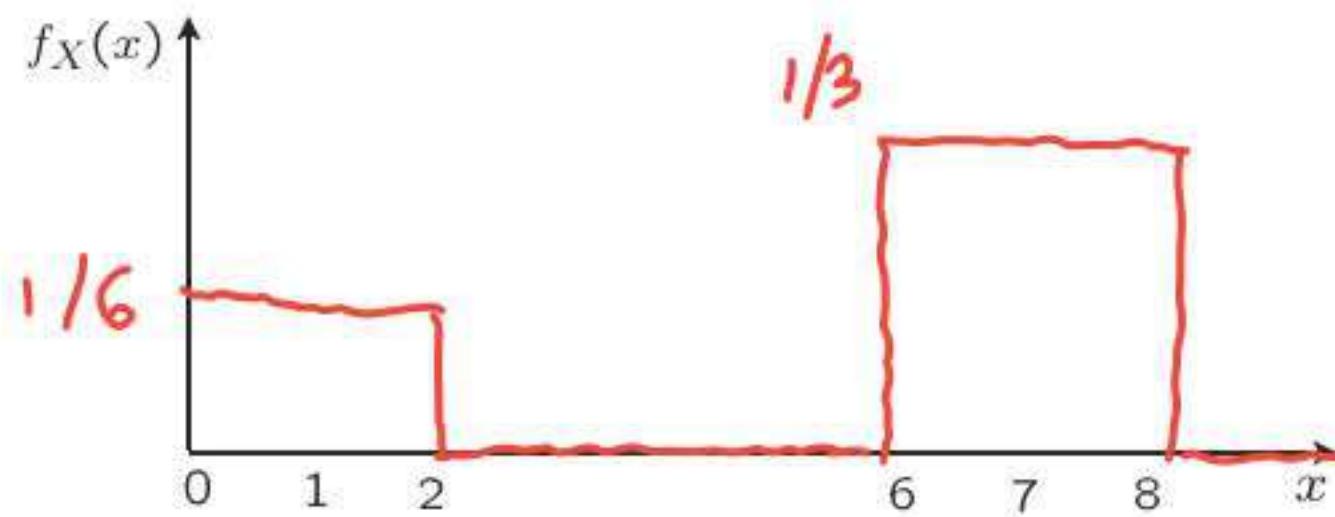
$$\int x f_X(x) dx = P(A_1) \int x f_{X|A_1}(x) dx + \cdots$$

$$E[X] = P(A_1)E[X|A_1] + \cdots + P(A_n)E[X|A_n]$$

Example

- Bill goes to the supermarket shortly, with probability $1/3$, at a time uniformly distributed between 0 and 2 hours from now; or with probability $2/3$, later in the day at a time uniformly distributed between 6 and 8 hours from now

$$\Pr(A_1) = \frac{1}{3} \quad f_{X|A_1} \sim \text{unif}[0, 2] \quad \Pr(A_2) = \frac{2}{3} \quad f_{X|A_2} \sim U[6, 8]$$



$$f_X(x) = \Pr(A_1)f_{X|A_1}(x) + \dots + \Pr(A_n)f_{X|A_n}(x)$$

$$\bullet \quad E[X] = \Pr(A_1)E[X | A_1] + \dots + \Pr(A_n)E[X | A_n]$$

$$\frac{1}{3} \cdot 1 + \frac{2}{3} \cdot 7$$

Mixed distributions

$X = \begin{cases} \text{uniform on } [0, 2], & \text{with probability } 1/2 \\ 1, & \text{with probability } 1/2 \end{cases}$	Is X discrete? No
Y discrete Z continuous	$X = \begin{cases} Y, & \text{with probability } p \\ Z, & \text{with probability } 1 - p \end{cases}$ Is X continuous? No
	$P(X=1) = 1/2$ X is mixed

$$F_X(x) = P \cdot P(Y \leq x) + (1-p) P(Z \leq x)$$

$$= p F_Y(x) + (1-p) F_Z(x)$$

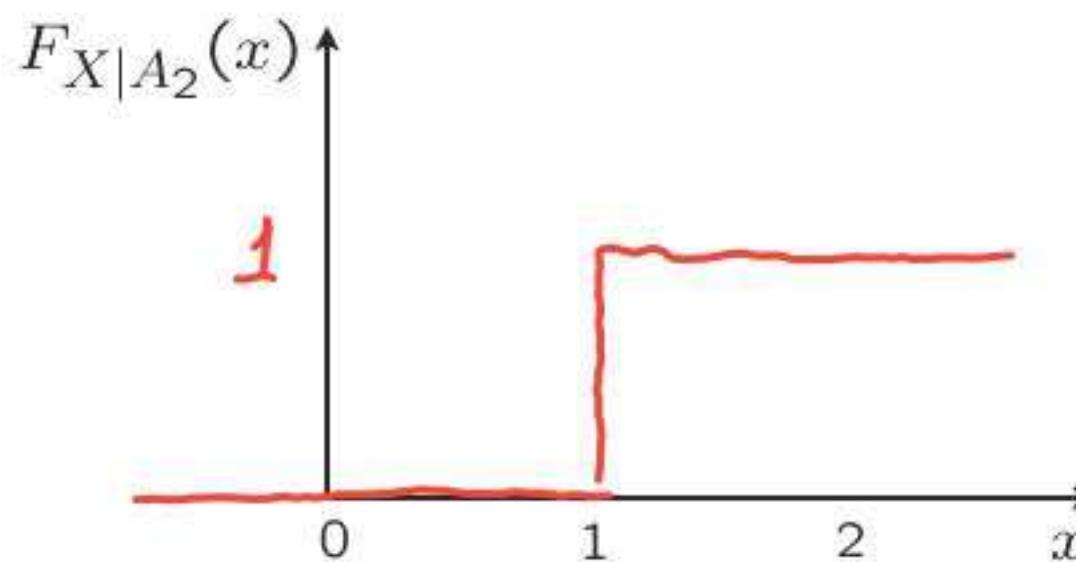
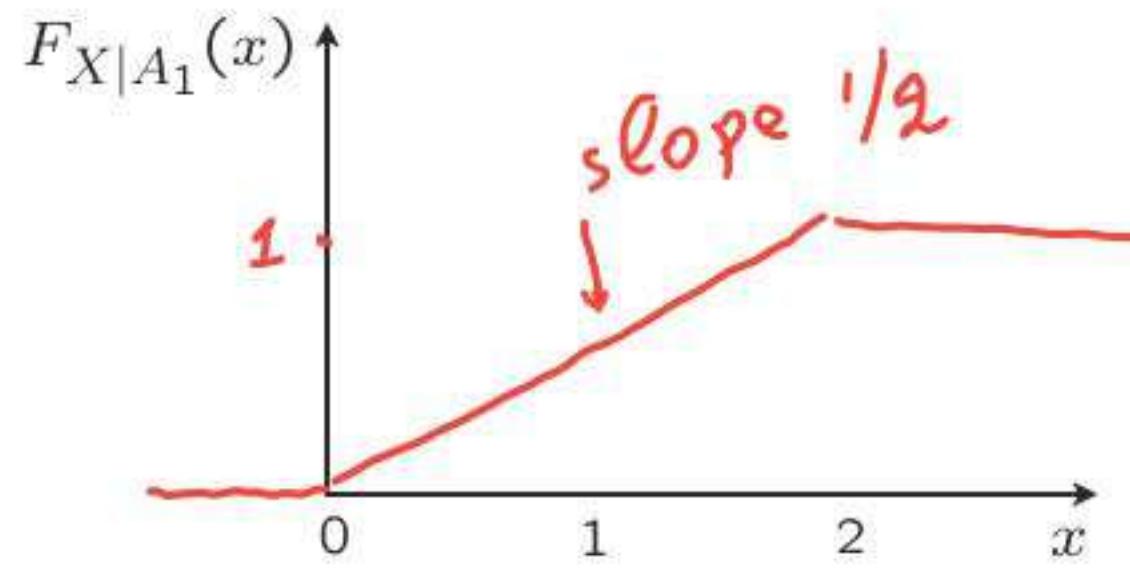
$$E[X] = p E[Y] + (1-p) E[Z]$$

Mixed distributions

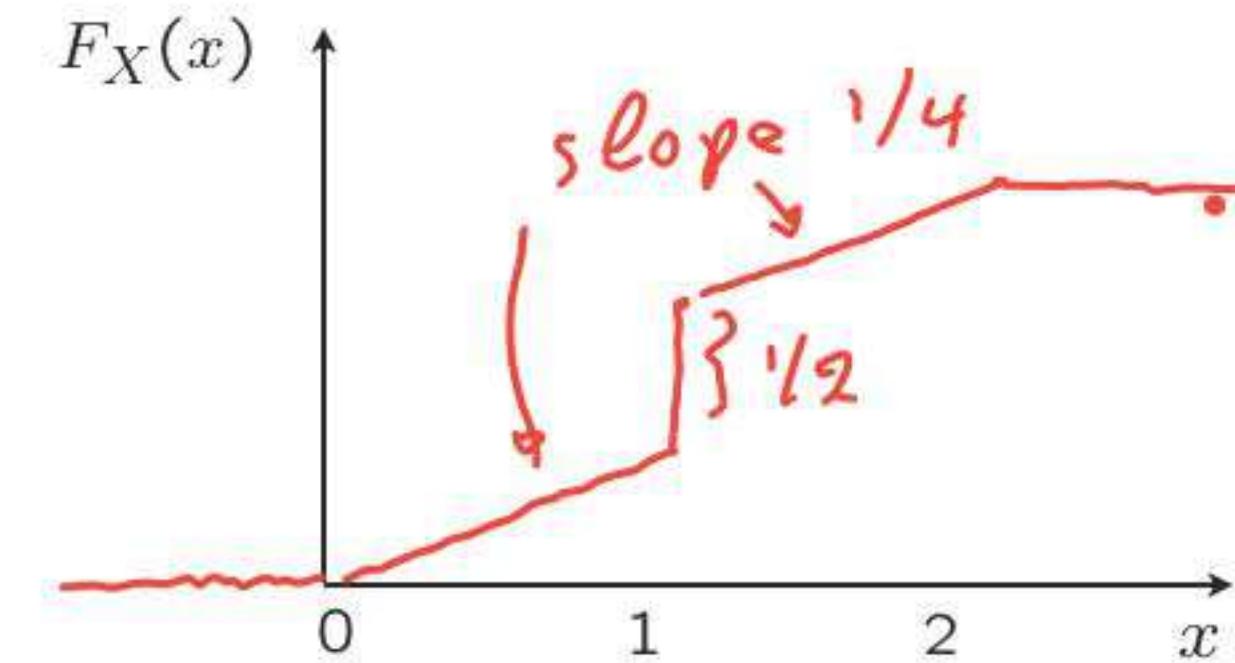
$X = \begin{cases} \text{uniform on } [0, 2], & \text{with probability } 1/2 \\ 1, & \text{with probability } 1/2 \end{cases}$

A_1

A_2



$$F_X(x) = P(A_1)F_{X|A_1}(x) + P(A_2)F_{X|A_2}(x)$$



Jointly continuous r.v.'s and joint PDFs

$$p_X(x) \quad f_X(x)$$

$$p_{X,Y}(x,y) \quad f_{X,Y}(x,y)$$

$$p_{X,Y}(x,y) = \mathbf{P}(X = x \text{ and } Y = y) \geq 0$$

$$f_{X,Y}(x,y) \geq 0$$

$$\mathbf{P}((X,Y) \in B) = \sum_{(x,y) \in B} p_{X,Y}(x,y)$$

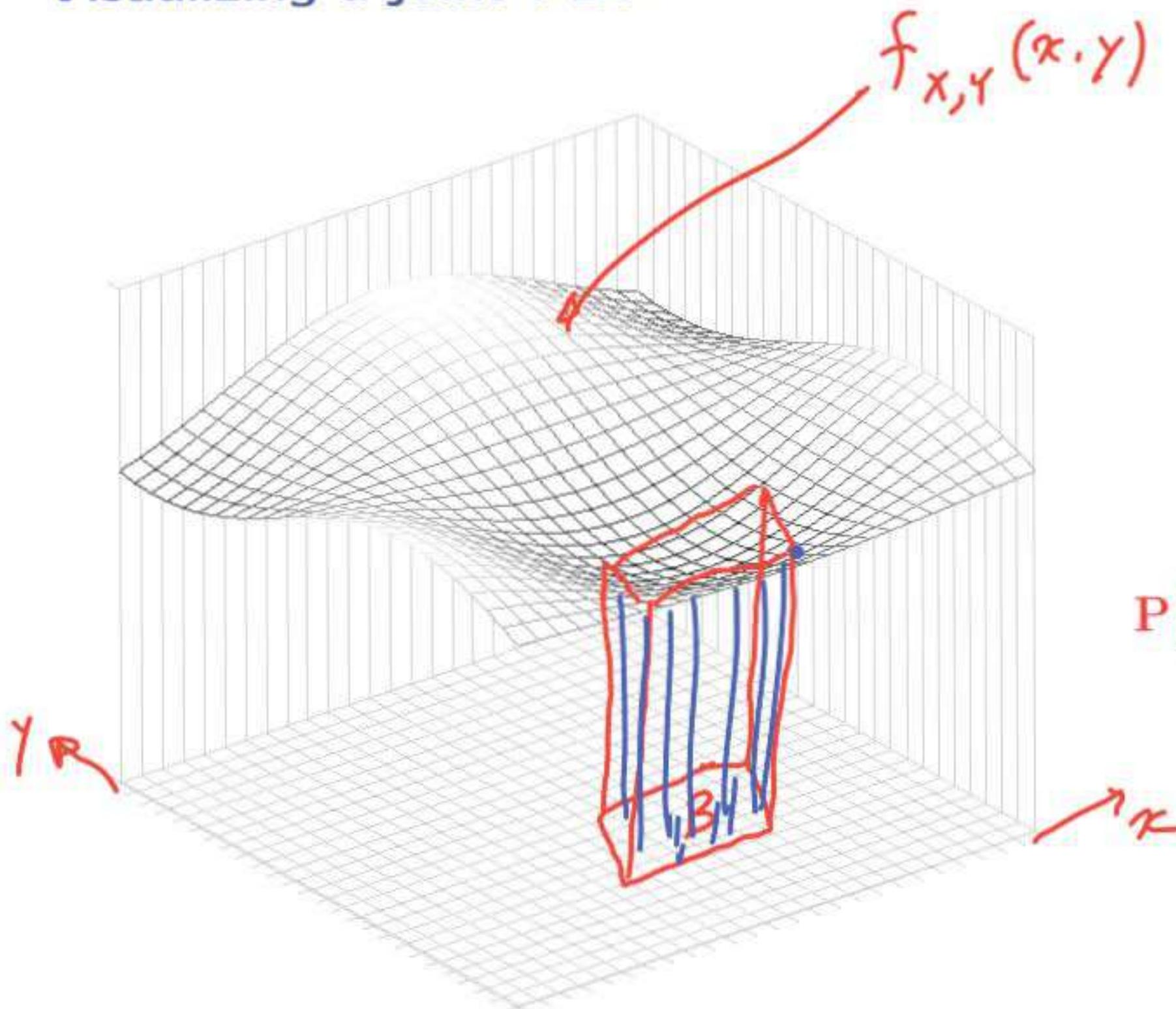
$$\mathbf{P}((X,Y) \in B) = \iint_{(x,y) \in B} f_{X,Y}(x,y) dx dy \bullet$$

$$\sum_x \sum_y p_{X,Y}(x,y) = 1$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$$

Definition: Two random variables are **jointly continuous** if they can be described by a joint PDF

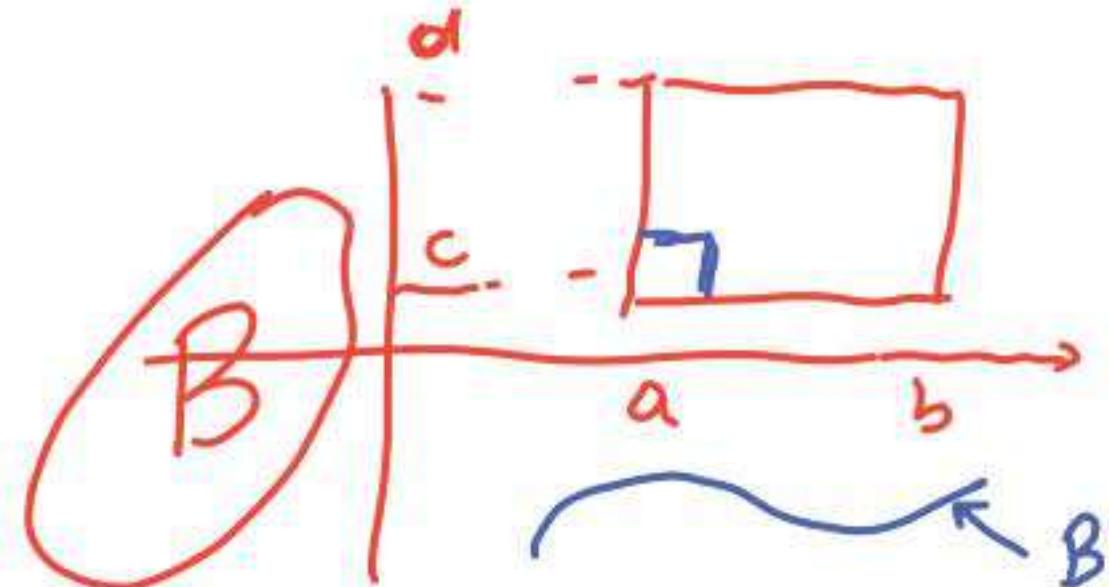
Visualizing a joint PDF



$$P((X, Y) \in B) = \iint_{(x,y) \in B} f_{X,Y}(x, y) dx dy$$

On joint PDFs

$$\mathbf{P}((X, Y) \in B) = \iint_{(x,y) \in B} f_{X,Y}(x, y) dx dy$$



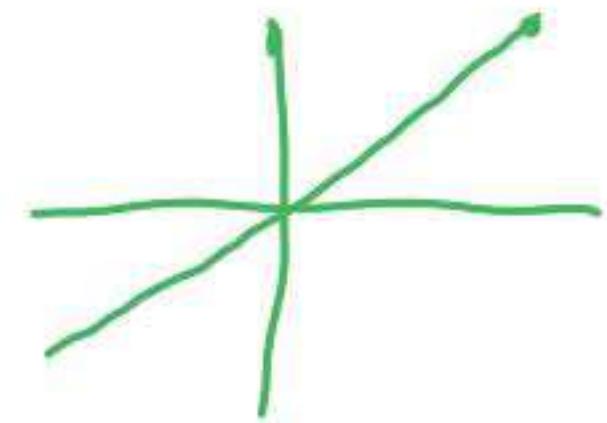
$$\mathbf{P}(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x, y) dx dy$$

$$\mathbf{P}(a \leq X \leq a + \delta, c \leq Y \leq c + \delta) \approx f_{X,Y}(a, c) \cdot \delta^2$$

$$Y = X$$

$f_{X,Y}(x, y)$: probability per unit area

$$\text{area}(B) = 0 \Rightarrow \mathbf{P}((X, Y) \in B) = 0$$



From the joint to the marginals

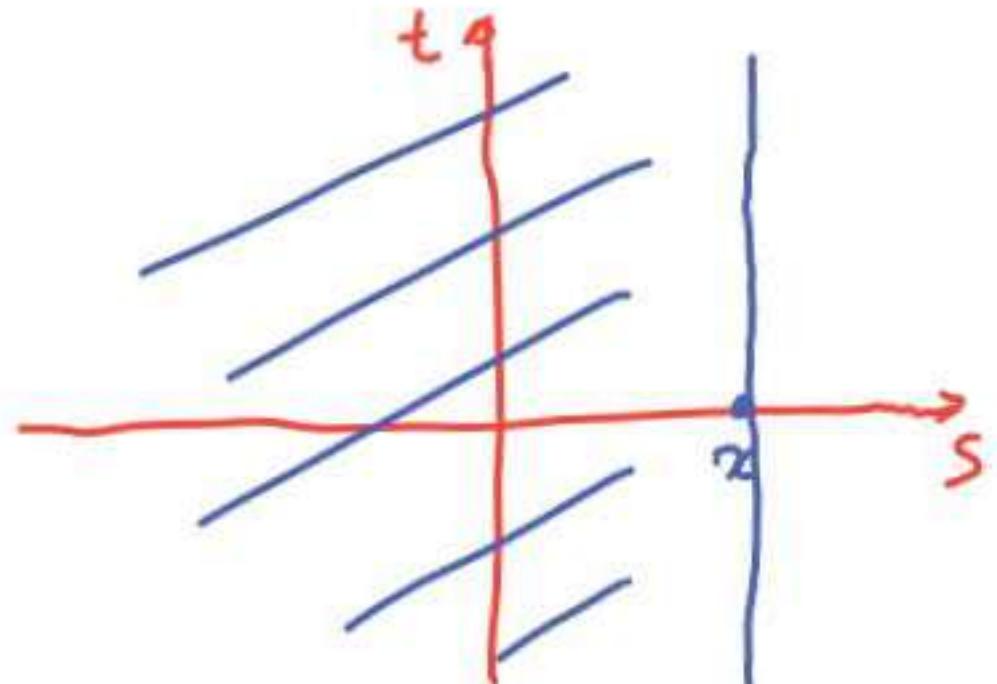
$$p_X(x) = \sum_y p_{X,Y}(x,y)$$

$$f_X(x) = \int f_{X,Y}(x,y) dy$$

$$p_Y(y) = \sum_x p_{X,Y}(x,y)$$

$$f_Y(y) = \int f_{X,Y}(x,y) dx$$

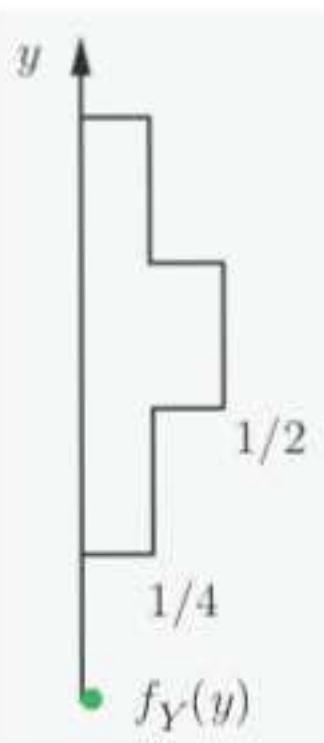
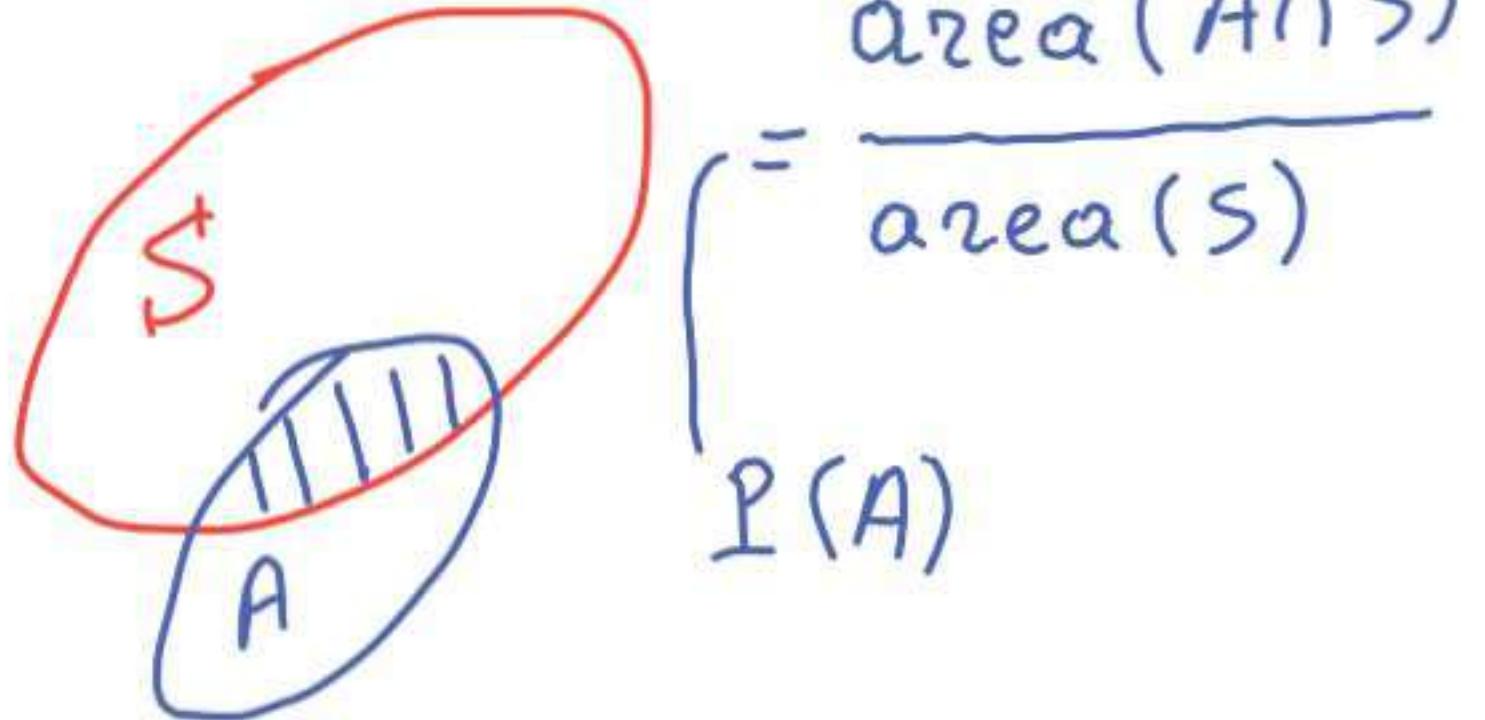
$$F_X(x) = P(X \leq x) = \int_{-\infty}^x \left[\int_{-\infty}^{\infty} f_{X,Y}(s,t) dt \right] ds$$



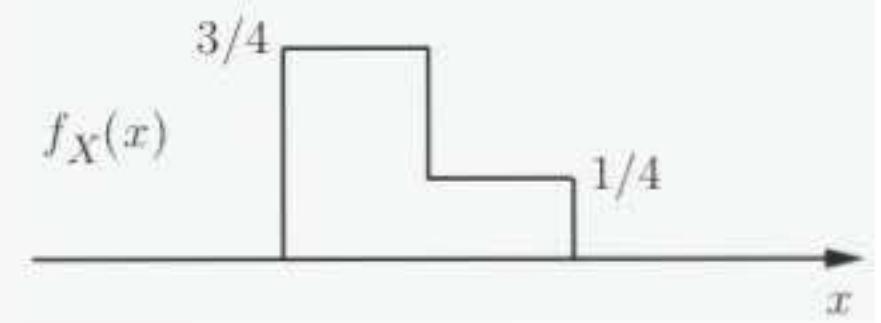
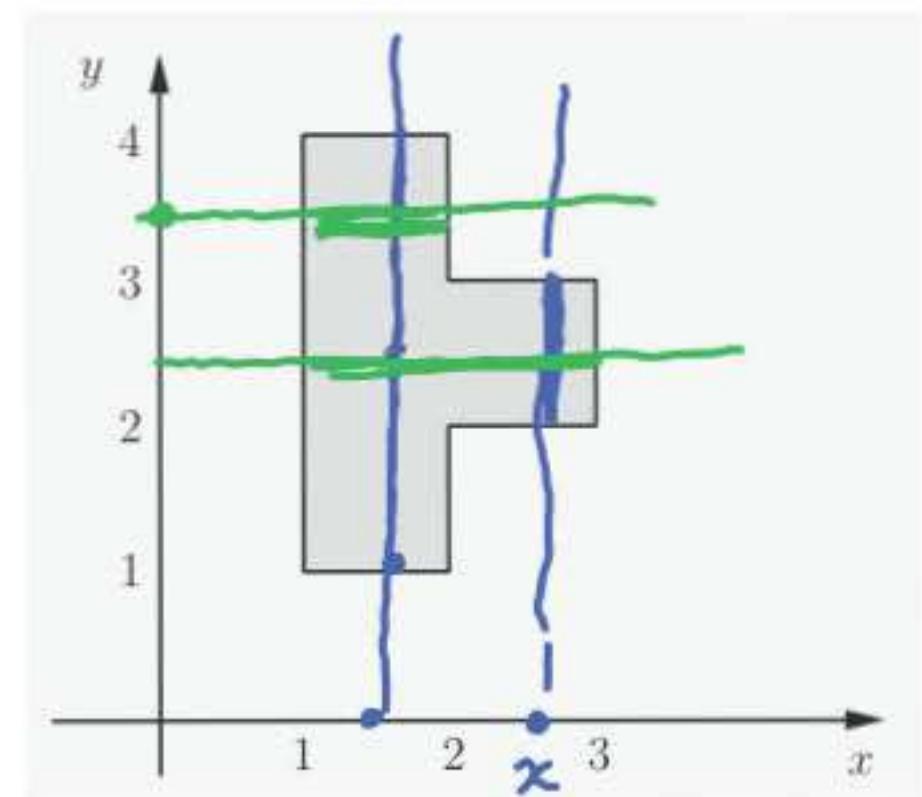
$$f_X(x) = \frac{dF_X(x)}{dx} = []$$

Uniform joint PDF on a set S

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\text{area of } S}, & \text{if } (x, y) \in S, \\ 0, & \text{otherwise.} \end{cases}$$



$$f_{X,Y} = \frac{1}{4}$$



More than two random variables

$$p_{X,Y,Z}(x, y, z)$$

$$f_{X,Y,Z}(x, y, z)$$

$$\sum_x \sum_y \sum_z p_{X,Y,Z}(x, y, z) = 1$$

$$p_X(x) = \sum_y \sum_z p_{X,Y,Z}(x, y, z)$$

$$p_{X,Y}(x, y) = \sum_z p_{X,Y,Z}(x, y, z)$$

Functions of multiple random variables

$$Z = g(X, Y)$$

Expected value rule:

$$\mathbb{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y)$$

$$\mathbb{E}[g(X, Y)] = \int \int g(x, y) f_{X,Y}(x, y) dx dy$$

Linearity of expectations

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[X_1 + \dots + X_n] = \mathbb{E}[X_1] + \dots + \mathbb{E}[X_n]$$

The joint CDF

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

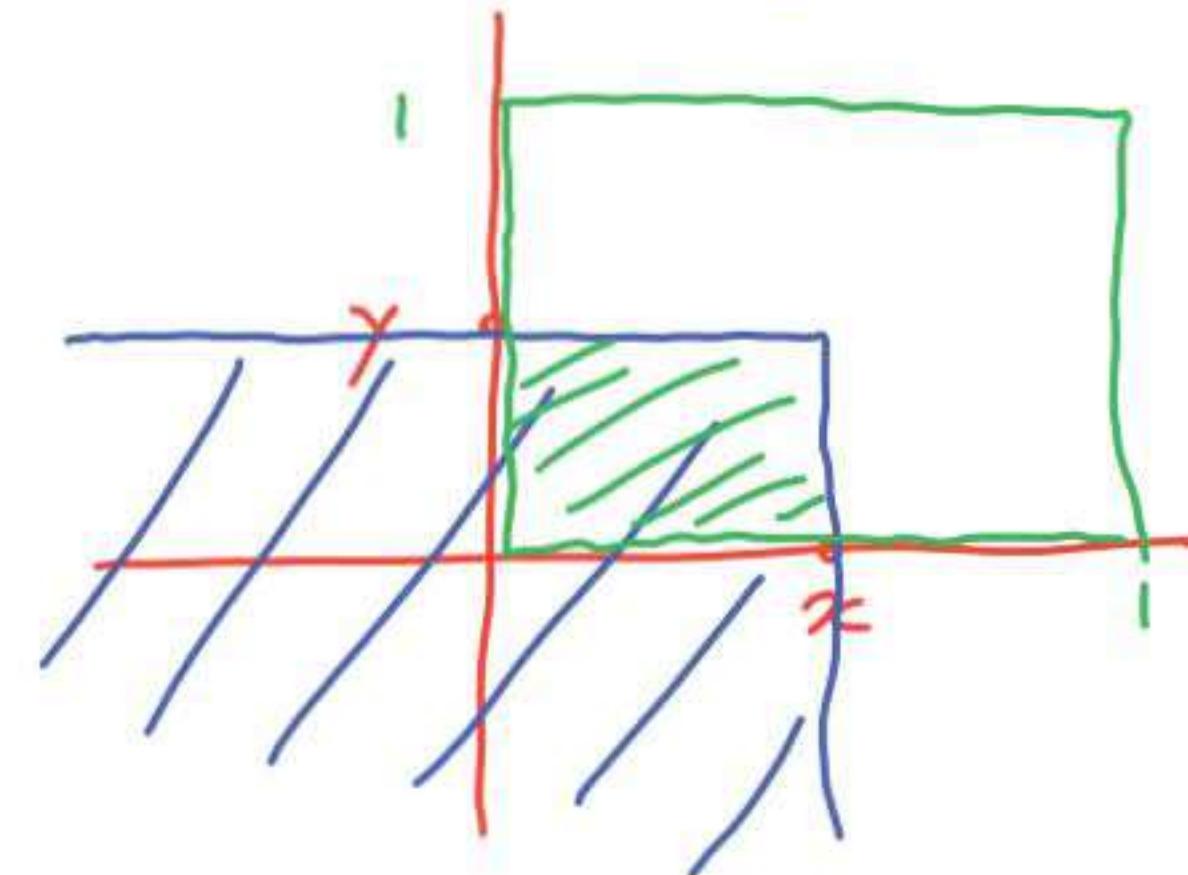
$$f_X(x) = \frac{dF_X}{dx}(x)$$

$$F_{X,Y}(x, y) = \mathbf{P}(X \leq x, Y \leq y) = \int_{-\infty}^y \left[\int_{-\infty}^x f_{X,Y}(s, t) ds \right] dt$$

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x, y)$$

$$F_{X,Y}(x, y) = xy$$

$$f_{X,Y}(x, y) = 1$$



MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 10: Conditioning on a random variable; Independence; Bayes' rule

- Conditioning X on Y
 - Total probability theorem
 - Total expectation theorem
- Independence
 - independent normals
- A comprehensive example
- Four variants of the Bayes rule

Conditional PDFs, given another r.v.

$$p_{X|Y}(x | y) = \mathbf{P}(X = x | Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}, \quad \text{if } p_Y(y) > 0$$

Definition: $f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$ if $f_Y(y) > 0$

$$\mathbf{P}(x \leq X \leq x + \delta | A) \approx f_{X|A}(x) \cdot \delta, \quad \text{where } \mathbf{P}(A) > 0$$

$$\overbrace{Y=y}^{\gamma=y} \quad \overbrace{Y \approx y}^{\gamma \approx y}$$

$$\mathbf{P}(x \leq X \leq x + \delta | y \leq Y \leq y + \epsilon) \approx \frac{f_{x,y}(x, y) \delta}{f_y(y) \delta} = f_{x|y}(x | y) \delta$$

Definition: $\mathbf{P}(X \in A | Y = y) = \int_A f_{X|Y}(x | y) dx$

$p_{X,Y}(x, y)$	$f_{X,Y}(x, y)$
$p_{X A}(x)$	$f_{X A}(x)$
$p_{X Y}(x y)$	$f_{X Y}(x y)$

Comments on conditional PDFs

$$f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

- $f_{X|Y}(x | y) \geq 0$

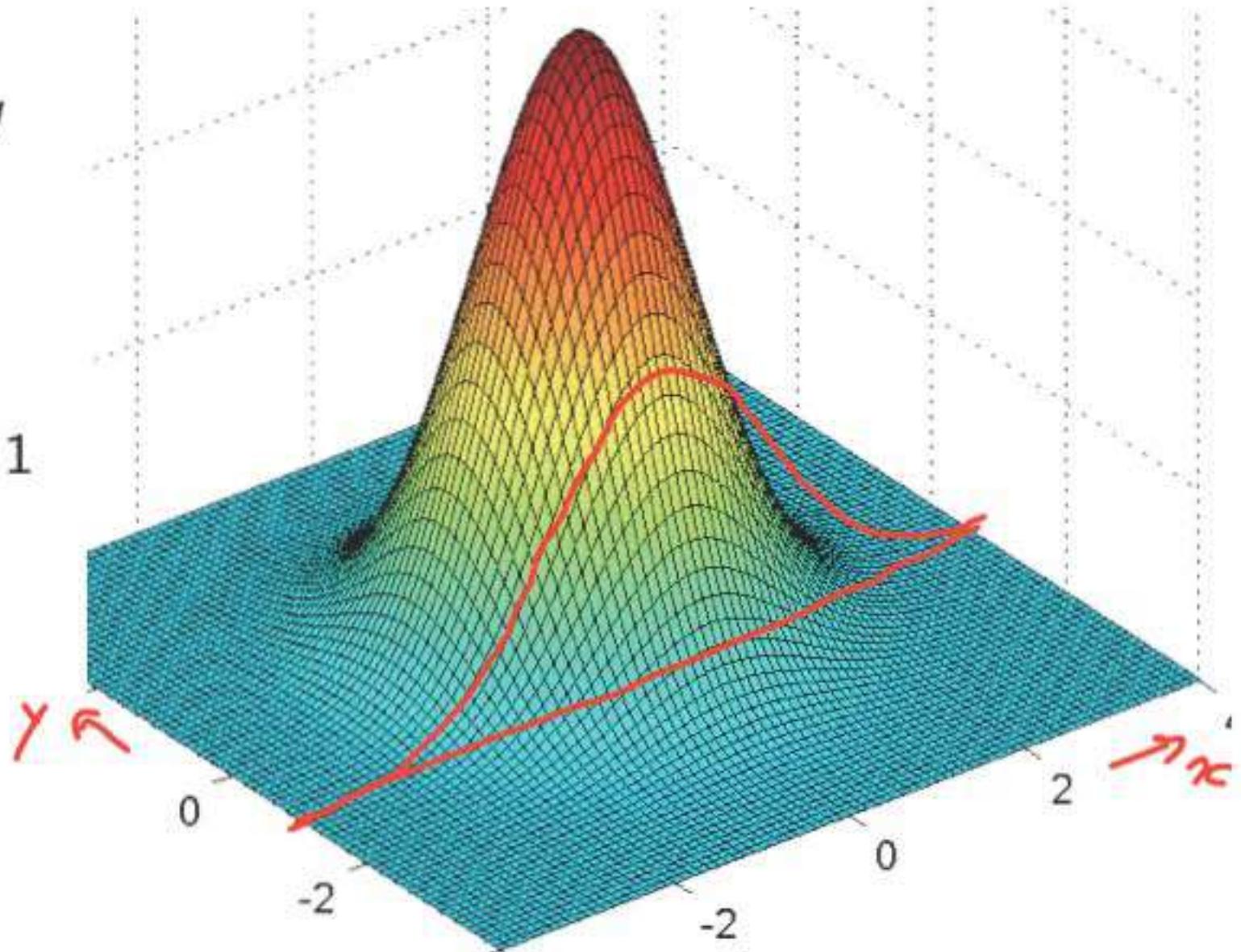
- Think of value of Y as fixed at some y
shape of $f_{X|Y}(\cdot | y)$: slice of the joint

- $\int_{-\infty}^{\infty} f_{X|Y}(x | y) dx = \frac{\int_{-\infty}^{\infty} f_{X,Y}(x, y) dx}{f_Y(y)} = 1$

- Multiplication rule:

$$f_{X,Y}(x, y) = f_Y(y) \cdot f_{X|Y}(x | y)$$

$$= f_X(x) \cdot f_{Y|X}(y | x)$$



Total probability and expectation theorems

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x|y)$$

$$\mathbb{E}[X | Y = y] = \sum_x x p_{X|Y}(x|y)$$

$$\mathbb{E}[X] = \sum_y p_Y(y) \mathbb{E}[X | Y = y]$$

- Expected value rule...

$$\mathbb{E}[g(x) | Y = y]$$

$$= \int_{-\infty}^{\infty} g(x) f_{x|y}(x|y) dx$$

$$f_X(x) = \underbrace{\int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x|y) dy}_{f_{X,Y}(x,y)} \quad \text{Thm.}$$

$$\mathbb{E}[X | Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx \quad \text{Def.}$$

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} f_Y(y) \mathbb{E}[X | Y = y] dy \\ &= \int_{-\infty}^{\infty} f_Y(y) \int_{-\infty}^{\infty} x f_{x|y}(x|y) dx dy \end{aligned}$$

$$= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} \cancel{f_Y(y)} f_{x|y}(x|y) dy dx$$

$$= \int_{-\infty}^{\infty} x f_X(x) dx = \mathbb{E}[X]$$

Independence

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \quad \text{for all } x, y$$

$$f_{X,Y}(x,y) = \underline{f_X(x)} f_Y(y), \quad \text{for all } x \text{ and } y$$

$$f_{Y|X} = f_Y$$

$$f_{X,Y}(x,y) = \underline{f_{X|Y}(x|y)} f_Y(y)$$

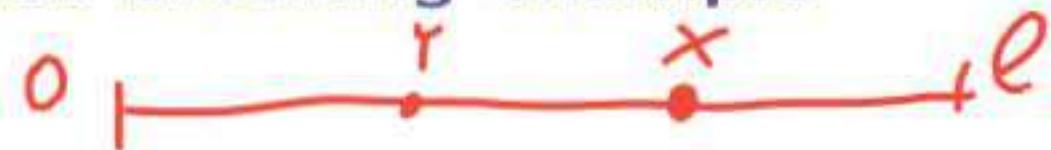
- equivalent to: $f_{X|Y}(x|y) = f_X(x)$, for all y with $f_Y(y) > 0$ and all x

If X, Y are **independent**: $E[XY] = E[X]E[Y]$

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$$

$g(X)$ and $h(Y)$ are also independent: $E[g(X)h(Y)] = E[g(X)] \cdot E[h(Y)]$

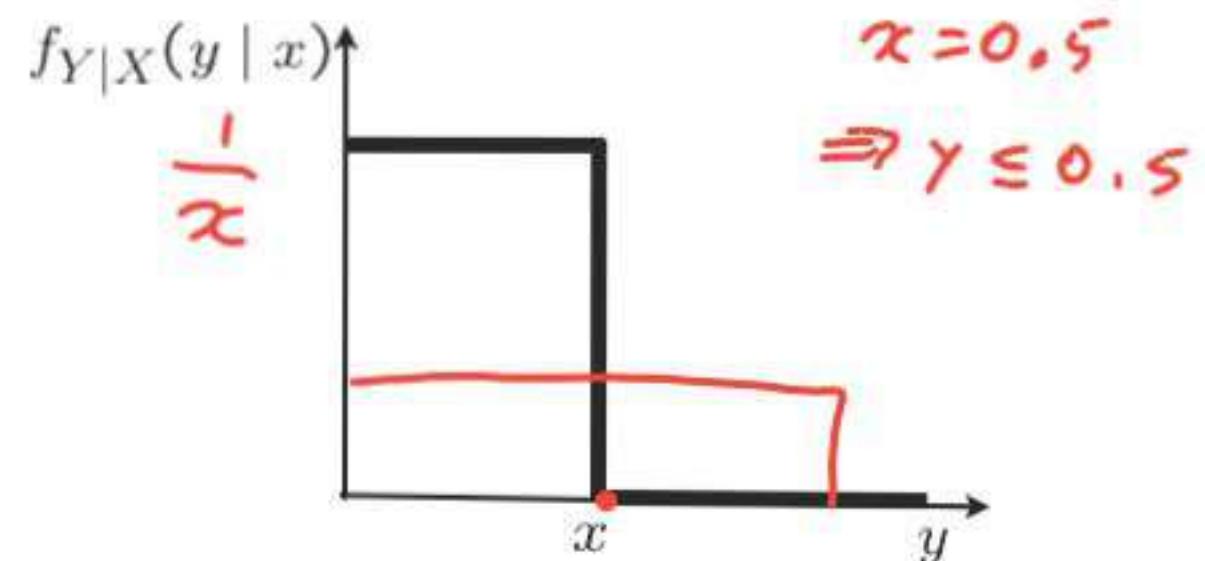
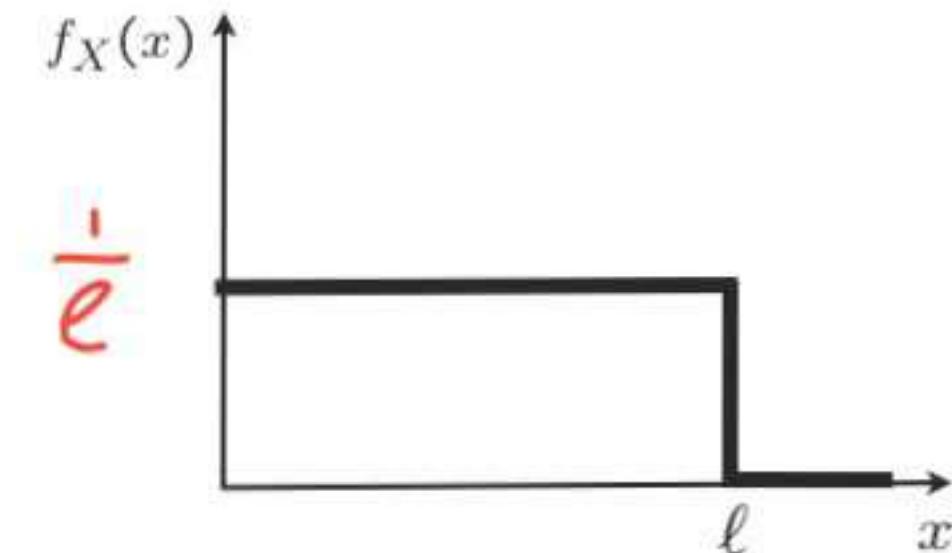
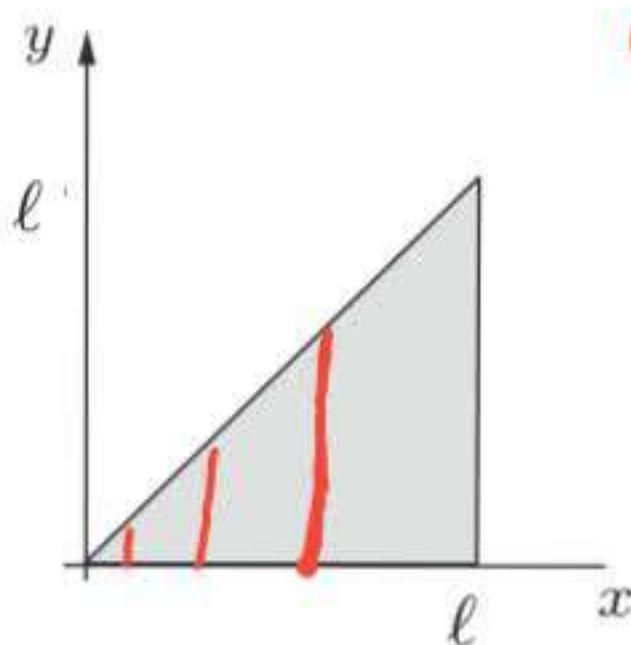
Stick-breaking example



- Break a stick of length ℓ twice
 - first break at X : uniform in $[0, \ell]$
 - second break at Y : uniform in $[0, X]$

$$f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y|x) = \frac{1}{\ell x}$$

$$0 \leq y \leq x \leq \ell$$

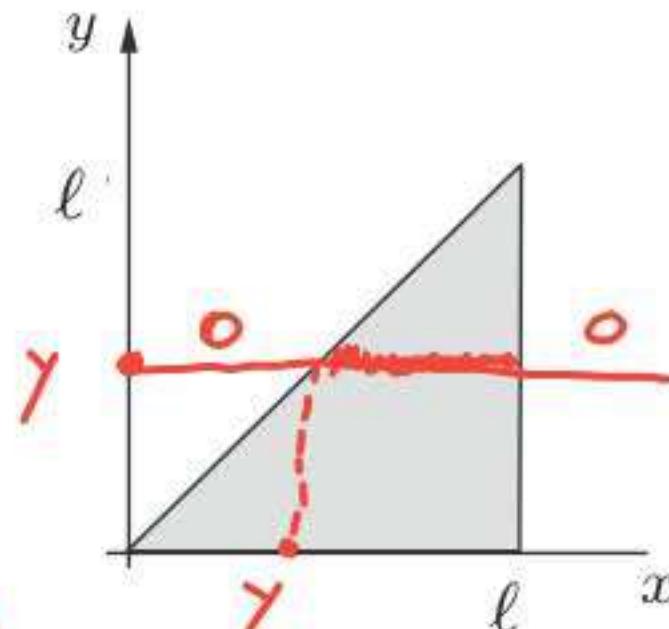


Stick-breaking example

$$f_{X,Y}(x,y) = \frac{1}{\ell x}, \quad 0 \leq y \leq x \leq \ell$$

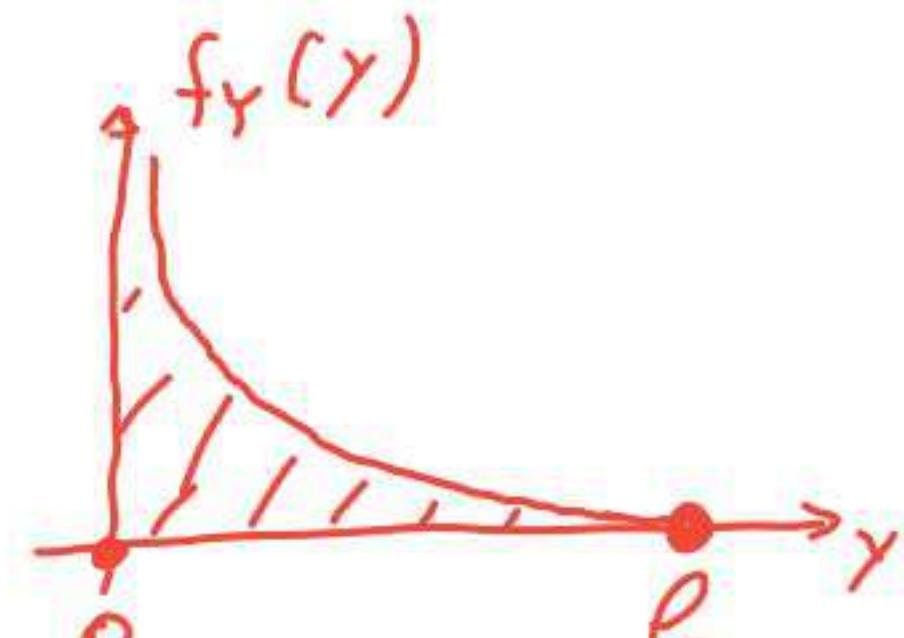
$$f_Y(y) = \int f_{x,y}(x,y) dx = \int_y^{\ell} \frac{1}{\ell x} dx = \frac{1}{\ell} \log\left(\frac{\ell}{y}\right)$$

$$E[Y] = \int_0^{\ell} y \frac{1}{\ell} \log\left(\frac{\ell}{y}\right) dy$$



- Using total expectation theorem:

$$E[Y] = \int_0^{\ell} \frac{1}{\ell} E[Y|x=x] dx = \int_0^{\ell} \left(\frac{1}{\ell} \right) \frac{x}{2} dx = \frac{1}{2} E[X] = \frac{1}{2} \cdot \frac{\ell}{2} = \frac{\ell}{4}$$



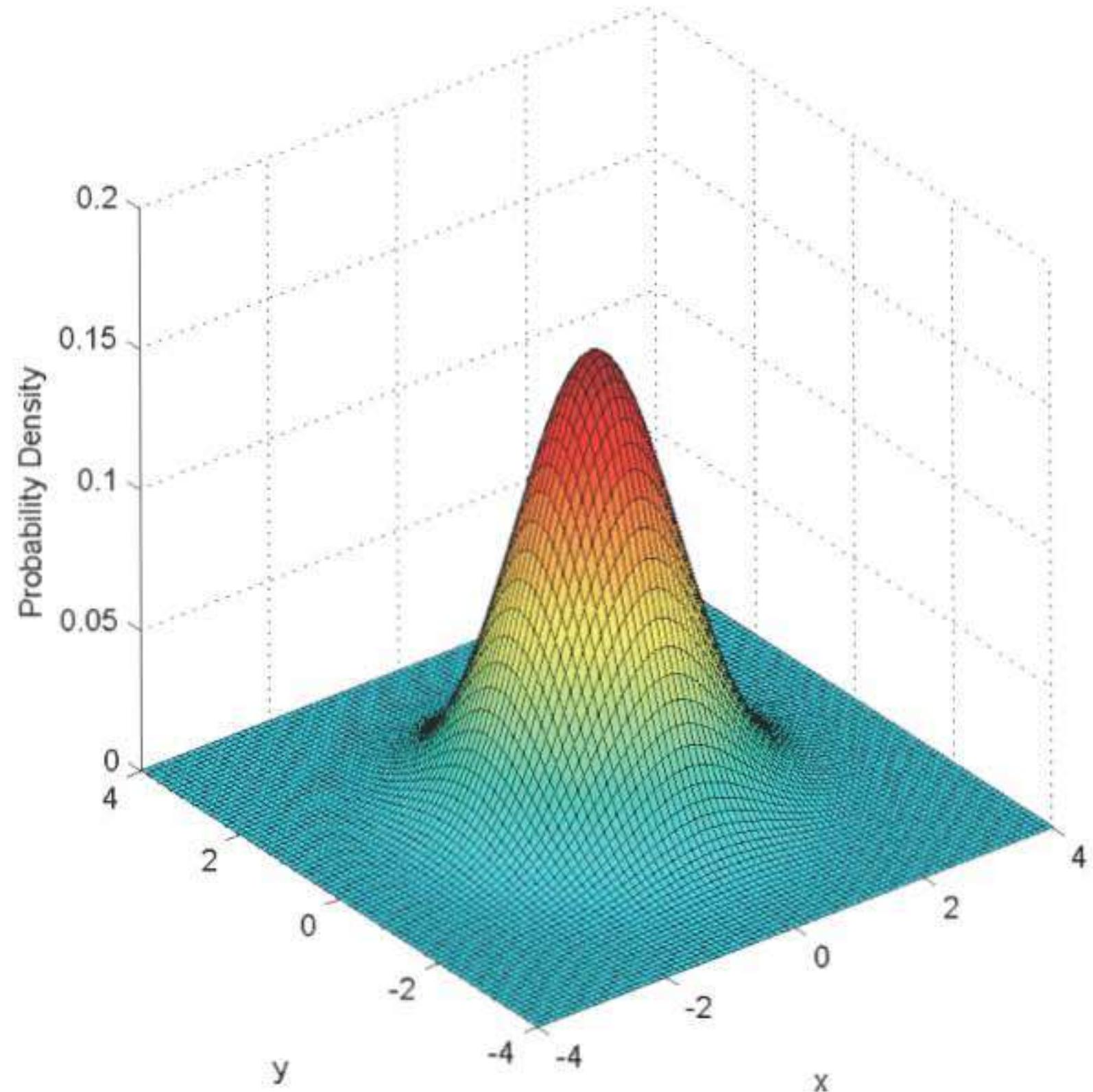
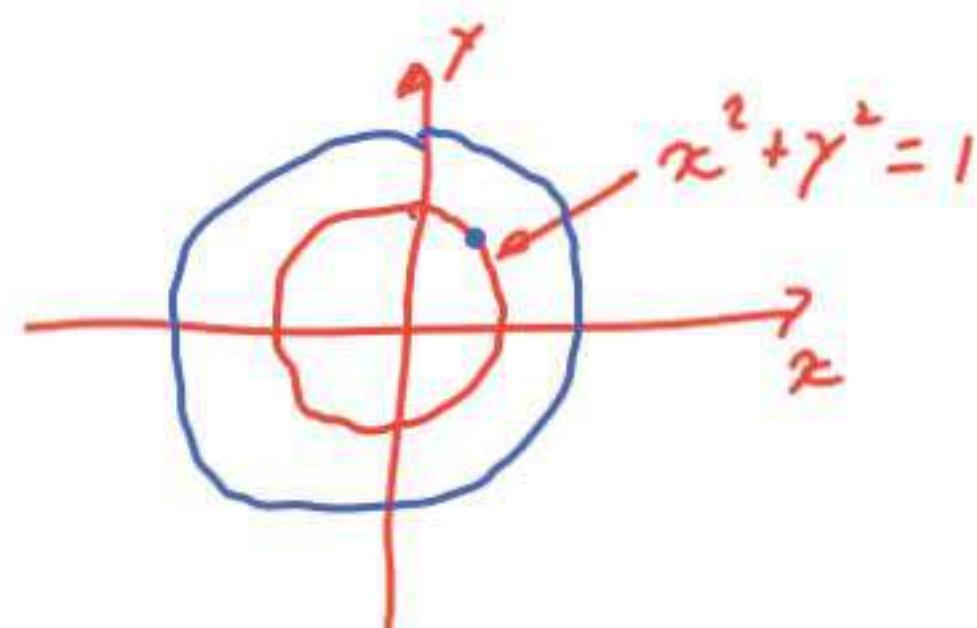
Independent standard normals

$\mu_X = \mu_Y = 0; \sigma_X^2 = \sigma_Y^2 = 1$

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

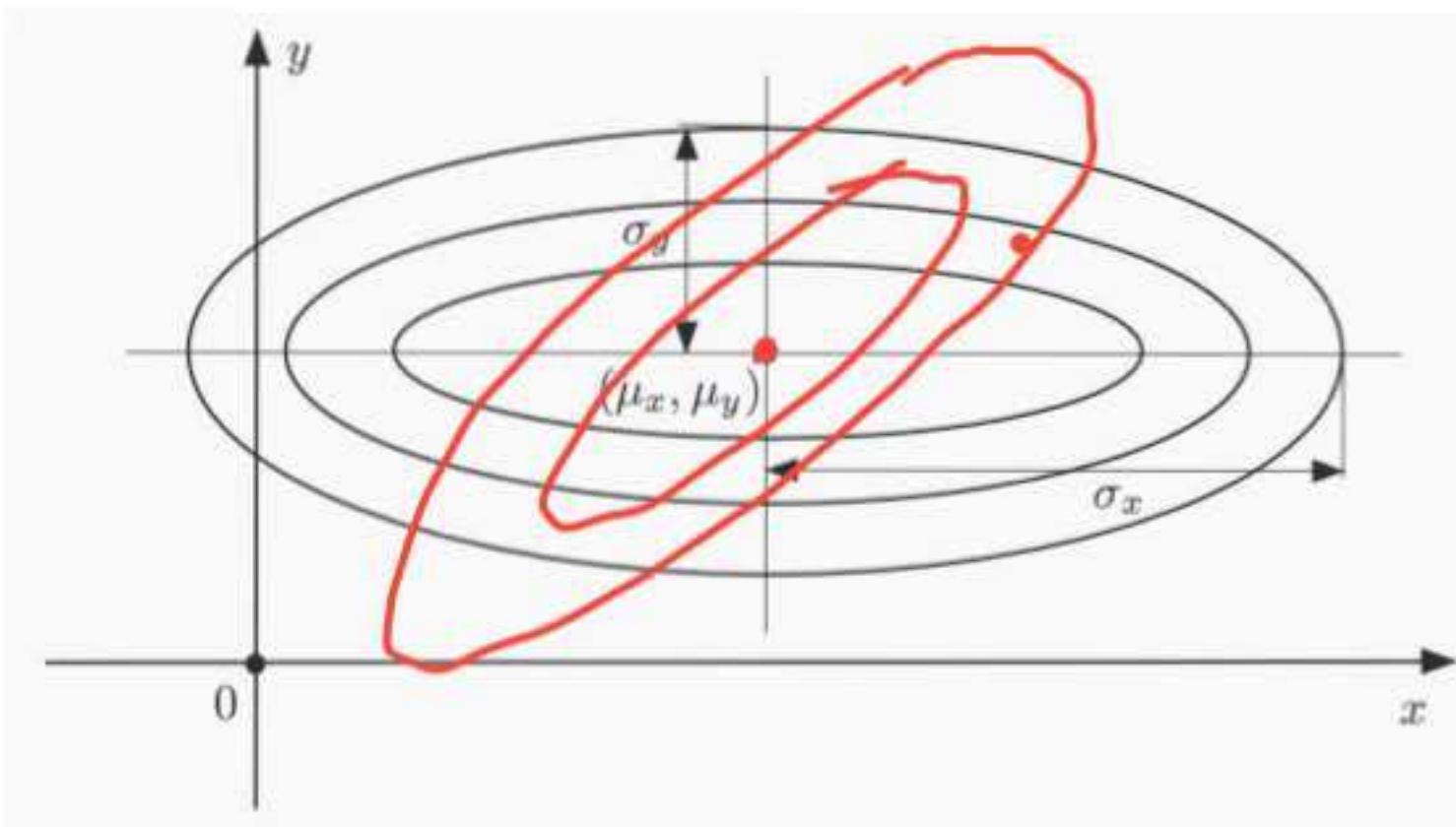
$$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\}$$

$$= \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(x^2 + y^2)\right\}$$

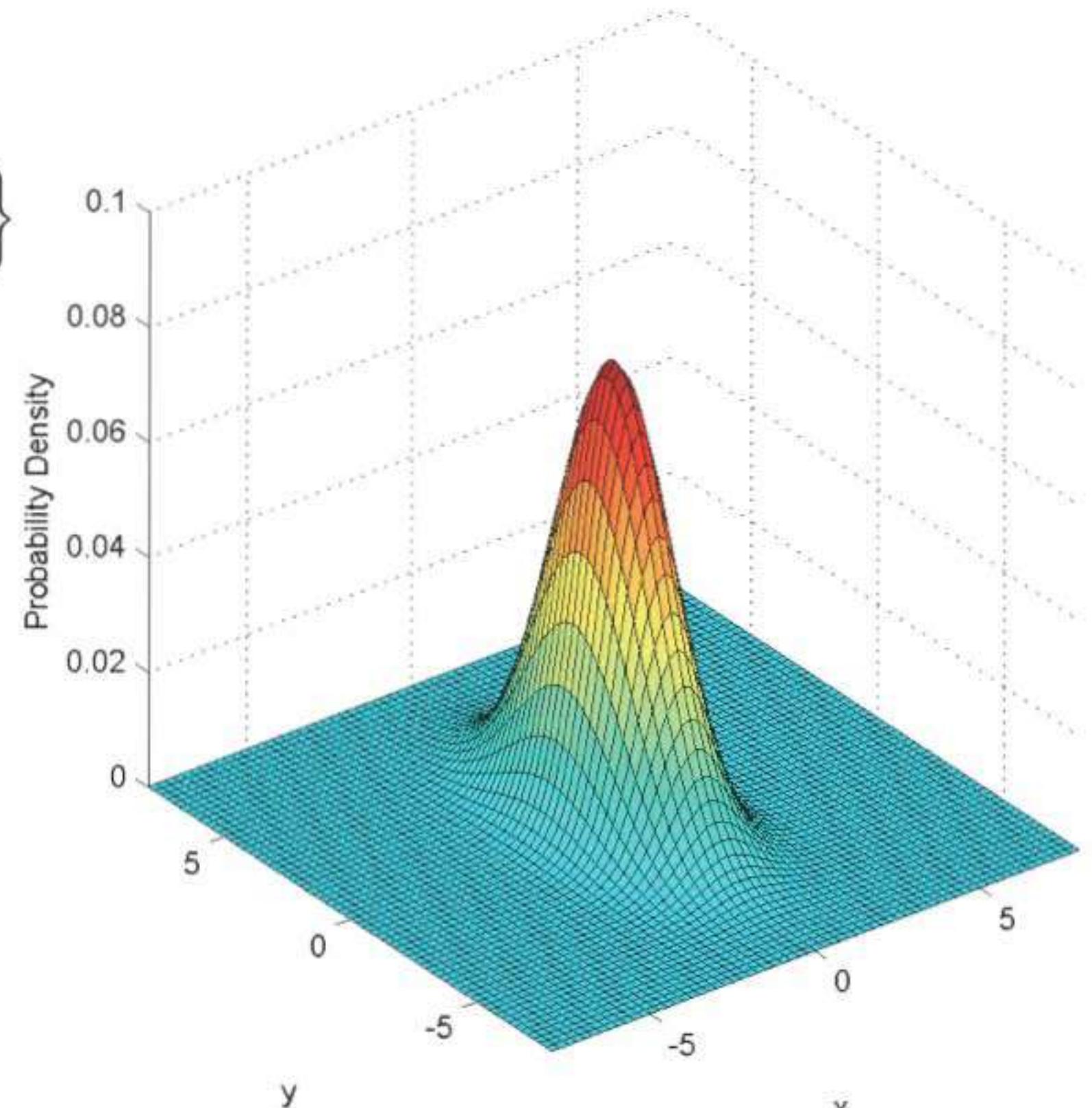


Independent normals

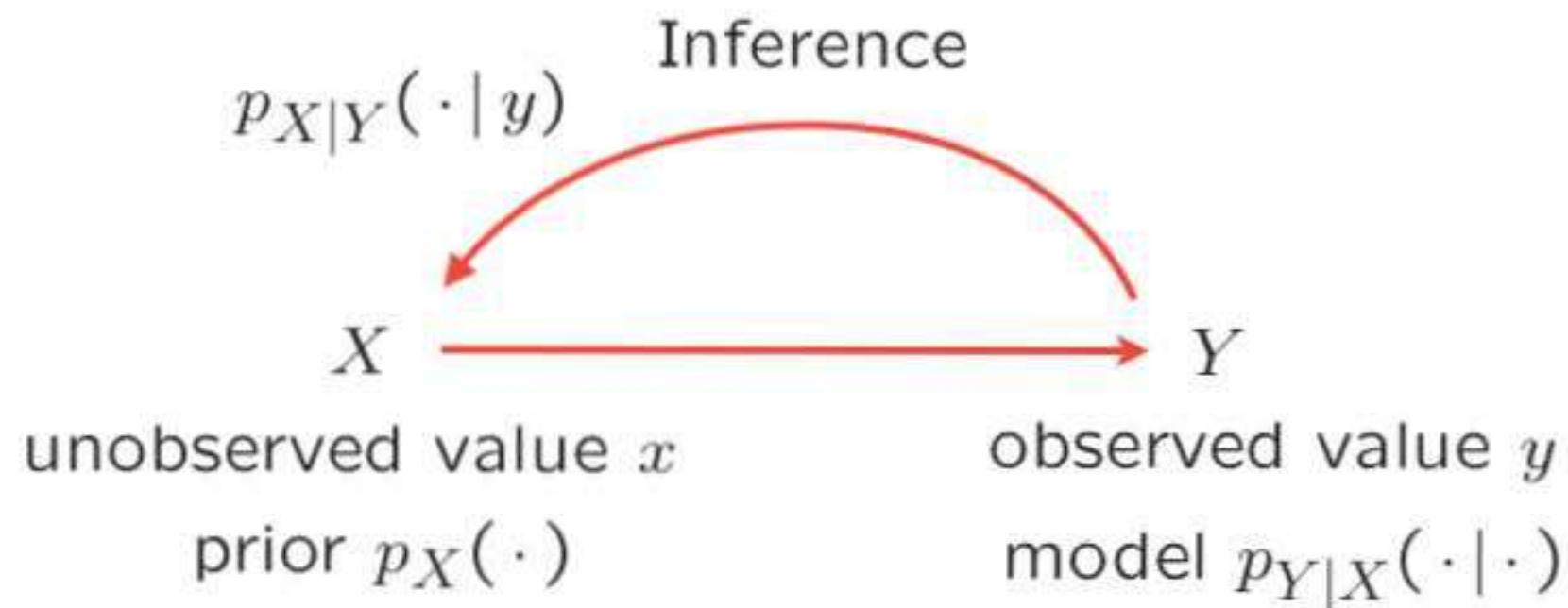
$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$
$$= \frac{1}{2\pi\sigma_x\sigma_y} \exp \left\{ -\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2} \right\}$$



$$\mu_X=\mu_Y=0; \sigma_X^2=1, \sigma_Y^2=4$$



The Bayes rule — a theme with variations



$$\begin{aligned} p_{X,Y}(x,y) &= p_X(x) p_{Y|X}(y|x) \\ &= p_Y(y) p_{X|Y}(x|y) \end{aligned}$$

$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x) f_{Y|X}(y|x) \\ &= f_Y(y) f_{X|Y}(x|y) \end{aligned}$$

$$p_{X|Y}(x|y) = \frac{p_X(x) p_{Y|X}(y|x)}{p_Y(y)}$$

$$f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{f_Y(y)}$$

Posterior $p_Y(y) = \sum_{x'} p_X(x') p_{Y|X}(y|x')$

$$f_Y(y) = \int f_X(x') f_{Y|X}(y|x') dx' \bullet$$

The Bayes rule — one discrete and one continuous random variable

K : discrete

Y : continuous

$$\begin{aligned}
 & P(K=k, y \leq Y \leq y+\delta) \quad \delta > 0, \delta \approx 0 \\
 & = P(K=k) P(y \leq Y \leq y+\delta | K=k) \quad \approx \quad p_K(k) f_{Y|K}(y|k) \\
 & = P(y \leq Y \leq y+\delta) P(K=k | y \leq Y \leq y+\delta) \approx f_Y(y) p_{K|Y}(k|y)
 \end{aligned}$$

$$p_{K|Y}(k|y) = \frac{p_K(k) f_{Y|K}(y|k)}{f_Y(y)}$$

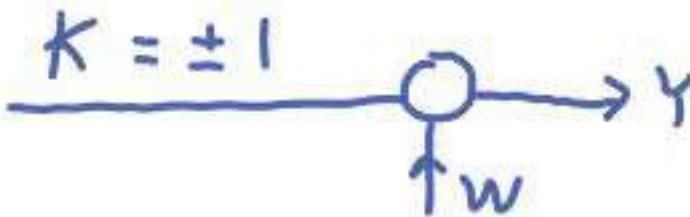
$$f_{Y|K}(y|k) = \frac{f_Y(y) p_{K|Y}(k|y)}{\bullet p_K(k)}$$

$$f_Y(y) = \sum_{k'} p_K(k') f_{Y|K}(y|k')$$

$$p_K(k) = \int f_Y(y') p_{K|Y}(k|y') dy'$$

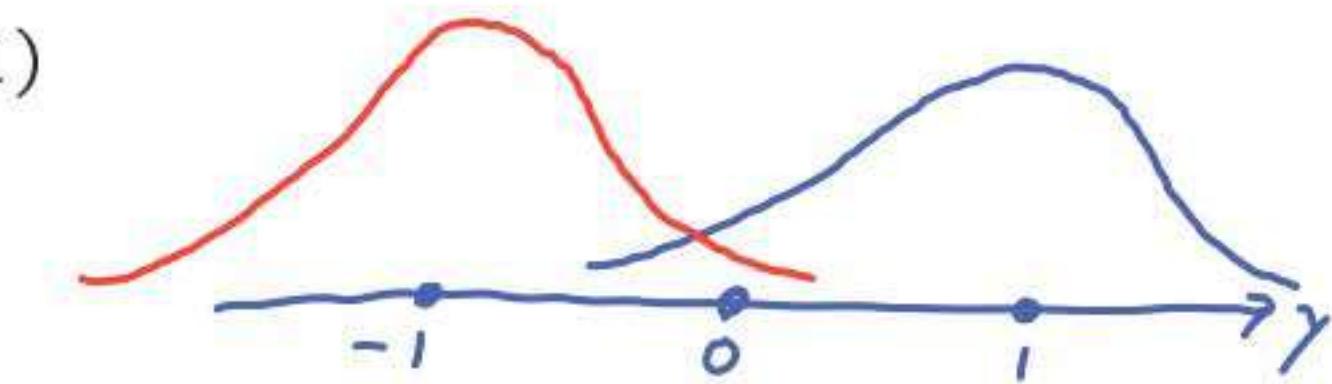
The Bayes rule — discrete unknown, continuous measurement

- unkown K : equally likely to be -1 or $+1$
- measurement Y : $Y = K + W; W \sim \mathcal{N}(0, 1)$



$$Y|K=1 \sim \mathcal{N}(1, 1)$$

$$Y|K=-1 \sim \mathcal{N}(-1, 1)$$



- Probability that $K = 1$, given that $Y = y$? $P_{K|Y}(1|y)$

$$p_K(k) = 1/2 \quad f_{Y|K}(y|k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-k)^2}$$

$k = -1, +1$

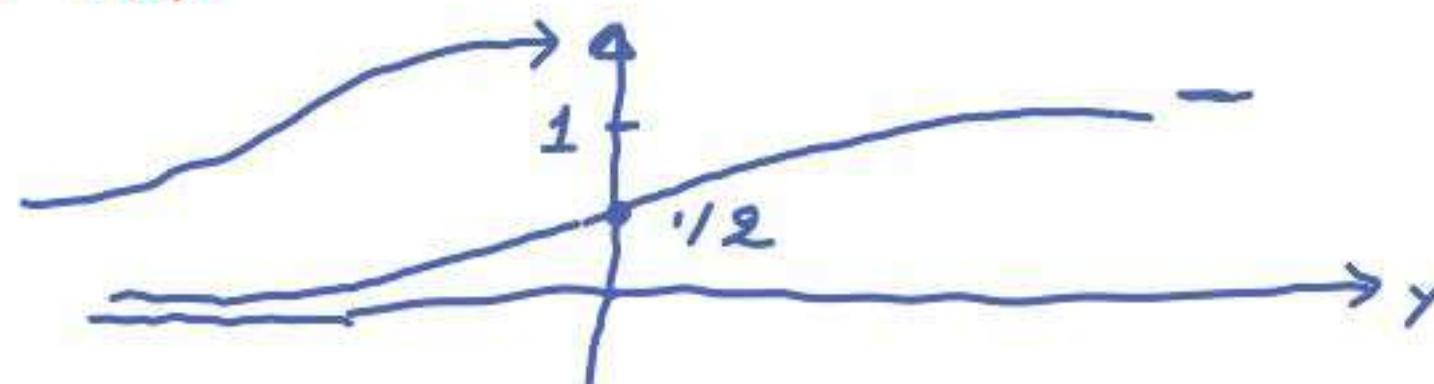
$$f_Y(y) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y+1)^2} + \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-1)^2}$$

$$p_{K|Y}(k|y) = \frac{p_K(k) f_{Y|K}(y|k)}{f_Y(y)}$$

$$f_Y(y) = \sum_{k'} p_K(k') f_{Y|K}(y|k')$$

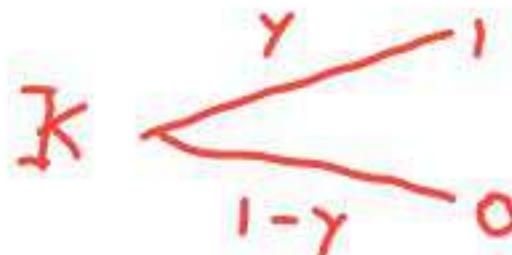
$$p_{K|Y}(1|y) = \frac{1}{1 + e^{-2y}}$$

algebra

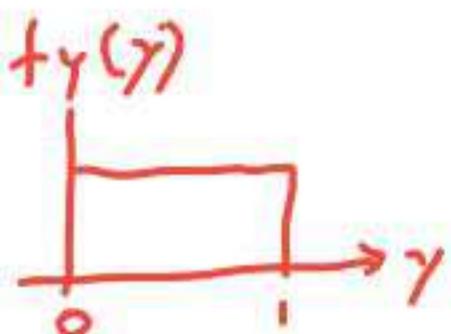


The Bayes rule — continuous unknown, discrete measurement

- measurement K : Bernoulli with parameter Y



- unkown Y : uniform on $[0, 1]$



- Distribution of Y given that $K = 1$?

$$f_{Y|K}(y|1)$$

$$f_Y(y) = \begin{cases} 1 & y \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

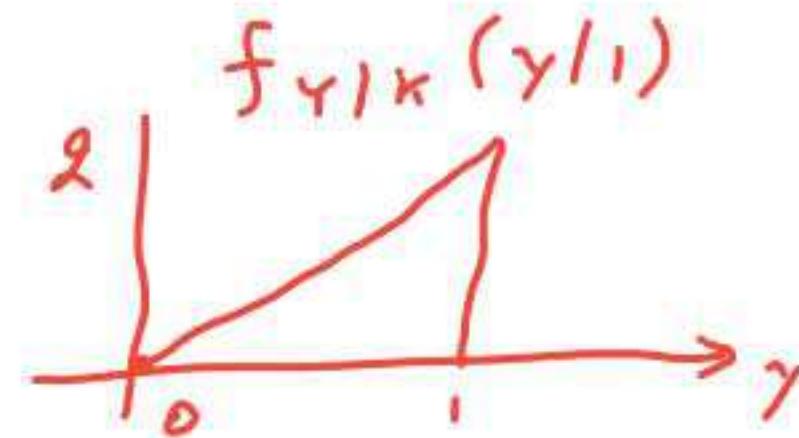
$$p_{K|Y}(1|y) =$$

$$p_K(1) = \int_0^1 1 \cdot y \, dy = \frac{y^2}{2} \Big|_0^1 = \frac{1}{2}$$

$$f_{Y|K}(y|1) = \frac{1 \cdot y}{1/2} = 2y, \quad y \in [0, 1]$$

$$f_{Y|K}(y|k) = \frac{f_Y(y) p_{K|Y}(k|y)}{p_K(k)}$$

$$p_K(k) = \int f_Y(y') p_{K|Y}(k|y') dy'$$



MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

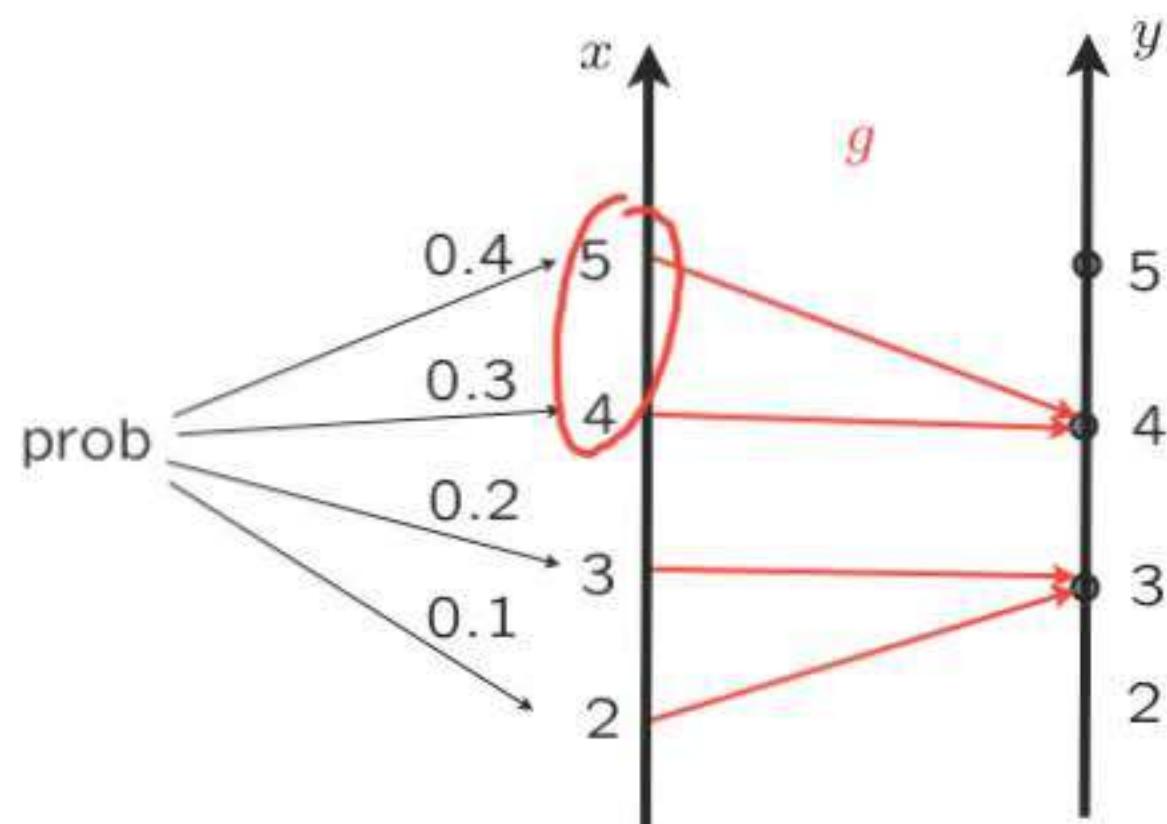
For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 11: Derived distributions

- Given the distribution of X ,
find the distribution of $Y = g(X)$
 - the discrete case
 - the continuous case
 - general approach, using CDFs
 - the linear case: $Y = aX + b$
 - general formula when g is monotonic
- Given the (joint) distribution of X and Y ,
find the distribution of $Z = g(X, Y)$

Derived distributions — the discrete case

$$Y = g(X)$$



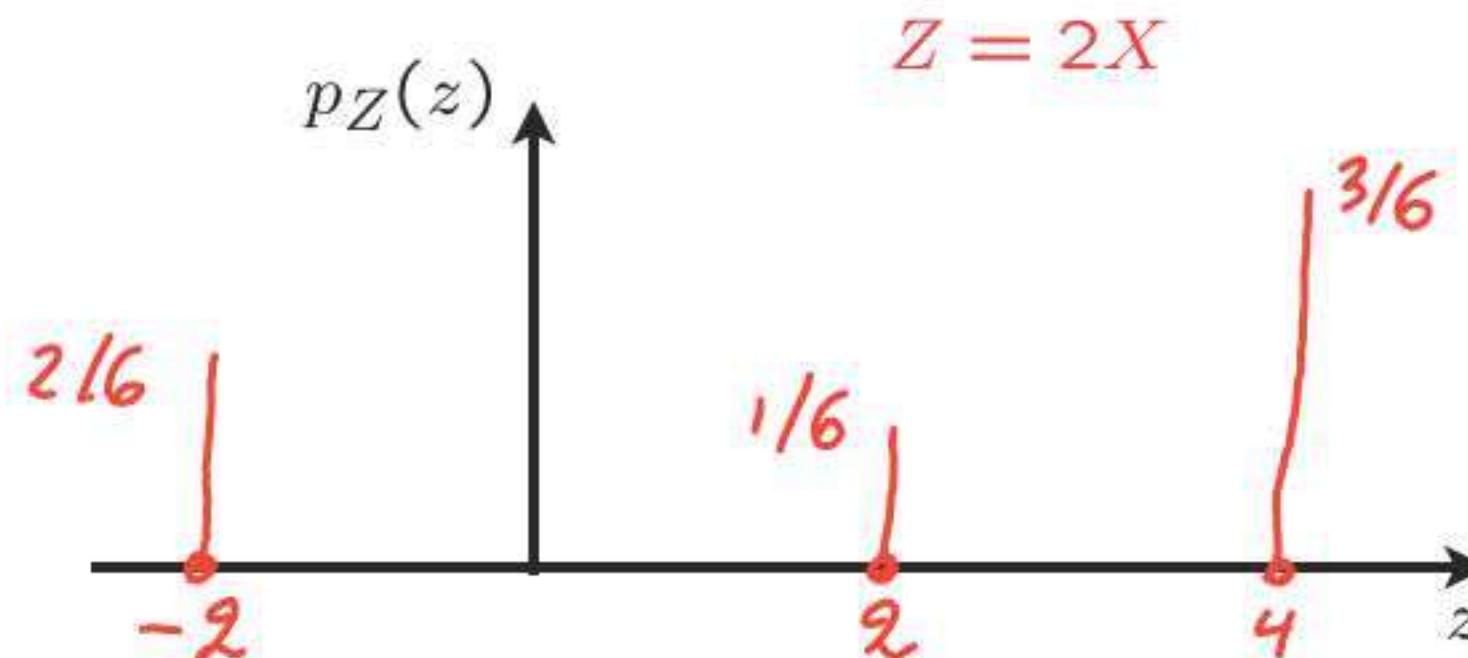
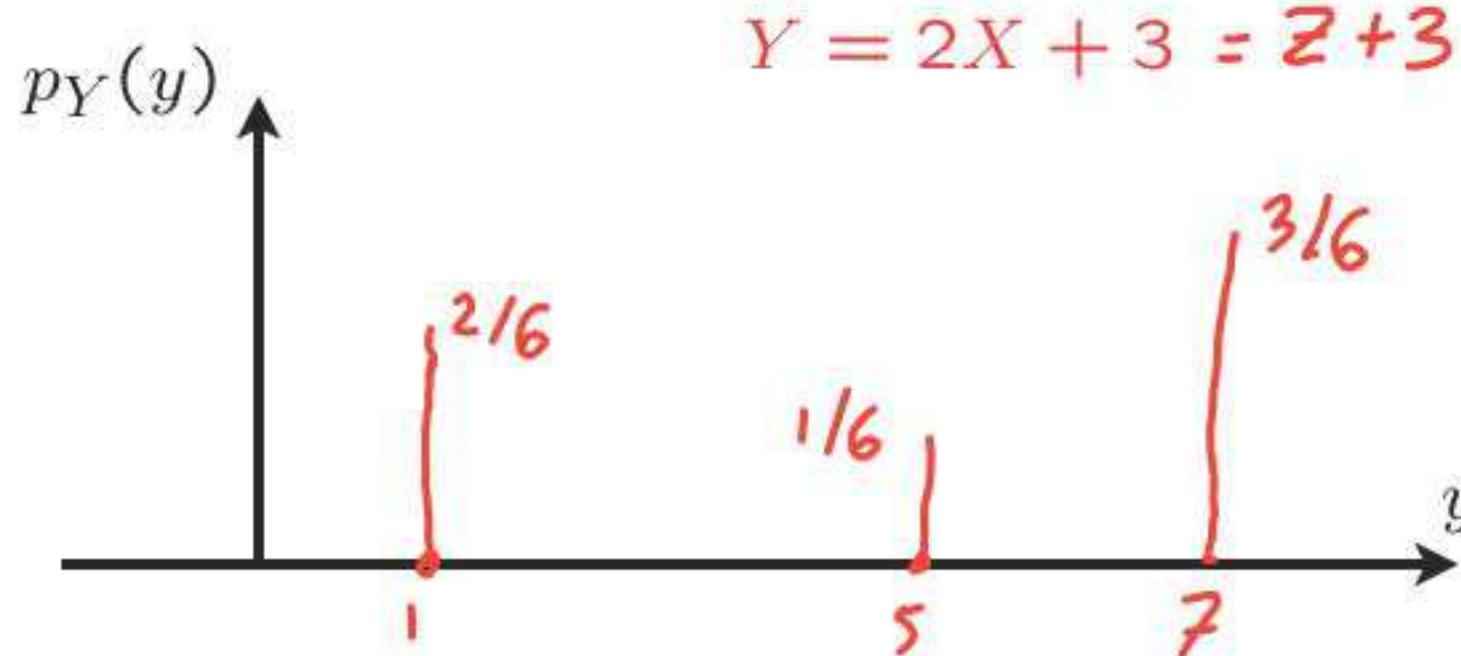
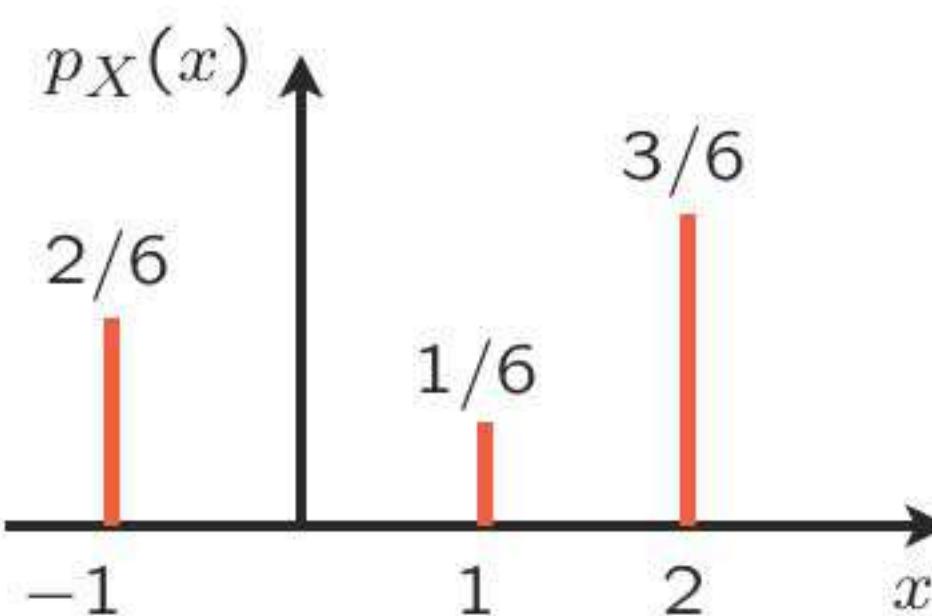
$$P_Y(4) = P(Y=4)$$

$$= P(X=4) + P(X=5)$$

$$= p_X(4) + p_X(5) = 0.3 + 0.4$$

$$\begin{aligned} p_Y(y) &= P(g(X) = y) \\ &= \sum_{x: g(x)=y} p_X(x) \end{aligned}$$

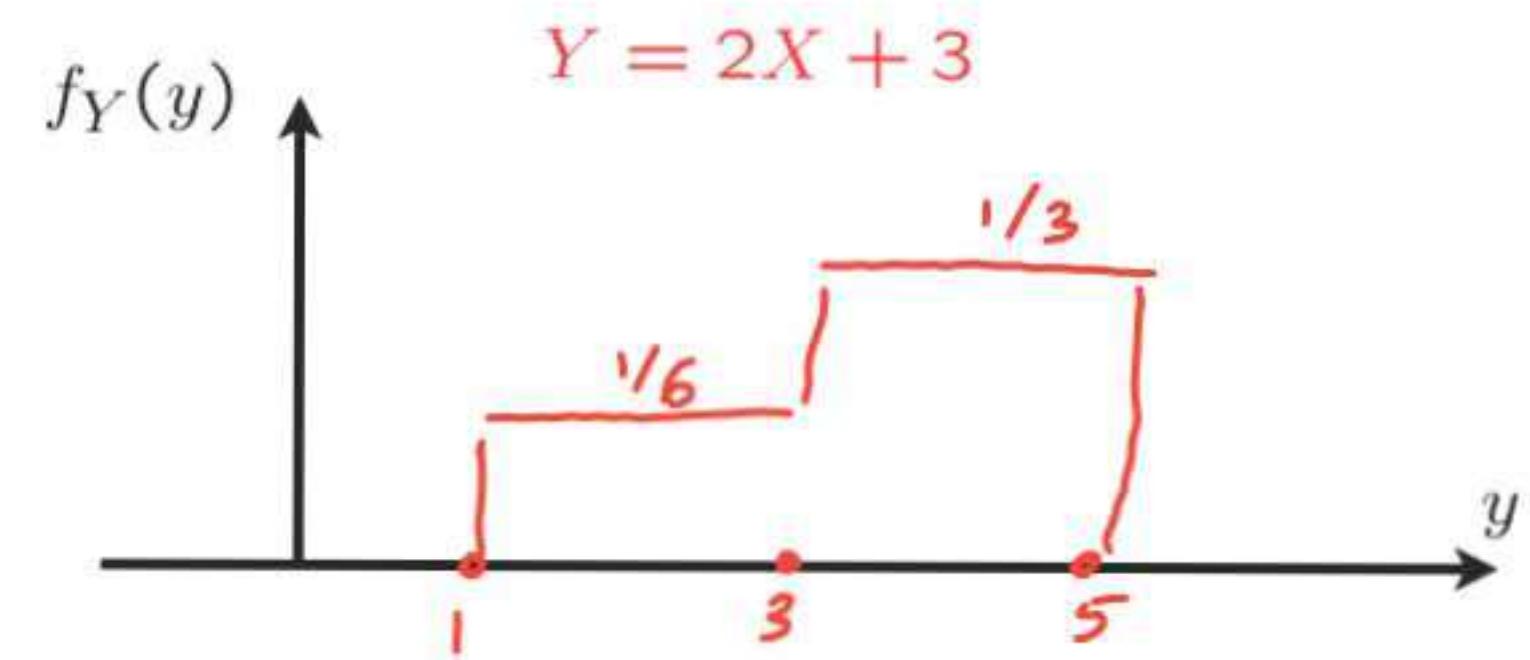
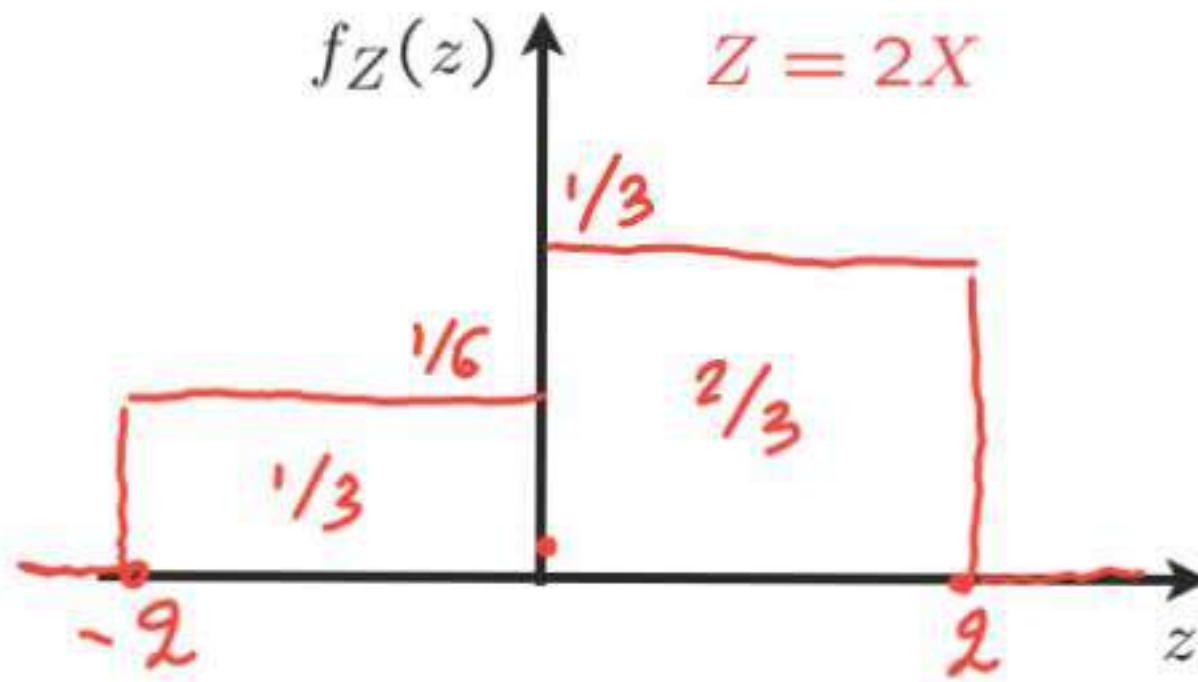
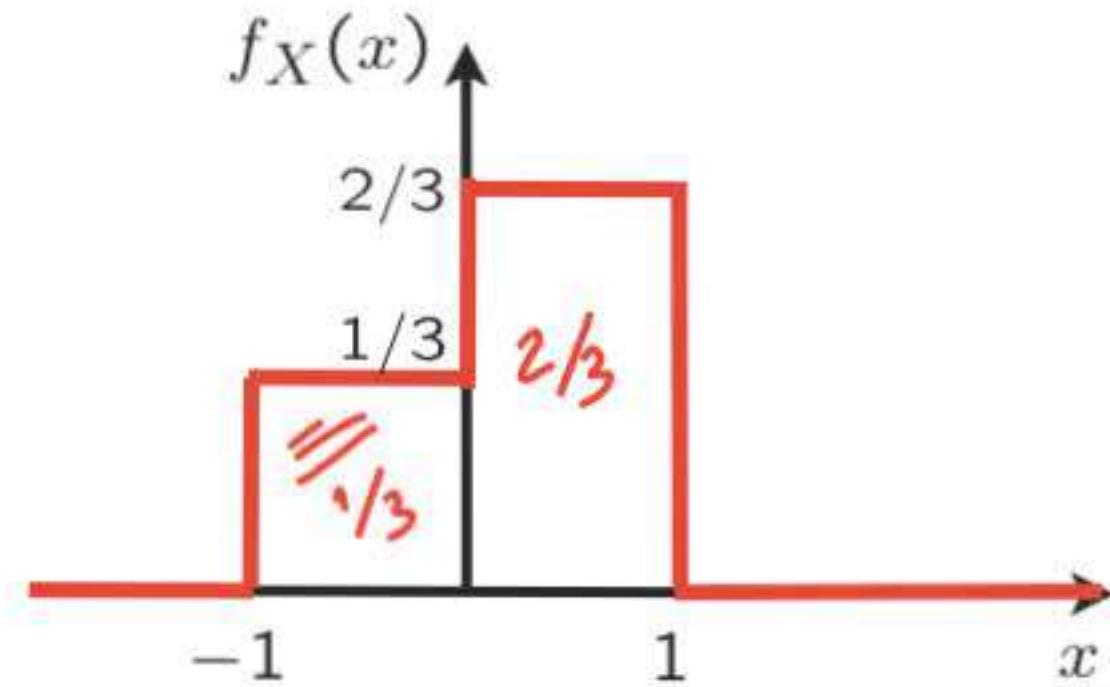
A linear function of a discrete r.v.



$$\begin{aligned}
 P_Y(y) &= P(Y=y) = P(2X+3=y) \\
 &= P\left(X=\frac{y-3}{2}\right) = P_X\left(\frac{y-3}{2}\right)
 \end{aligned}$$

$$Y = aX + b : \quad p_Y(y) = p_X\left(\frac{y-b}{a}\right)$$

A linear function of a continuous r.v.



A linear function of a continuous r.v.

$$Y = aX + b$$

$$a > 0$$

$$\mathbb{P}(Y=y) = \mathbb{P}(aX+b=y) = \mathbb{P}\left(X=\frac{y-b}{a}\right)$$

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX+b \leq y)$$

$$= \mathbb{P}\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a}$$

$$a < 0$$

$$= \mathbb{P}\left(X \geq \frac{y-b}{a}\right)$$

$$= 1 - \mathbb{P}\left(X \leq \frac{y-b}{a}\right)$$

$$= 1 - F_X\left(\frac{y-b}{a}\right)$$

$$f_Y(y) = -f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a}$$

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

$$p_Y(y) = p_X\left(\frac{y-b}{a}\right) \cdot$$

A linear function of a normal r.v. is normal

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

$$Y = aX + b, \quad a \neq 0$$

$$\begin{aligned} f_Y(y) &= \frac{1}{|a|} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{y-b}{a}-\mu\right)^2/2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi}\sigma|a|} e^{-\frac{(y-b-a\mu)^2}{2\sigma^2 a^2}} \end{aligned}$$

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

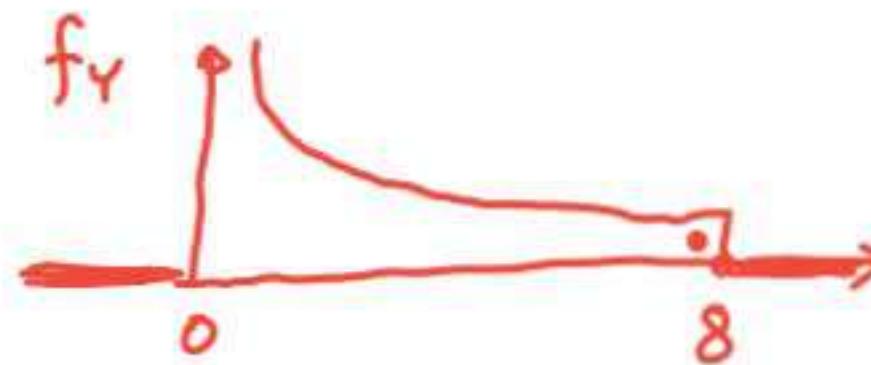
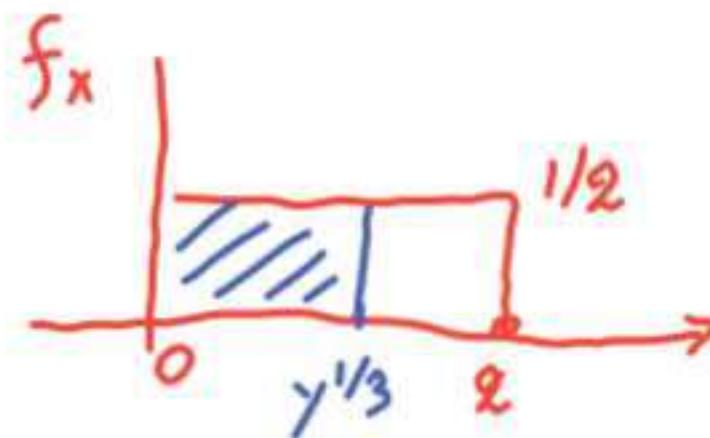
If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

A general function $g(X)$ of a continuous r.v.

- Two-step procedure:

- Find the CDF of Y : $F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(g(X) \leq y)$
- Differentiate: $f_Y(y) = \frac{dF_Y}{dy}(y)$

Example: $Y = X^3$; X uniform on $[0, 2]$



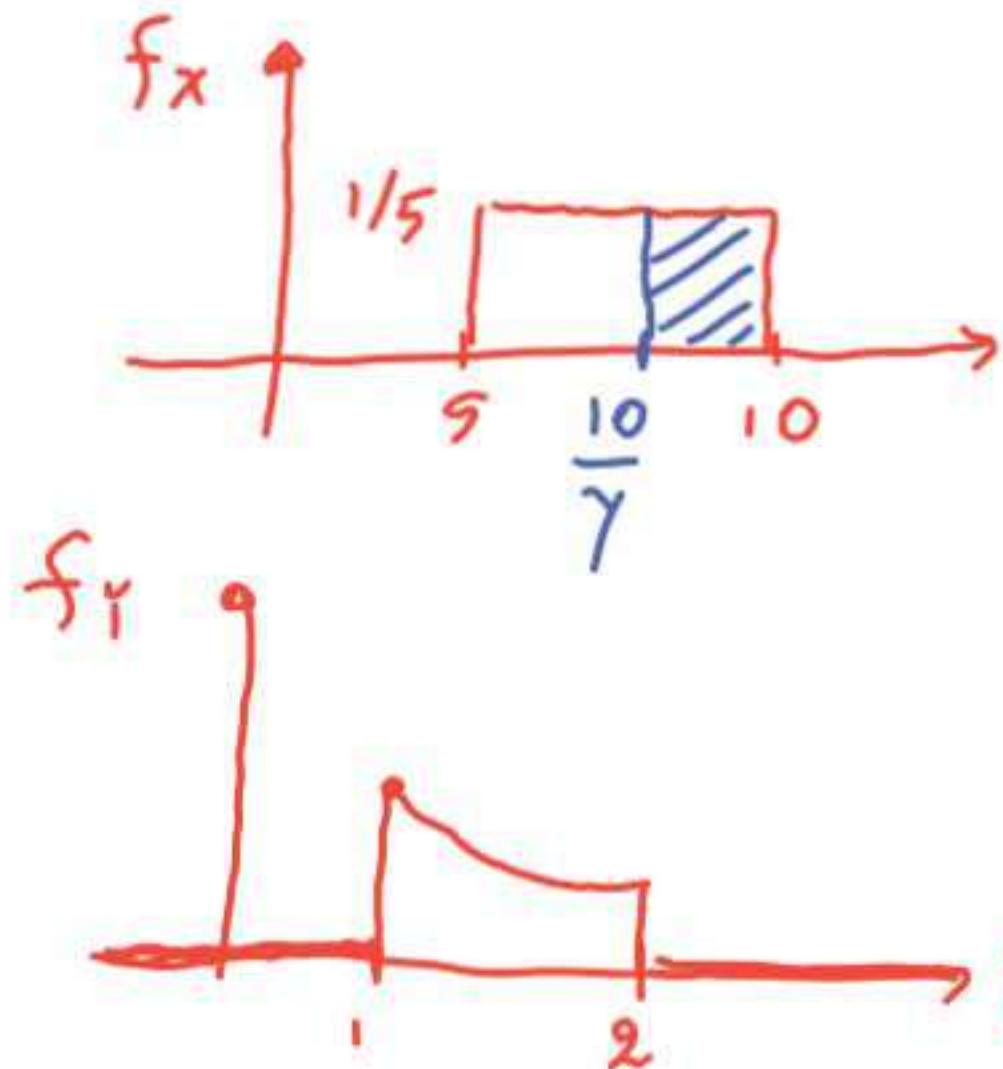
$$0 \leq y \leq 8$$

$$F_Y(y) = P(Y \leq y) = P(X^3 \leq y) = P(X \leq y^{1/3}) = \frac{1}{2} y^{1/3}$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{2} \cdot \frac{1}{3} y^{-2/3} = \frac{1}{6} \cdot \frac{1}{y^{2/3}}$$

Example: $Y = a/X$

- You go to the gym and set the speed X of the treadmill to a number between 5 and 10 km/hr (with a uniform distribution). Find the PDF of the time it takes to run 10km.



$$\text{time} = Y = \frac{10}{X} \quad 1 \leq Y \leq 2$$

$$F_Y(y) = P(Y \leq y) = P\left(\frac{10}{X} \leq y\right)$$

$$= P\left(X \geq \frac{10}{y}\right) = \frac{1}{5} \left(10 - \frac{10}{y}\right)$$

$$f_Y(y) = \frac{1}{5} \frac{(-10)}{-y^2} = \frac{2}{y^2}, \quad 1 \leq y \leq 2$$

$$= 0, \quad \text{otherwise}$$

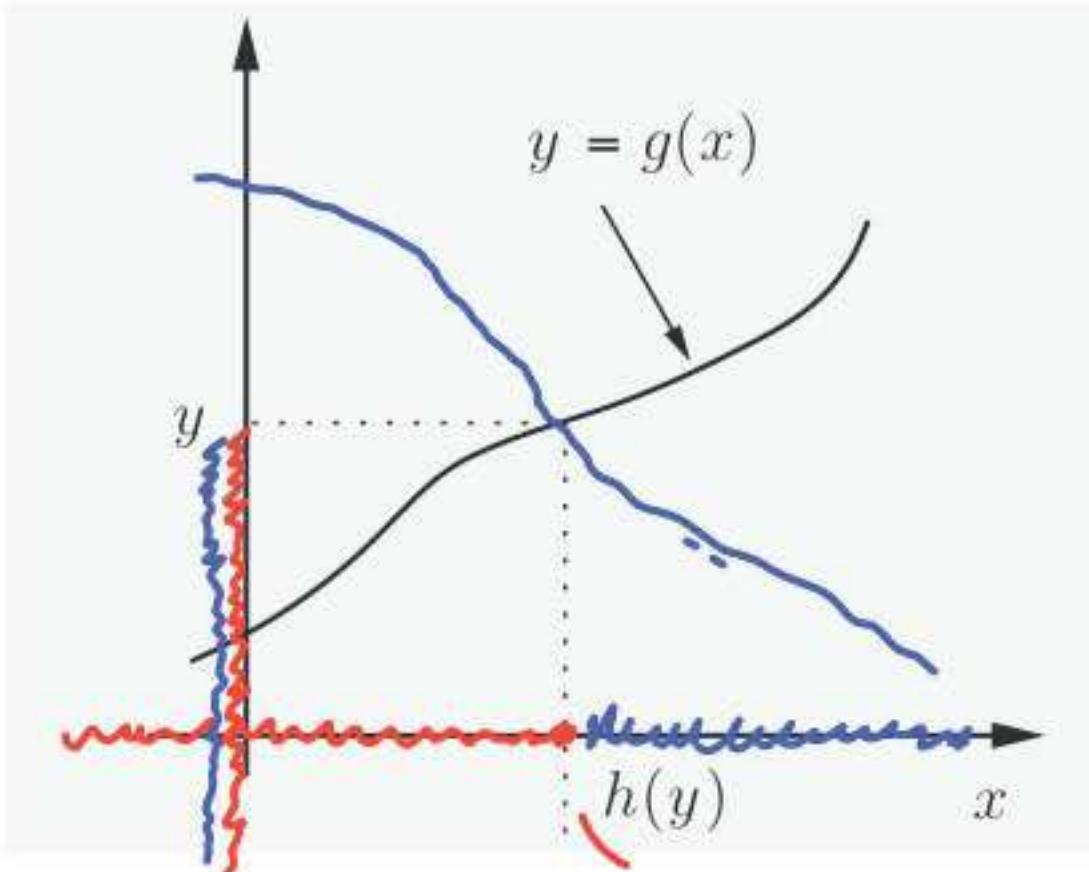
A general formula for the PDF of $Y = g(X)$ when g is monotonic

$$x^3 \frac{\alpha}{x}$$

~~decreasing $x < x' \Rightarrow g(x) < g(x')$~~

Assume g strictly increasing

and differentiable



inverse function $h \rightarrow$ decreasing

$$F_Y(y) = P(Y \leq y) = P(X \leq h(y)) = F_X(h(y))$$

$$f_Y(y) = f_X(h(y)) \left| \frac{d h}{d y}(y) \right|$$

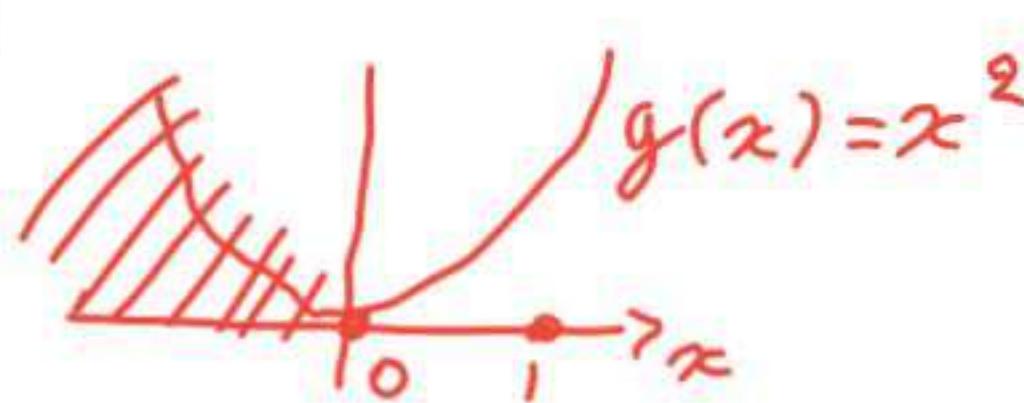
$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X \geq h(y)) \\ &= 1 - P(X \leq h(y)) = 1 - F_X(h(y)) \end{aligned}$$

$$f_Y(y) = f_X(h(y)) \left| \frac{d h}{d y}(y) \right|$$

$$f_Y(y) = f_X(h(y)) \left| \frac{d h}{d y}(y) \right|$$

Example: $Y = X^2$; X uniform on $[0, 1]$

$$f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy}(y) \right|$$

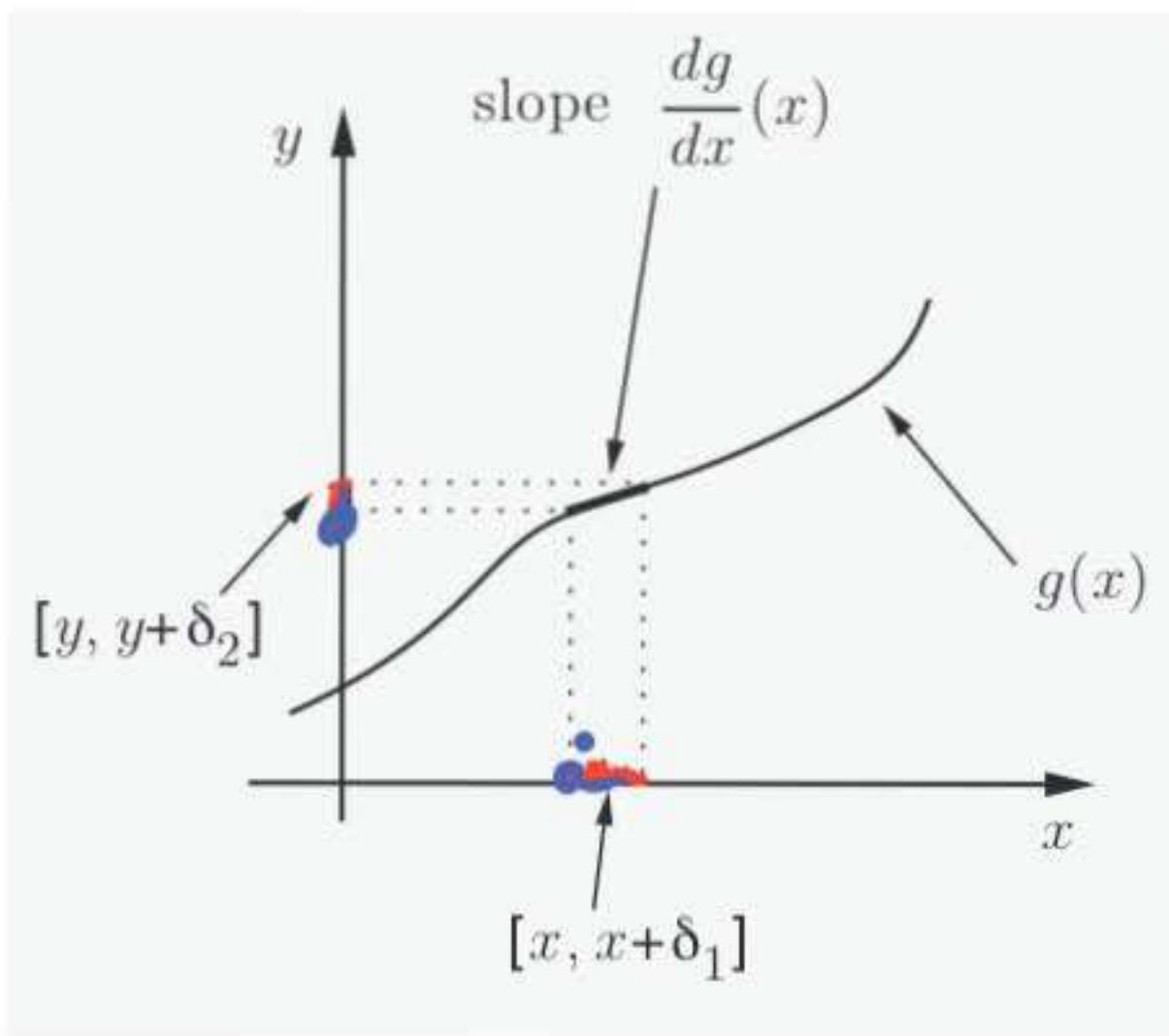


$$y = x^2 \Leftrightarrow x = \sqrt{y} \quad h(y) = \sqrt{y}$$

$$f_Y(y) = \frac{1}{2\sqrt{y}}$$

$$0 \leq y \leq 1$$

An intuitive explanation for the monotonic case



$$Y = g(X)$$

$$X = h(Y)$$

$$\delta_2 \approx \delta_1 \frac{\partial g}{\partial X}(x)$$

$$\delta_1 \approx \delta_2 \cdot \frac{\partial h}{\partial Y}(y) \quad \text{④}$$

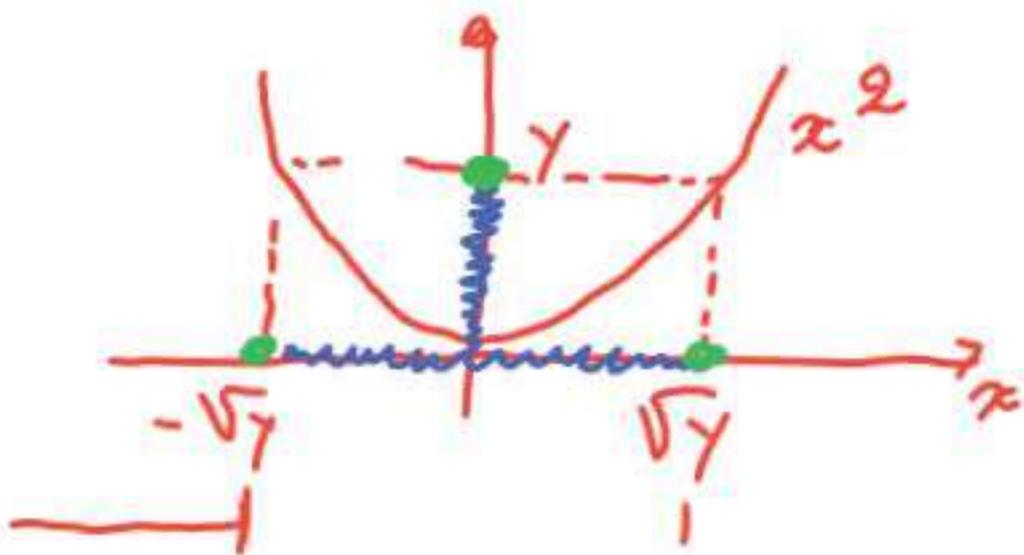
$$f_Y(y) \delta_2 \approx P(Y \leq Y \leq y + \delta_2) = P(X \leq X \leq x + \delta_1)$$

$$\approx f_X(x) \delta_1 \approx f_X(x) \delta_2 \frac{\partial h}{\partial Y}(y)$$

$$f_Y(y) = f_X(x) \frac{\partial h}{\partial Y}(y)$$

$$= f_X(h(y)) \frac{\partial h}{\partial Y}(y)$$

A nonmonotonic example: $Y = X^2$



- The discrete case:

$$p_Y(9) = P(X=3) + P(X=-3)$$

$$p_Y(y) = P_x(\sqrt{y}) + P_x(-\sqrt{y})$$

- The continuous case: $y \geq 0$

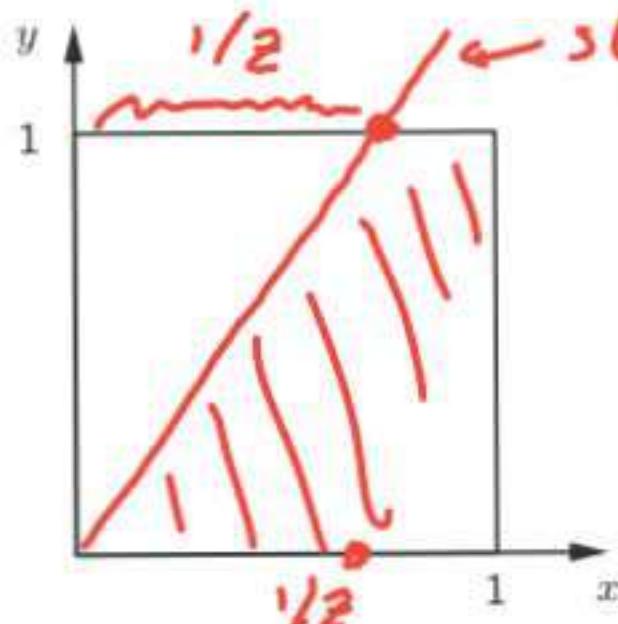
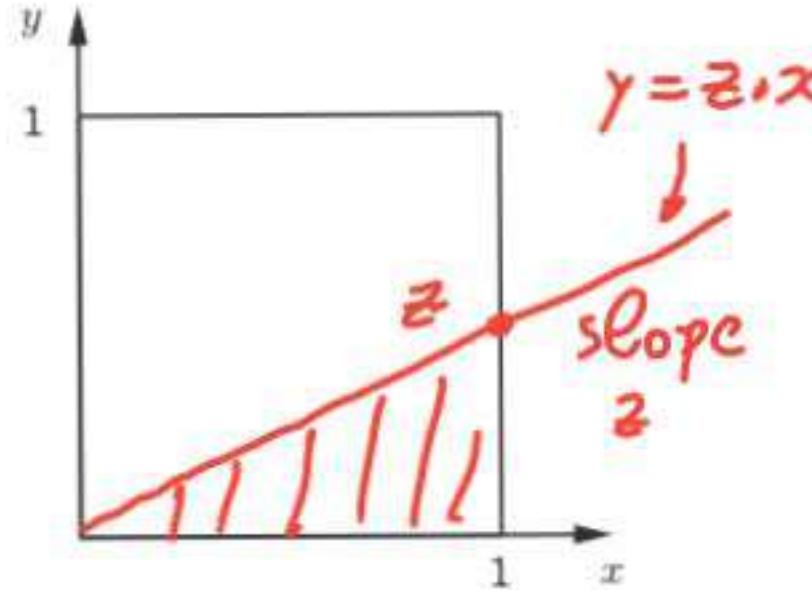
$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(|X| \leq \sqrt{y}) = P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \end{aligned}$$

$$f_Y(y) = f_X(\sqrt{y}) \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \frac{\cancel{-1}}{2\sqrt{y}}$$

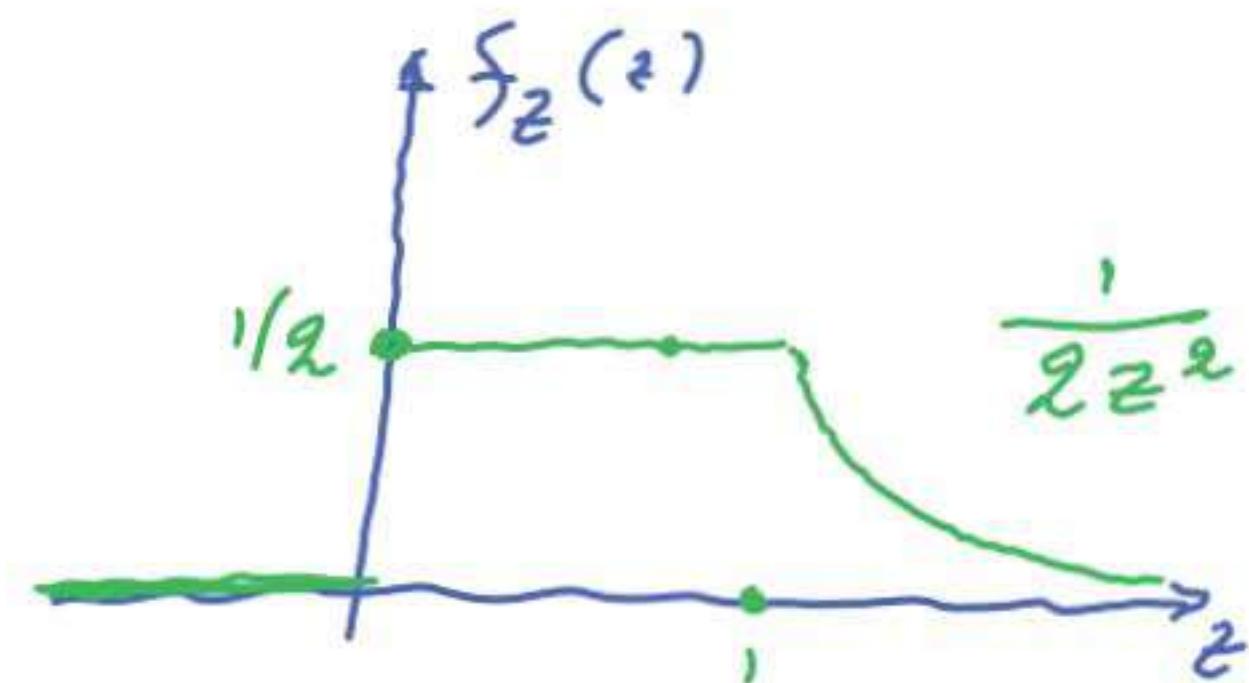
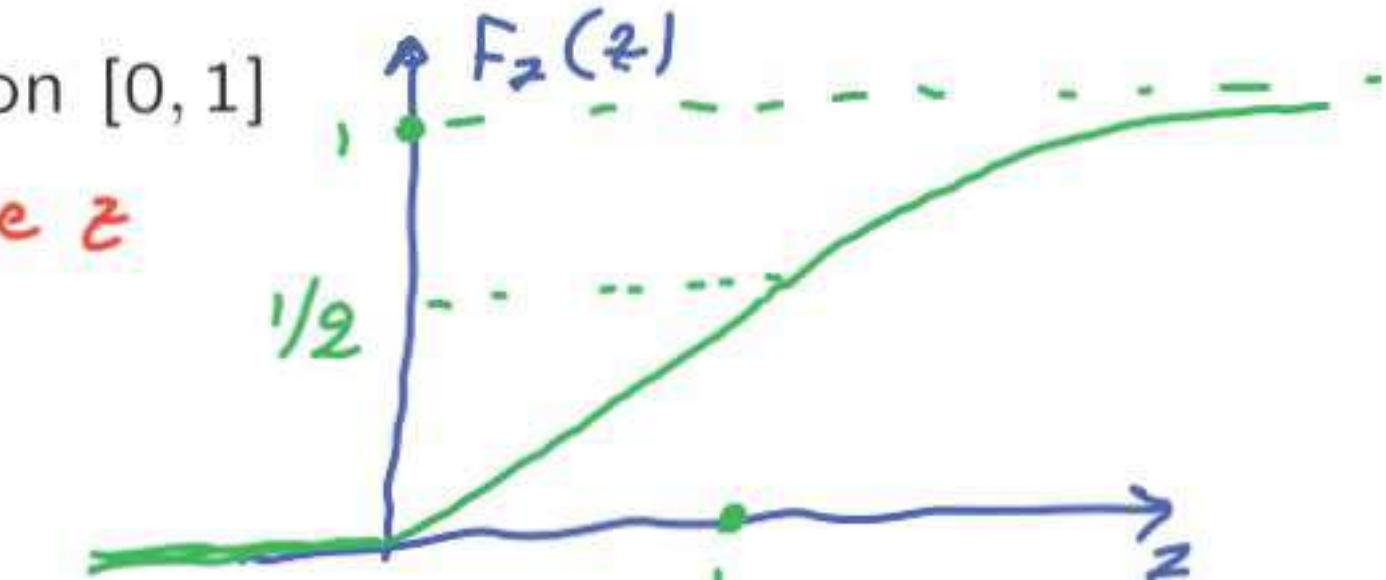
A function of multiple r.v.'s: $Z = g(X, Y)$

- Same methodology: find CDF of Z

- Let $Z = Y/X$; X, Y independent, uniform on $[0, 1]$



$$\begin{aligned}
 F_Z(z) &= P\left(\frac{Y}{X} \leq z\right) = 0, \quad z < 0 \\
 &= \frac{1}{2} \cdot z, \quad 0 \leq z \leq 1 \\
 &= 1 - \frac{1}{2z}, \quad z > 1
 \end{aligned}$$



MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 12: Sums of independent random variables; Covariance and correlation

- The PMF/PDF of $X + Y$ (X and Y independent)
 - the discrete case
 - the continuous case
 - the mechanics
 - the sum of independent normals
- Covariance and correlation
 - definitions
 - mathematical properties
 - interpretation

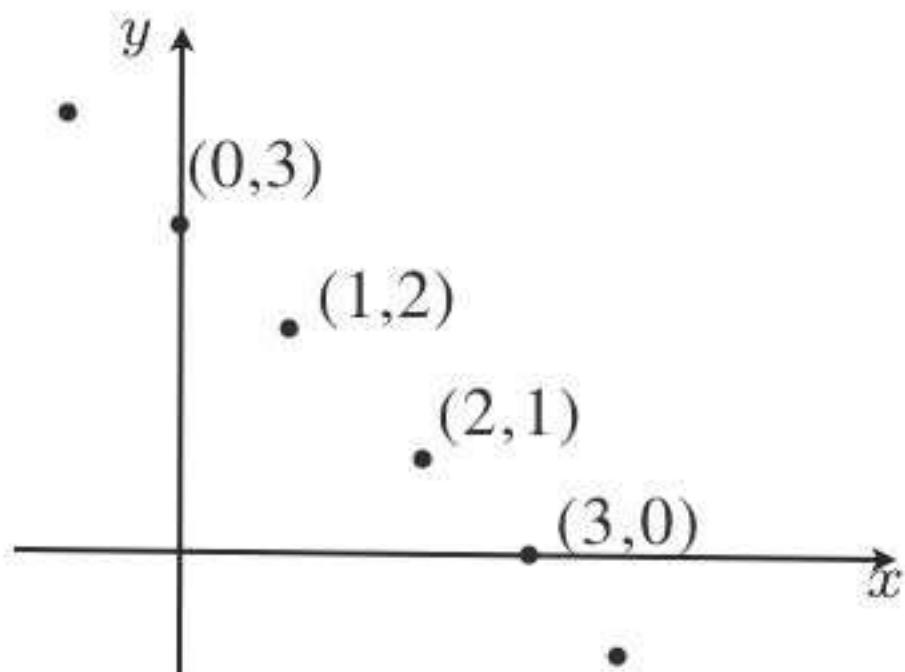
The distribution of $X + Y$: the discrete case

- $Z = X + Y$; X, Y independent, discrete

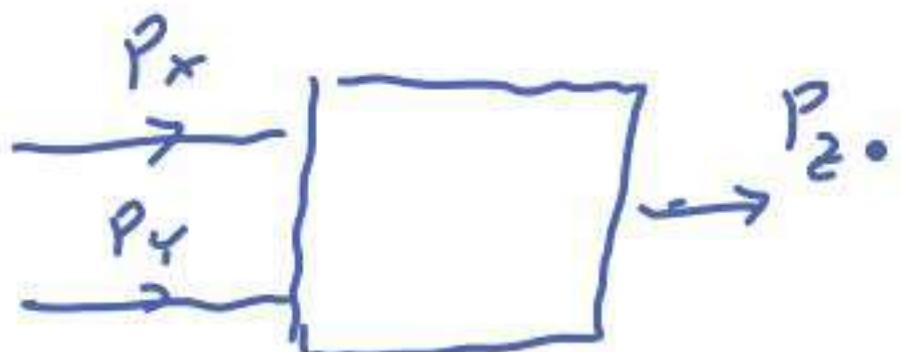
$g(x,y)$

known PMFs

$$p_Z(3) = \dots + P(X=0, Y=3) + P(X=1, Y=2) + \dots \\ = \dots + P_X(0) P_Y(3) + P_X(1) P_Y(2) + \dots$$



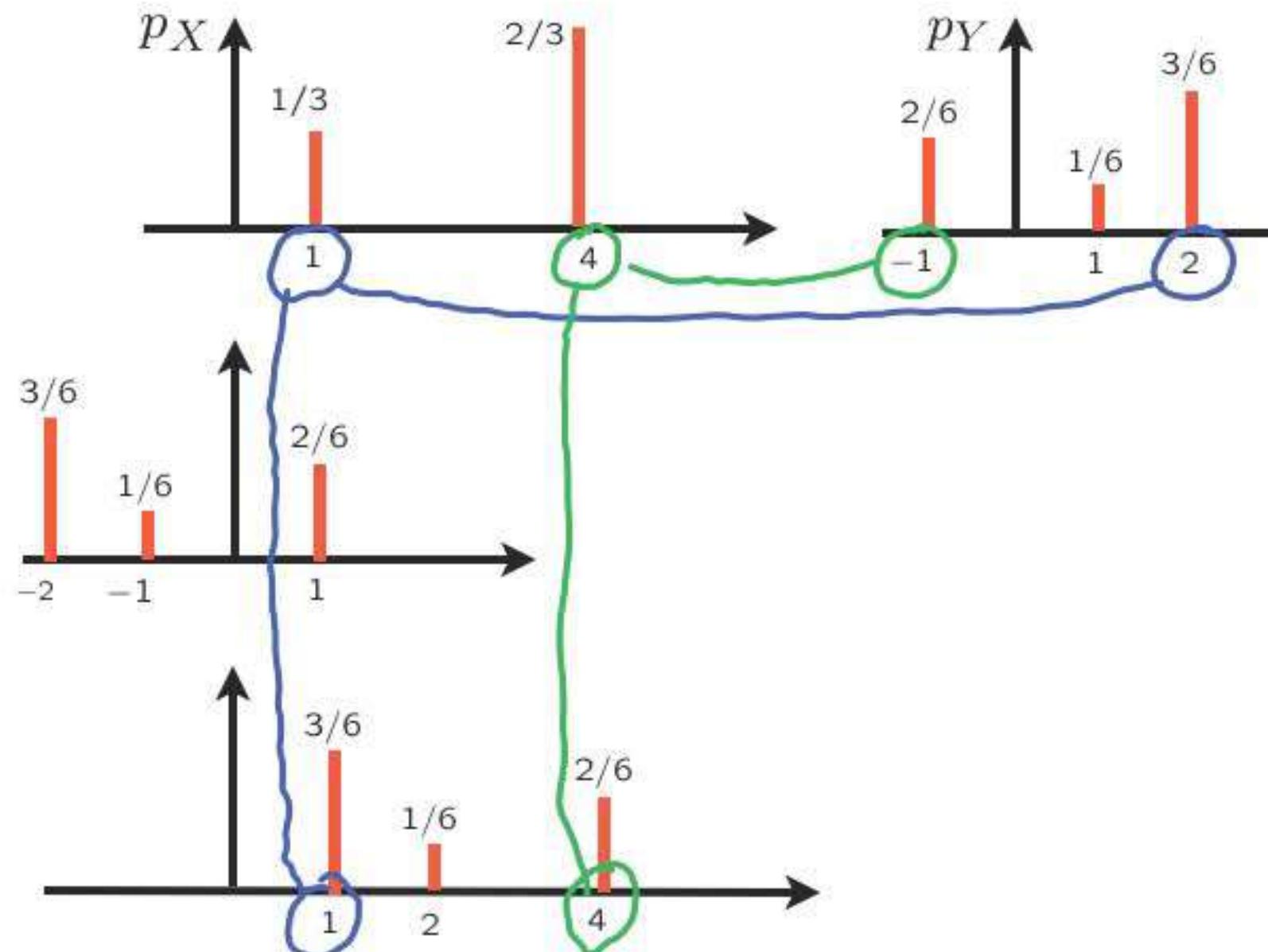
$$p_Z(z) = \sum_x p_X(x) p_Y(z-x)$$



$$P_Z(z) = \sum_x P(X=x, Y=z-x) \\ = \sum_x P_X(x) P_Y(z-x)$$

Discrete convolution mechanics

$$p_Z(z) = \sum_x p_X(x) p_Y(z - x)$$



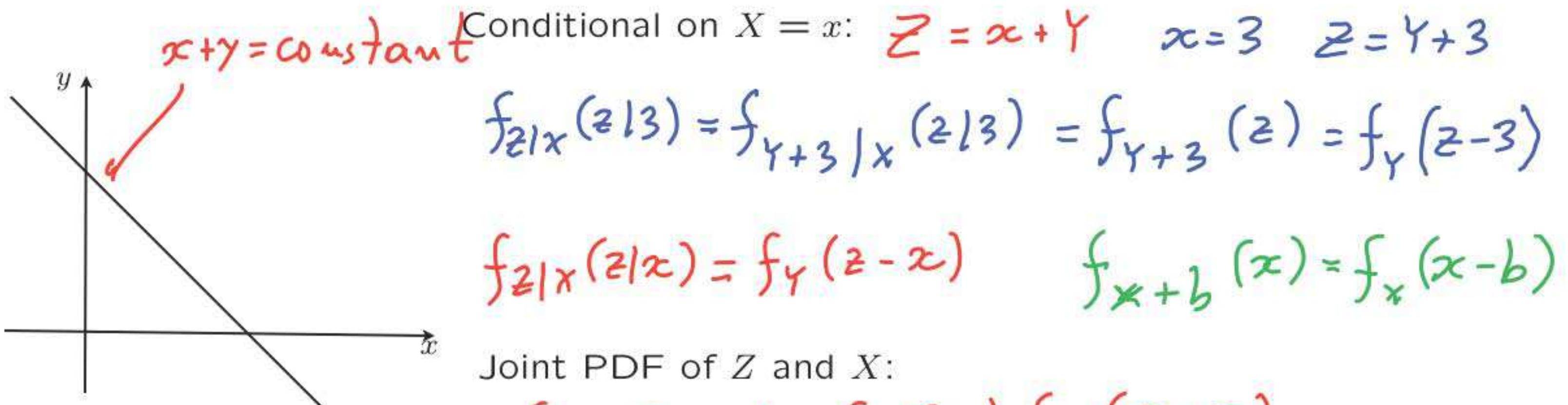
- To find $p_Z(3)$:
 - Flip (horizontally) the PMF of Y
 - Put it underneath the PMF of X
 - Right-shift the flipped PMF by 3
 - Cross-multiply and add
 - Repeat for other values of z

The distribution of $X + Y$: the continuous case

- $Z = X + Y$; X, Y independent, continuous known PDFs

$$p_Z(z) = \sum_x p_X(x) p_Y(z - x)$$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx$$



From joint to the marginal: $f_Z(z) = \int_{-\infty}^{\infty} f_{X,Z}(x,z) dx$

- Same mechanics as in discrete case (flip, shift, etc.)

The sum of independent normal r.v.'s

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x) dx$$

- $X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2), \text{ independent}$

$$Z = X + Y$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-(x-\mu_x)^2/2\sigma_x^2}$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-(y-\mu_y)^2/2\sigma_y^2}$$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\} \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left\{-\frac{(z-x-\mu_y)^2}{2\sigma_y^2}\right\} dx$$

$$(\text{algebra}) = \frac{1}{\sqrt{2\pi(\sigma_x^2 + \sigma_y^2)}} \exp\left\{-\frac{(z-\mu_x-\mu_y)^2}{2(\sigma_x^2 + \sigma_y^2)}\right\}$$

$$N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

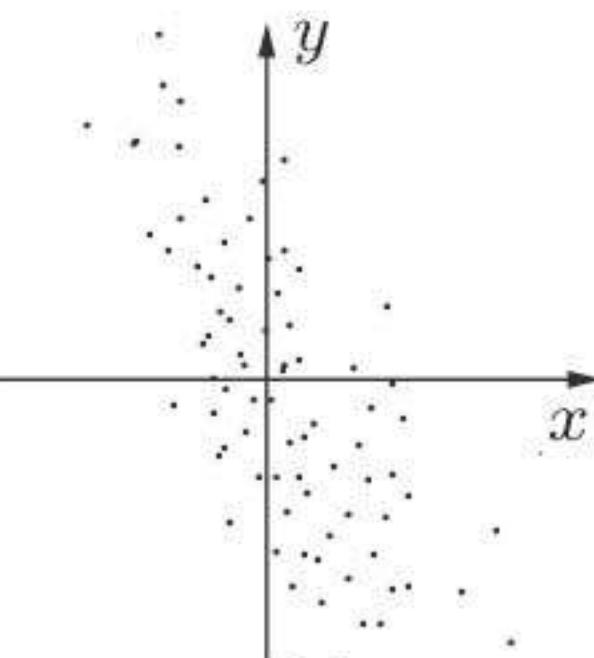
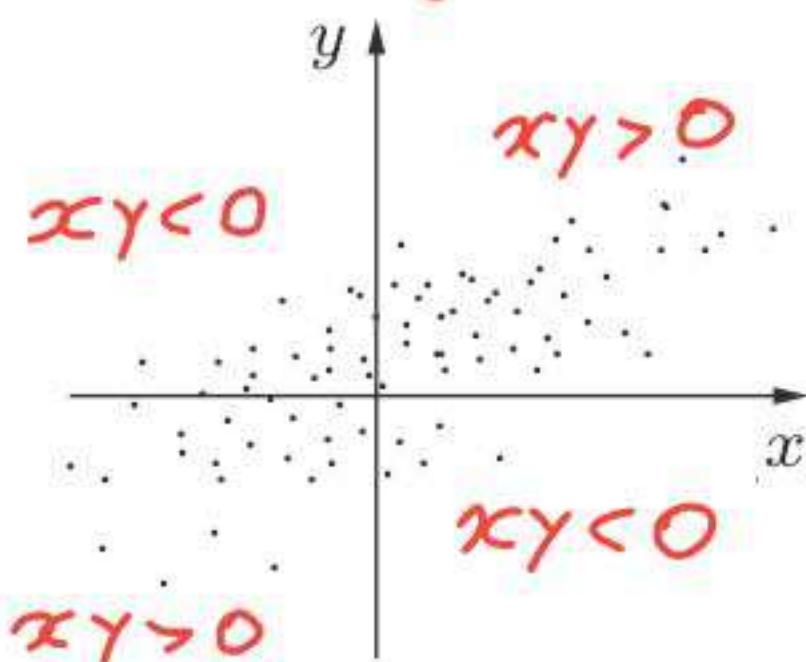
$X + Y + W$

The sum of finitely many independent normals is normal

Covariance

- Zero-mean, discrete X and Y
 - if independent: $E[XY] =$

$$= E[X]E[Y] = 0$$

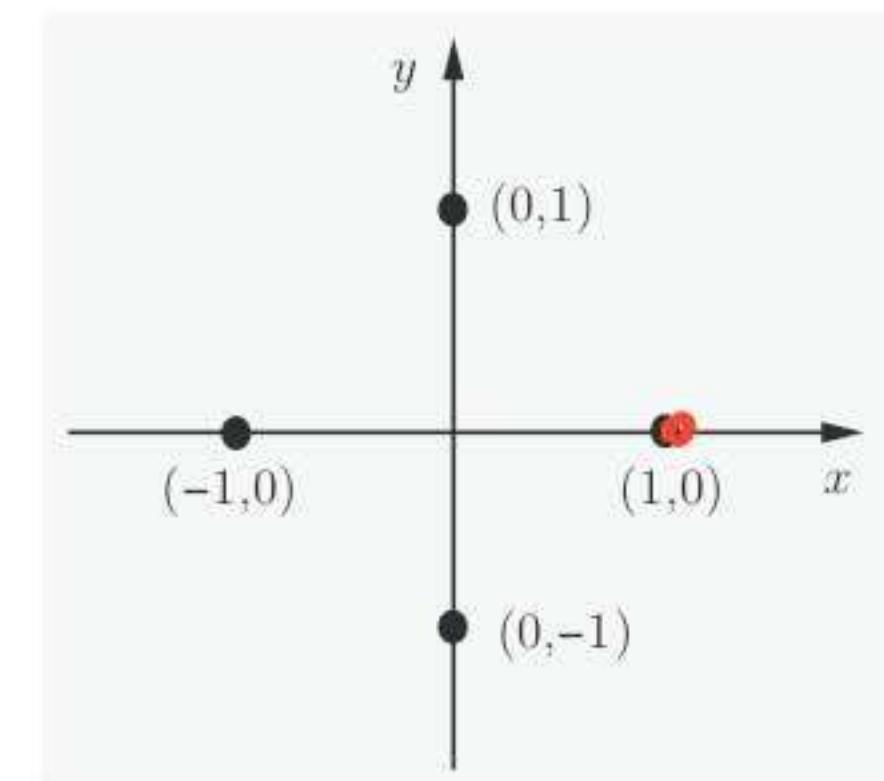


Definition for general case:

$$\text{cov}(X, Y) = E[(\underline{X - E[X]}) \cdot (\underline{Y - E[Y]})]$$

$$\text{and } \text{cov}(X, Y) = E[(X - E[X])] E[Y - E[Y]]$$

- independent $\Rightarrow \text{cov}(X, Y) = 0$
(converse is not true)



Covariance properties

$$\text{cov}(X, X) = E[(X - E[X])^2]$$
$$= \text{var}(X) = E[X^2] - (E[X])^2$$

$$\text{cov}(aX + b, Y) =$$

(assume 0 means)

$$= E[(aX+b)Y] = aE[XY] + bE[Y]$$
$$= a \cdot \text{cov}(X, Y)$$

$$\text{cov}(X, Y + Z) = E[X(Y+Z)]$$
$$= E[XY] + E[XZ] = \text{cov}(X, Y) + \text{cov}(X, Z)$$

$$\text{cov}(X, Y) = E[(X - E[X]) \cdot (Y - E[Y])]$$

$$= E[XY] - E[X]E[Y]$$
$$- E[E[X]Y] + E[E[X]E[Y]]$$
$$= E[XY] - E[X]E[Y]$$
$$- E[X]E[Y] + E[X]E[Y]$$

$$\text{cov}(X, Y) = E[XY] - E[X]E[Y]$$

The variance of a sum of random variables

$$\begin{aligned}\text{var}(X_1 + X_2) &= E[(X_1 + X_2 - E[X_1 + X_2])^2] \\&= E[(\underline{(X_1 - E[X_1])} + \underline{(X_2 - E[X_2])})^2] \\&= E[(X_1 - E[X_1])^2 + (X_2 - E[X_2])^2 \\&\quad + 2(X_1 - E[X_1])(X_2 - E[X_2])] \\&= \text{Var}(X_1) + \text{Var}(X_2) + 2 \text{cov}(X_1, X_2).\end{aligned}$$

The variance of a sum of random variables

$$\text{var}(X_1 + X_2) = \text{var}(X_1) + \text{var}(X_2) + 2 \text{cov}(X_1, X_2)$$

$$\begin{aligned}\text{var}(X_1 + \dots + X_n) &= E[(X_1 + \dots + X_n)^2] \\ (\text{assume 0 means}) &= E\left[\sum_{i=1}^n X_i^2 + \sum_{\substack{i=1, \dots, n \\ j=1, \dots, n \\ i \neq j}} X_i X_j\right] \\ &= \sum_i \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)\end{aligned}$$

$\left. \begin{array}{c} i=1, \dots, n \\ j=1, \dots, n \\ i \neq j \end{array} \right\} n^2 - n \text{ terms}$

$$\text{var}(X_1 + \dots + X_n) = \sum_{i=1}^n \text{var}(X_i) + \sum_{\{(i,j): i \neq j\}} \text{cov}(X_i, X_j)$$

The Correlation coefficient

- Dimensionless version of covariance:

$$-1 \leq \rho \leq 1$$

$$\begin{aligned}\rho(X, Y) &= E\left[\frac{(X - E[X]) \cdot (Y - E[Y])}{\sigma_X \sigma_Y}\right] \\ &= \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}\end{aligned}$$

- Measure of the degree of “association” between X and Y
- Independent $\Rightarrow \rho = 0$, “uncorrelated”
(converse is not true)
- $|\rho| = 1 \Leftrightarrow (X - E[X]) = c(Y - E[Y])$ (linearly related)

$$\bullet \rho(X, X) = \frac{\text{var}(X)}{\sigma_X^2} = 1$$

$$\bullet \text{cov}(aX + b, Y) = a \cdot \text{cov}(X, Y) \Rightarrow \rho(aX + b, Y) = \frac{a \text{cov}(X, Y)}{|a| \sigma_X \sigma_Y} = \underbrace{\frac{a}{|a|} \text{sign}(a)}_{\text{• } \rho(X, Y)}$$

Proof of key properties of the correlation coefficient

$$\rho(X, Y) = E \left[\frac{(X - E[X])}{\sigma_X} \cdot \frac{(Y - E[Y])}{\sigma_Y} \right]$$

$$-1 \leq \rho \leq 1$$

- Assume, for simplicity, zero means and unit variances, so that $\rho(X, Y) = E[XY]$

$$\begin{aligned} E[(X - \rho Y)^2] &= E[X^2] - 2\rho E[XY] + \rho^2 E[Y^2] \\ 0 \leq &= 1 - 2\rho^2 + \rho^2 = \underline{\underline{1 - \rho^2}} \quad 1 - \rho^2 \geq 0 \Rightarrow \rho^2 \leq 1 \end{aligned}$$

If $|\rho| = 1$, then $X = \rho Y \Rightarrow X = Y$ or $X = -Y$

Interpreting the correlation coefficient

- Association does not imply causation or influence

X : math aptitude

Y : musical ability

- Correlation often reflects underlying, common, hidden factor

- Assume, Z, V, W are independent

$$X = \underline{\underline{Z}} + V$$

$$Y = \underline{\underline{Z}} + W$$

Assume, for simplicity, that Z, V, W have zero means, unit variances

$$\text{var}(x) = \text{var}(Z) + \text{var}(v) = 2 \Rightarrow \sigma_x = \sqrt{2} \quad \sigma_y = \sqrt{2}$$

$$\begin{aligned}\text{cov}(x, y) &= E[(Z+V)(Z+W)] = E[Z^2] + E[梓] + E[ZW] + E[VW] \\ &= 1 + 0 + 0 + 0\end{aligned}$$

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Correlations matter...

- A real-estate investment company invests \$10M in each of 10 states. At each state i , the return on its investment is a random variable X_i , with mean 1 and standard deviation 1.3 (in millions).

$$\text{var}(X_1 + \dots + X_{10}) = \sum_{i=1}^{10} \text{var}(X_i) + \sum_{\{(i,j): i \neq j\}} \text{cov}(X_i, X_j)$$

$$E[X_1 + \dots + X_{10}] = 10$$

- If the X_i are uncorrelated, then:

$$\text{var}(X_1 + \dots + X_{10}) = 10 \cdot (1.3)^2 = 16.9 \quad \sigma(X_1 + \dots + X_{10}) = 4.1$$

- If for $i \neq j$, $\rho(X_i, X_j) = 0.9$:

$$\text{cov}(X_i, X_j) = \rho \sigma_{X_i} \sigma_{X_j} = 0.9 \times 1.3 \times 1.3 = 1.52$$

$$\text{var}(X_1 + \dots + X_{10}) = 10 \cdot (1.3)^2 + 90 \cdot 1.52 = 154$$

$$\sigma(X_1 + \dots + X_{10}) = 12.4$$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 13: Conditional expectation and variance revisited;

Application: Sum of a random number of independent r.v.'s

- A more abstract version of the conditional expectation
 - view it as a random variable
 - the law of iterated expectations
- A more abstract version of the conditional variance
 - view it as a random variable
 - the law of total variance
- Sum of a random number
of independent r.v.'s
 - mean
 - variance

Conditional expectation as a random variable

- Function h
e.g., $h(x) = x^2$, for all x
 - Random variable X ; what is $h(X)$?
 $\cancel{= X^2}$
 - $h(X)$ is the r.v. that takes the value x^2 , if X happens to take the value x
- $\cancel{g(y)} = \mathbb{E}[X | Y = y] = \sum_x x p_{X|Y}(x | y)$
 $=$ (integral in continuous case)
 - $g(Y)$: is the r.v. that takes the value $\underline{\mathbb{E}[X | Y = y]}$, if Y happens to take the value y
- Remarks:
 - It is a function of Y
 - It is a random variable
 - Has a distribution, mean, variance, etc.

Definition: $\underline{\mathbb{E}[X|Y]} = g(Y)$

The mean of $E[X | Y]$: Law of iterated expectations

- $g(y) = E[X | Y = y]$

$$E[X | Y] \stackrel{\Delta}{=} g(Y)$$

$$E[E[X | Y]] = E[X]$$

$$\underbrace{E[E[X | Y]]}_{\text{exp. value rule}} = E[g(Y)]$$

$$= \sum_y g(y) P_Y(y)$$

exp. value rule

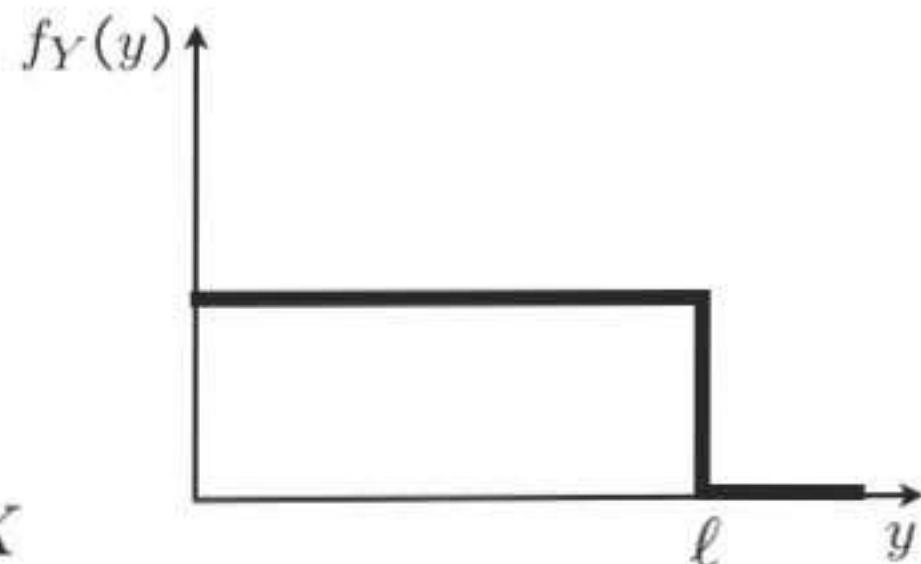
$$= \sum_y E[X | Y=y] P_Y(y)$$

• total exp then

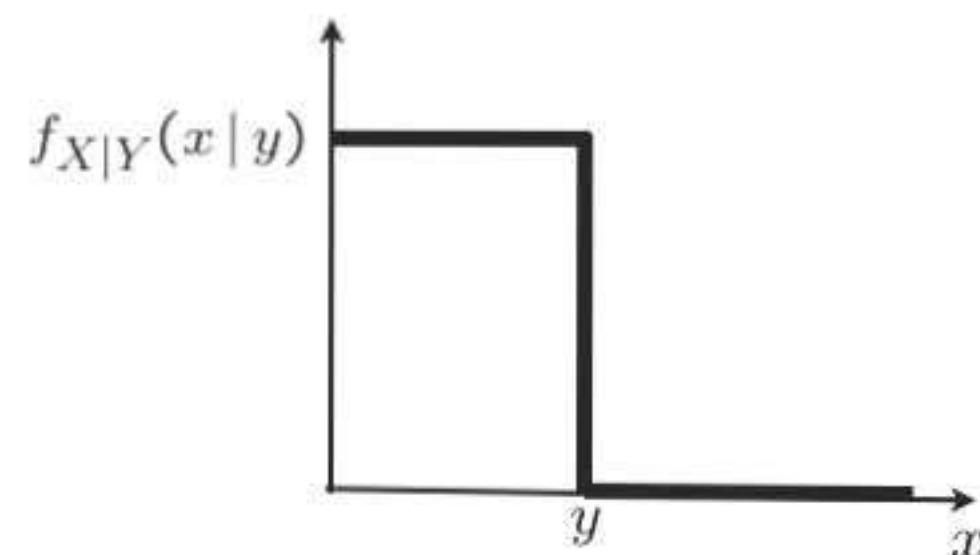
$$= E[X]$$

Stick-breaking example

- Stick example: stick of length ℓ
break at uniformly chosen point Y
break what is left at uniformly chosen point X



- $E[X | Y = y] = \frac{y}{2}$



- $E[X | Y] = \frac{\gamma}{2}$

$$E[X] = E[E[X | Y]] = E\left[\frac{Y}{2}\right] = \frac{1}{2} E[Y] = \frac{1}{2} \cdot \frac{\ell}{2} = \frac{\ell}{4}$$

Forecast revisions

$$E[E[X|Y]] = E[X]$$

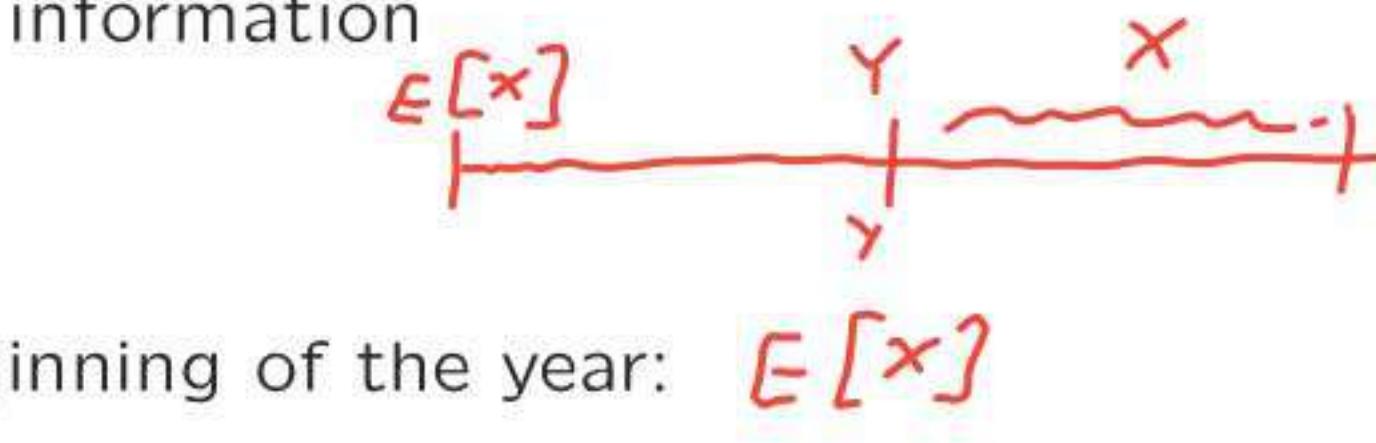
- Suppose forecasts are made by calculating expected value, given any available information

- X : February sales

- Forecast in the beginning of the year: $E[X]$

- End of January: will get new information, value y of Y

Revised forecast:



$$E[X|Y]$$

- Law of iterated expectations:

$$E[\text{revised forecast}] = E[X] = \text{original forecast}$$

The conditional variance as a random variable

$$\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

$$\text{var}(X | Y = y) = \mathbb{E}[(X - \underline{\mathbb{E}[X | Y = y]})^2 | Y = y]$$



$\text{var}(X | Y)$ is the r.v. that takes the value $\text{var}(X | Y = y)$, when $Y = y$

- Example: X uniform on $[0, Y]$

$$\text{var}(X | Y = y) = \frac{y^2}{12}$$

$$\text{var}(X | Y) = \frac{Y^2}{12}$$

Law of total variance: $\text{var}(X) = \mathbb{E}[\text{var}(X | Y)] + \text{var}(\mathbb{E}[X | Y])$

Derivation of the law of total variance

$$\bullet \quad \text{var}(X) = E[\text{var}(X | Y)] + \text{var}(E[X | Y])$$

$$\bullet \quad \text{var}(X) = E[X^2] - (E[X])^2$$

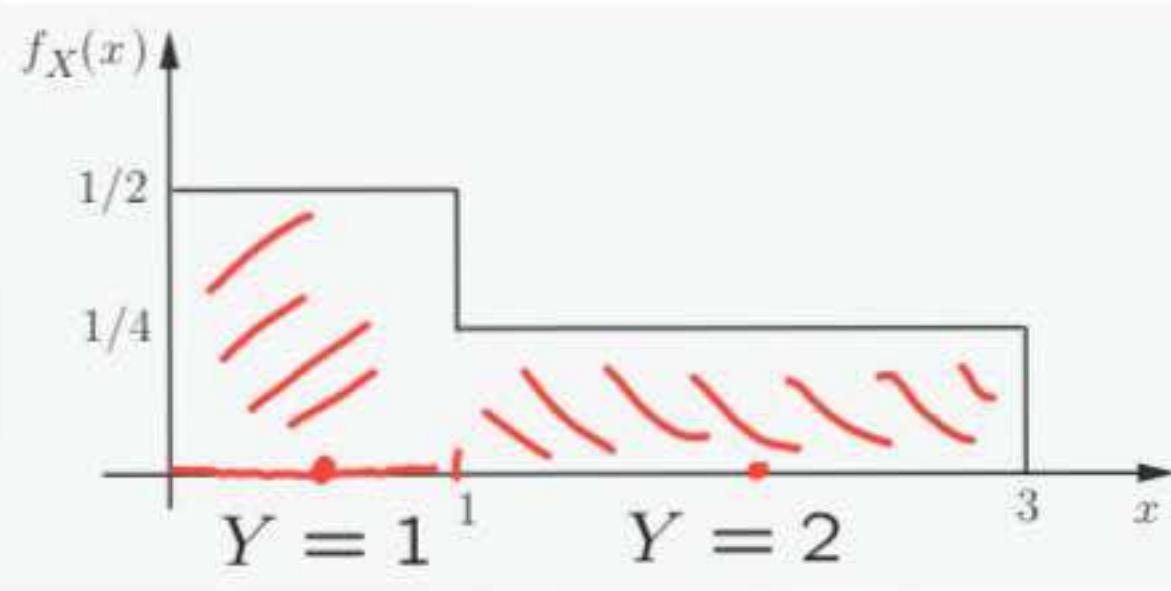
$$\text{var}(X | Y = y) = E[X^2 | Y = y] - (E[X | Y = y])^2 \text{ for all } y$$

$$\text{var}(X | Y) = E[X^2 | Y] - (E[X | Y])^2$$

$$E[\text{var}(X | Y)] = E[X^2] - E[(E[X | Y])^2]$$

$$+ \quad \text{var}(E[X | Y]) = E[(E[X | Y])^2] - (E[E[X | Y]])^2$$
$$= (E[X])^2$$

A simple example



$$\begin{aligned} \text{var}(X) &= \mathbb{E}[\text{var}(X | Y)] + \text{var}(\mathbb{E}[X | Y]) = \frac{37}{48} \\ &= 5/24 + 9/16 \end{aligned}$$

$$\begin{aligned} \text{var}(X | Y) &= \frac{1/2}{1/2} \quad \text{var}(X | Y=1) = \frac{1/12}{1/12} \\ &\quad \text{var}(X | Y=2) = \frac{2^2/12}{1/12} = \frac{4}{12} \end{aligned}$$

$$\mathbb{E}[\text{var}(X | Y)] = \frac{1}{2} \cdot \frac{1}{12} + \frac{1}{2} \cdot \frac{4}{12} = \frac{5}{24}$$

$$\begin{aligned} \mathbb{E}[X | Y] &= \frac{1/2}{1/2} \quad \mathbb{E}[X | Y=1] = \frac{1}{2} \\ &\quad \frac{1/2}{1/2} \quad \mathbb{E}[X | Y=2] = 2 \end{aligned}$$

$$\text{var}(\mathbb{E}[X | Y]) = \frac{1}{2} \left(\frac{1}{2} - \frac{5}{4} \right)^2$$

$$\begin{aligned} \mathbb{E}[\mathbb{E}[X | Y]] &= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot 2 = \frac{5}{4} = E[X] \\ &\quad + \frac{1}{2} \left(2 - \frac{5}{4} \right)^2 = \frac{9}{16} \end{aligned}$$

Section means and variances

- Two sections of a class: $y = 1$ (10 students); $y = 2$ (20 students)
 x_i : score of student i
- Experiment: pick a student at random (uniformly)
random variables: X and Y
- Data: $y = 1 : \frac{1}{10} \sum_{i=1}^{10} x_i = 90$ $y = 2 : \frac{1}{20} \sum_{i=11}^{30} x_i = 60$

$$\bullet E[X] = \frac{1}{30} \sum_{i=1}^{30} x_i = \frac{1}{30} (90 \cdot 10 + 60 \cdot 20) = 70$$

$$E[X | Y = 1] = 90$$

$$E[X | Y = 2] = 60$$

$$E[X | Y] = \begin{cases} 1/3 & 90 \\ 2/3 & 60 \end{cases}$$

$$\bullet E[E[X | Y]] = \frac{1}{3} \cdot 90 + \frac{2}{3} \cdot 60 = 70$$

Section means and variances (ctd.)

$$E[X | Y] = \begin{cases} 90, & \text{w.p. } 1/3 \\ 60, & \text{w.p. } 2/3 \end{cases}$$

$$E[E[X | Y]] = 70 = E[X]$$

$$\text{var}(E[X | Y]) = \frac{1}{3}(90 - 70)^2 + \frac{2}{3}(60 - 70)^2 = 200$$

- More data:

$$\frac{1}{10} \sum_{i=1}^{10} (x_i - 90)^2 = 10$$

$$\frac{1}{20} \sum_{i=11}^{30} (x_i - 60)^2 = 20$$

$$\text{var}(X | Y = 1) = 10$$

$$\text{var}(X | Y) = \frac{\frac{1}{3} \cdot 10}{\frac{2}{3} \cdot 20}$$

$$\text{var}(X | Y = 2) = 20$$

$$E[\text{var}(X | Y)] = \frac{1}{3} \cdot 10 + \frac{2}{3} \cdot 20 = \frac{50}{3}$$

$$\text{var}(X) = E[\text{var}(X | Y)] + \text{var}(E[X | Y]) = \frac{50}{3} + 200$$

$\text{var}(X)$ = (average variability **within** sections) + (variability **between** sections)

Sum of a random number of independent r.v.'s

$$E[Y] = E[N] \cdot E[X]$$

- N : number of stores visited
(N is a nonnegative integer r.v.)
- Let $Y = X_1 + \dots + X_N$
- X_i : money spent in store i
 - X_i independent, identically distributed
 - independent of N

$$\begin{aligned} E[Y | N = n] &= E[X_1 + \dots + X_n | N = n] = E[X_1 + \dots + X_n | N = n] \\ &= E[X_1 + \dots + X_n] = n E[X] \end{aligned}$$

$\xrightarrow{\text{E}[Y|N]} = N E[X]$

- Total expectation theorem:

$$E[Y] = \sum_n p_N(n) E[Y | N = n] = \underbrace{\sum_n p_N(n) n}_{\text{E}[N]} E[X] = E[N] E[X]$$

- Law of iterated expectations:

$$E[Y] = E[E[Y | N]] = E[NE[X]] = E[N] E[X]$$

Variance of sum of a random number of independent r.v.'s

$$Y = X_1 + \dots + X_N$$

$$\bullet \quad \text{var}(Y) = \mathbf{E}[\text{var}(Y | N)] + \text{var}(\mathbf{E}[Y | N])$$

$$\bullet \quad \mathbf{E}[Y | N] = N \mathbf{E}[X]$$

$$\text{var}(Y) = \mathbf{E}[N] \text{var}(X) + (\mathbf{E}[X])^2 \text{var}(N)$$

$$\bullet \quad \text{var}(\mathbf{E}[Y | N]) = \text{var}(N \mathbf{E}[X]) = (\mathbf{E}[X])^2 \text{var}(N)$$

$$\bullet \quad \text{var}(Y | N = n) = \text{var}(X_1 + \dots + X_n | N = n) = \text{var}(X_1 + \dots + X_n) = n \text{var}(X)$$

$$\text{var}(Y | N) = N \text{var}(X)$$

$$\bullet \quad \mathbf{E}[\text{var}(Y | N)] = \mathbf{E}[N \text{var}(X)] = \mathbf{E}[N] \text{var}(X)$$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

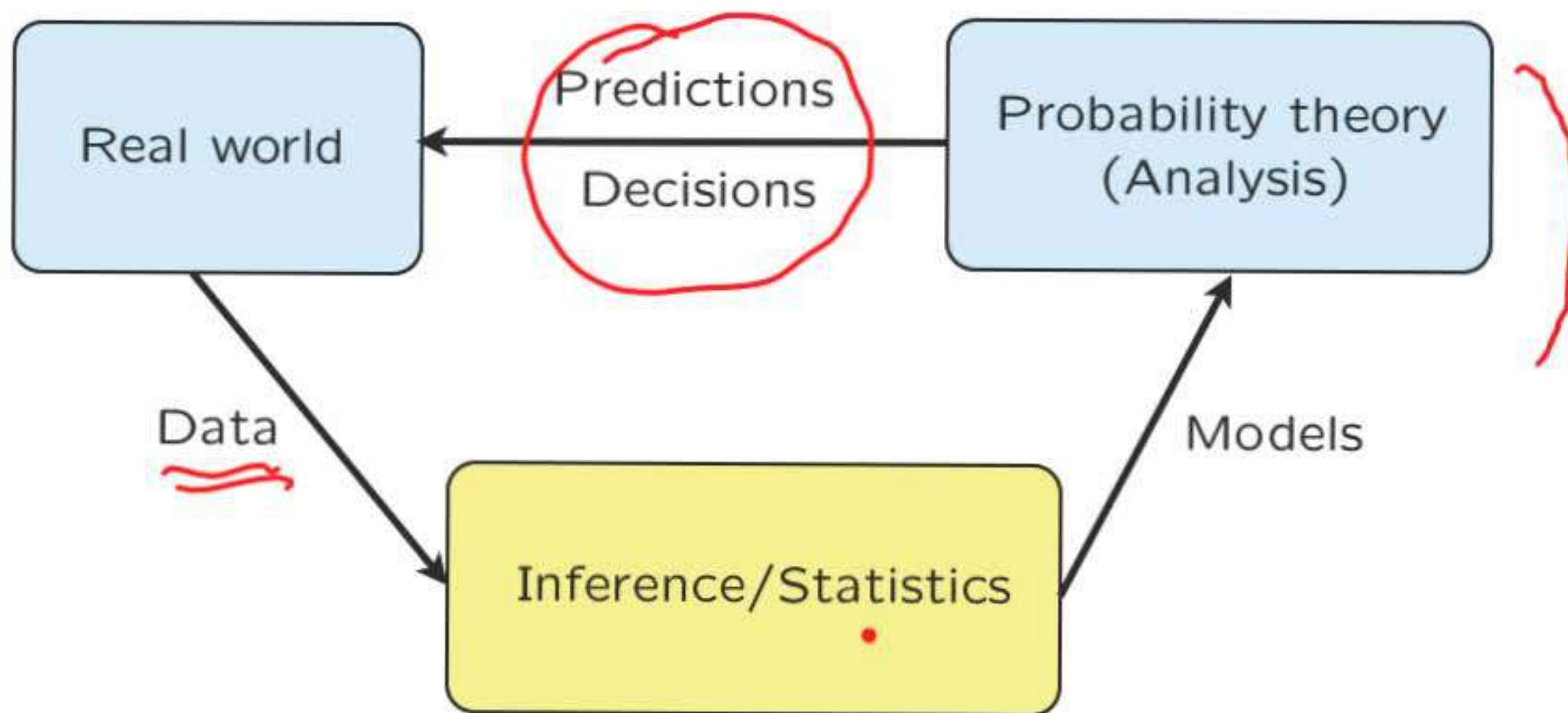
The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 14: Introduction to Bayesian inference

- The big picture
 - motivation, applications
 - problem types (hypothesis testing, estimation, etc.)
- The general framework
 - Bayes' rule → posterior
(4 versions)
 - point estimates (MAP, LMS)
 - performance measures)
(prob. of error; mean squared error)
 - examples

Inference: the big picture



Inference then and now

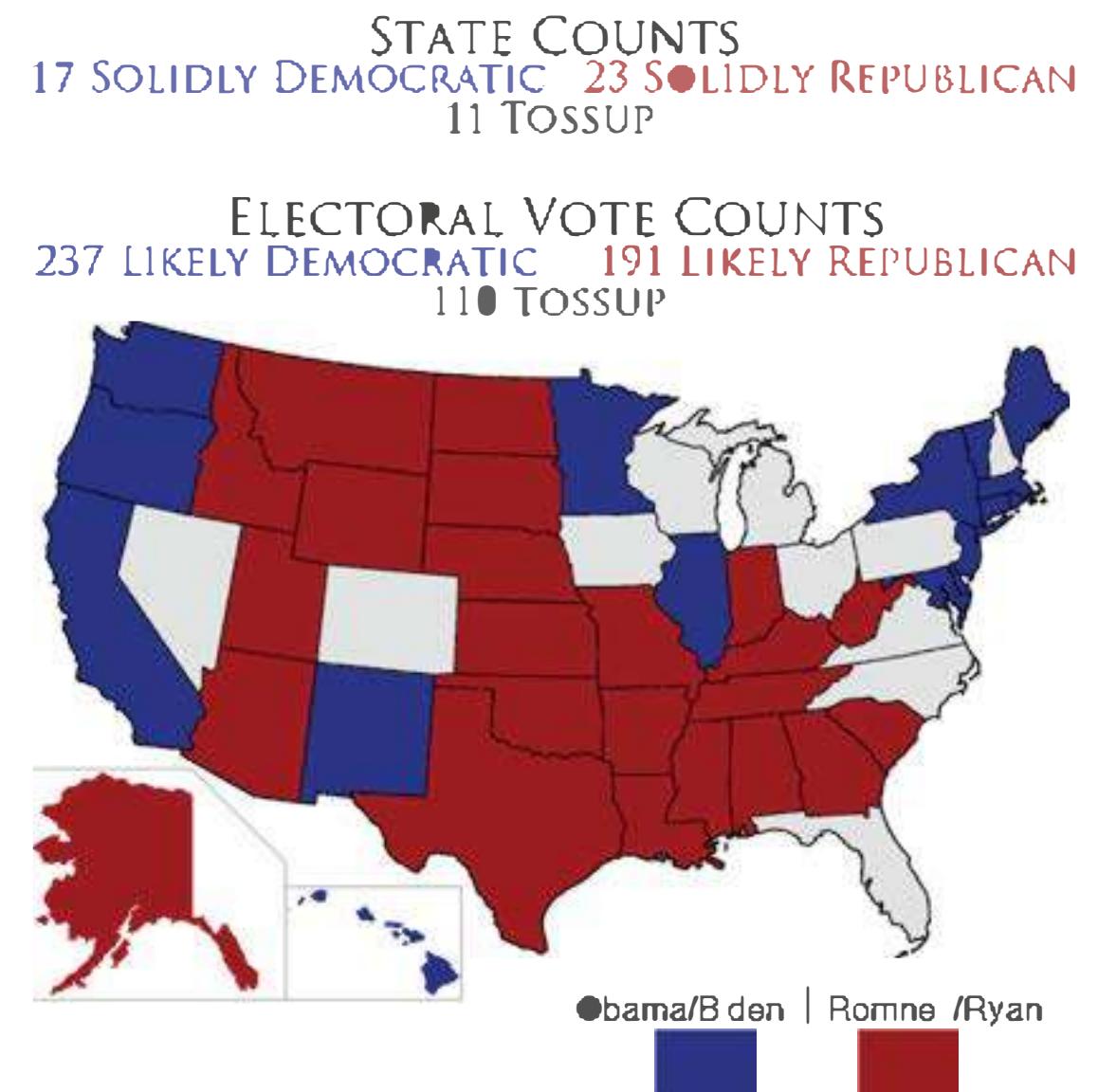
- Then:
 - 10 patients were treated: 3 died
 - 10 patients were not treated: 5 died
- Therefore ...

Now:

- Big data
- Big models
- Big computers

A sample of application domains

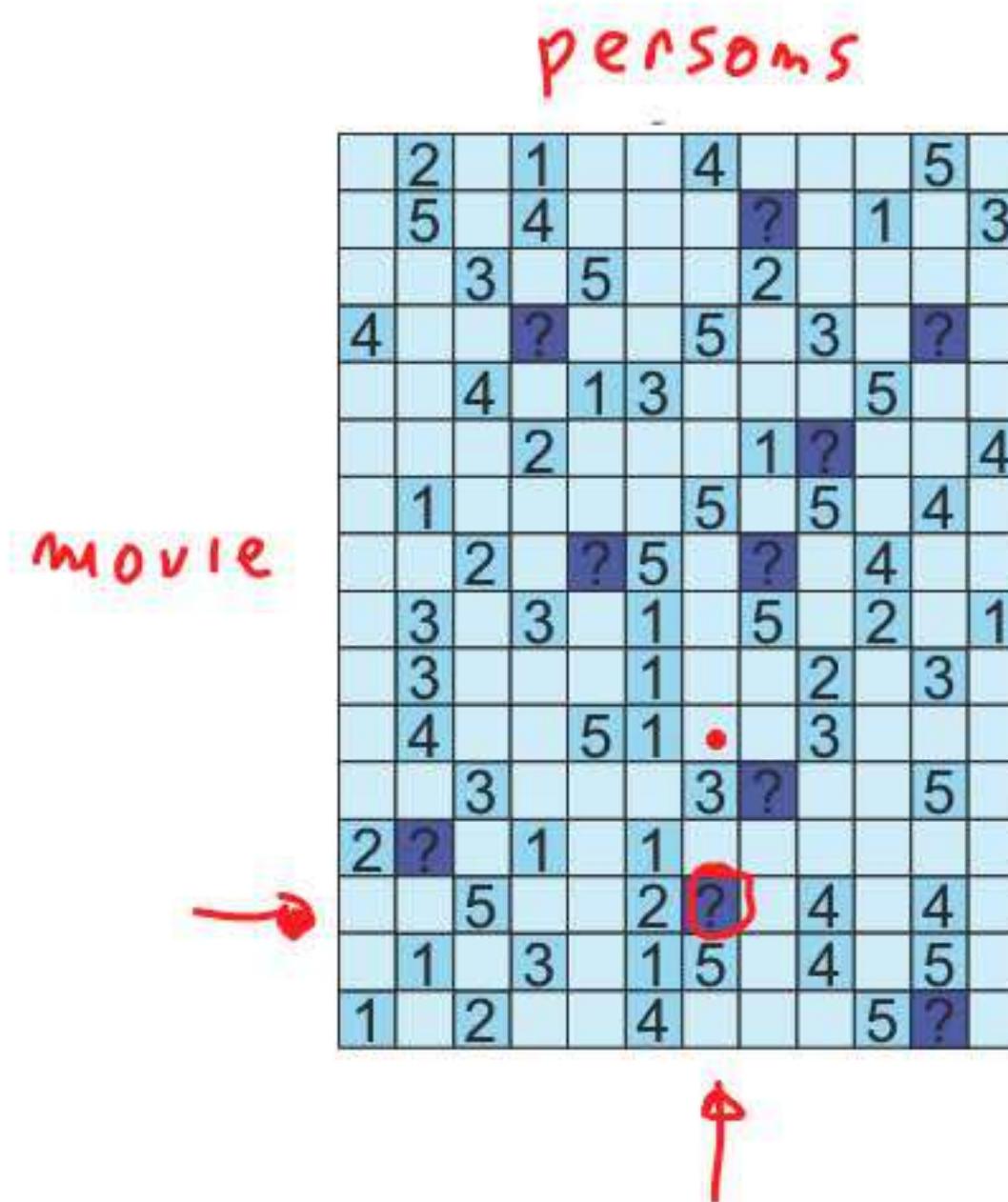
- Design and interpretation of experiments
 - polling •



© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use>.

A sample of application domains

- marketing, advertising
- recommendation systems
 - Netflix competition



A sample of application domains

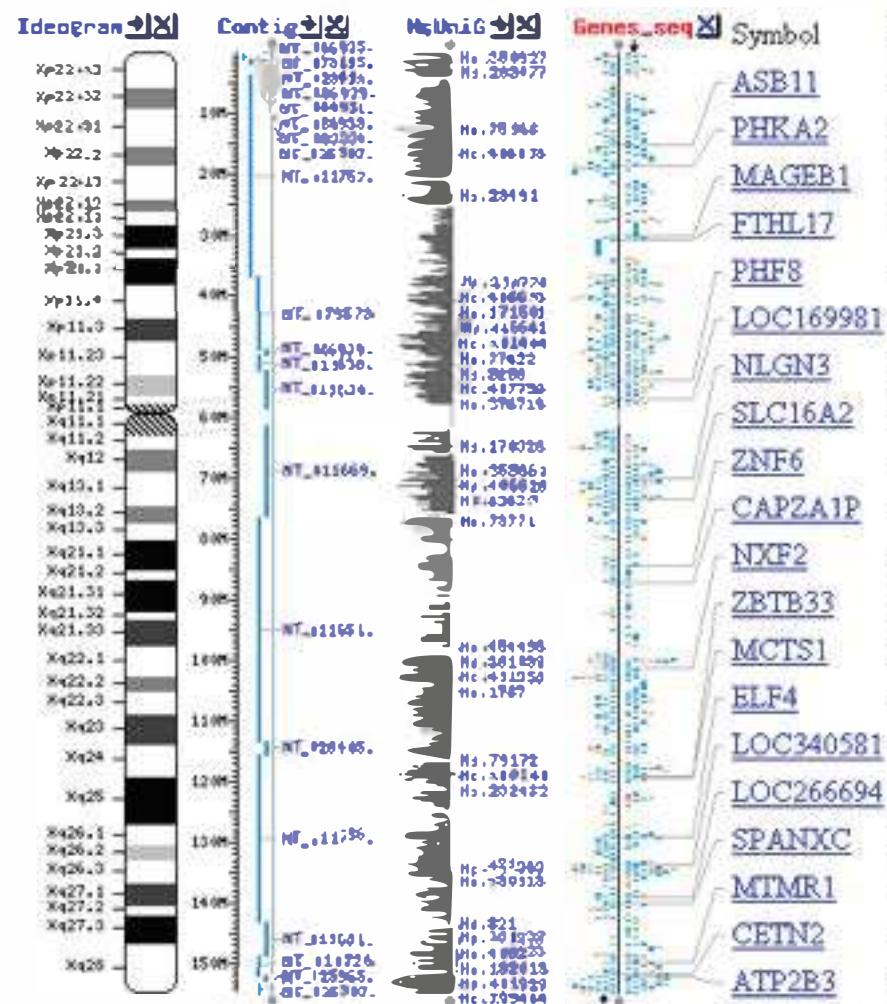
- Finance



© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use>.

A sample of application domains

- Life sciences
 - genomics

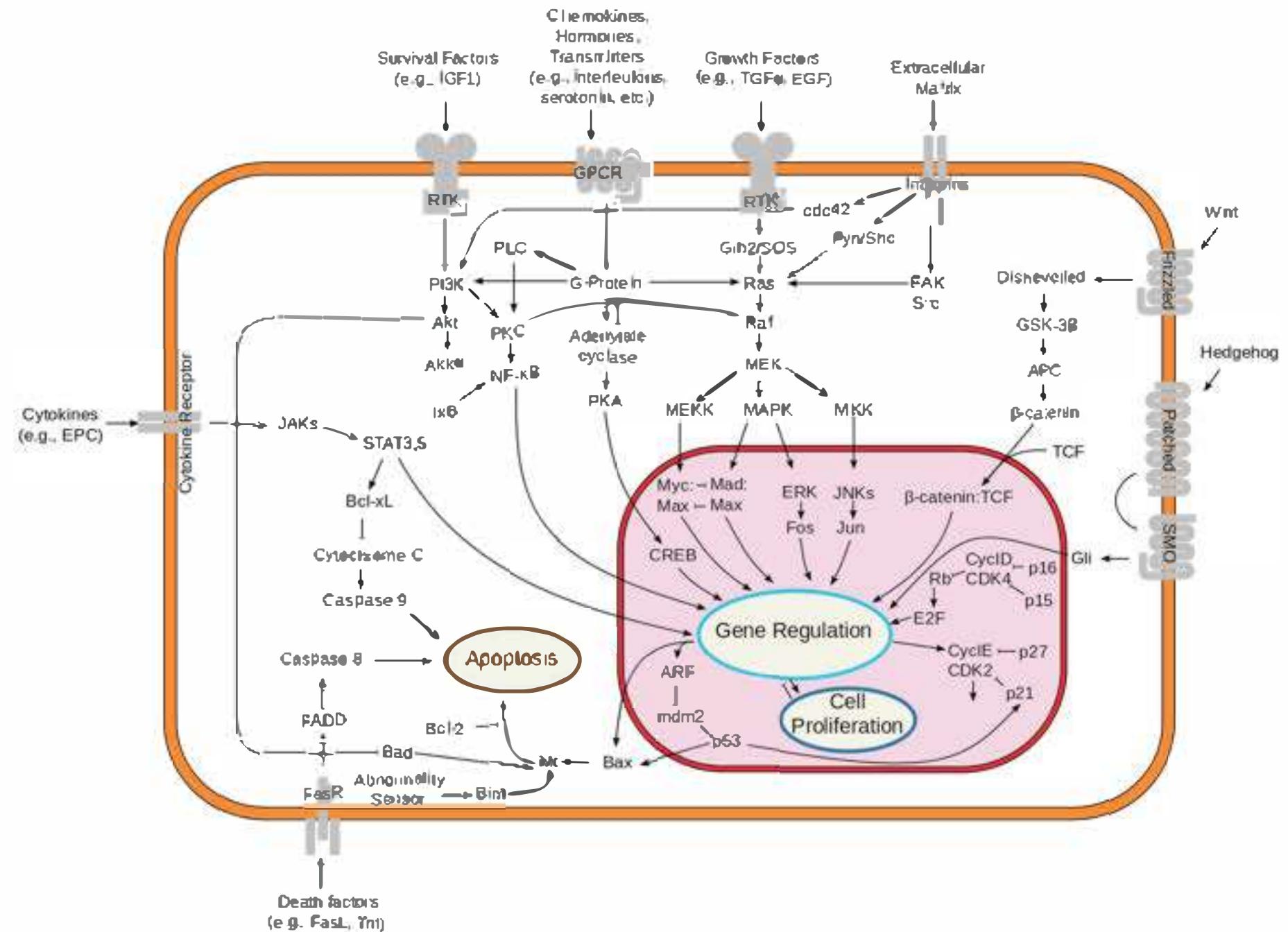


neuroscience, etc., etc.

This image is in the public domain.

Source: [Wikimedia](#).

- systems biology



This image is in the public domain.

Source: [Wikimedia](#).

A sample of application domains

- Modeling and monitoring the oceans
- Modeling and monitoring global climate
- Modeling and monitoring pollution
- Interpreting data from physics experiments
- Interpreting astronomy data

A sample of application domains

- Signal processing
 - communication systems (noisy ...)
 - speech processing and understanding
 - image processing and understanding
 - tracking of objects
 - positioning systems (e.g., GPS)
 - detection of abnormal events

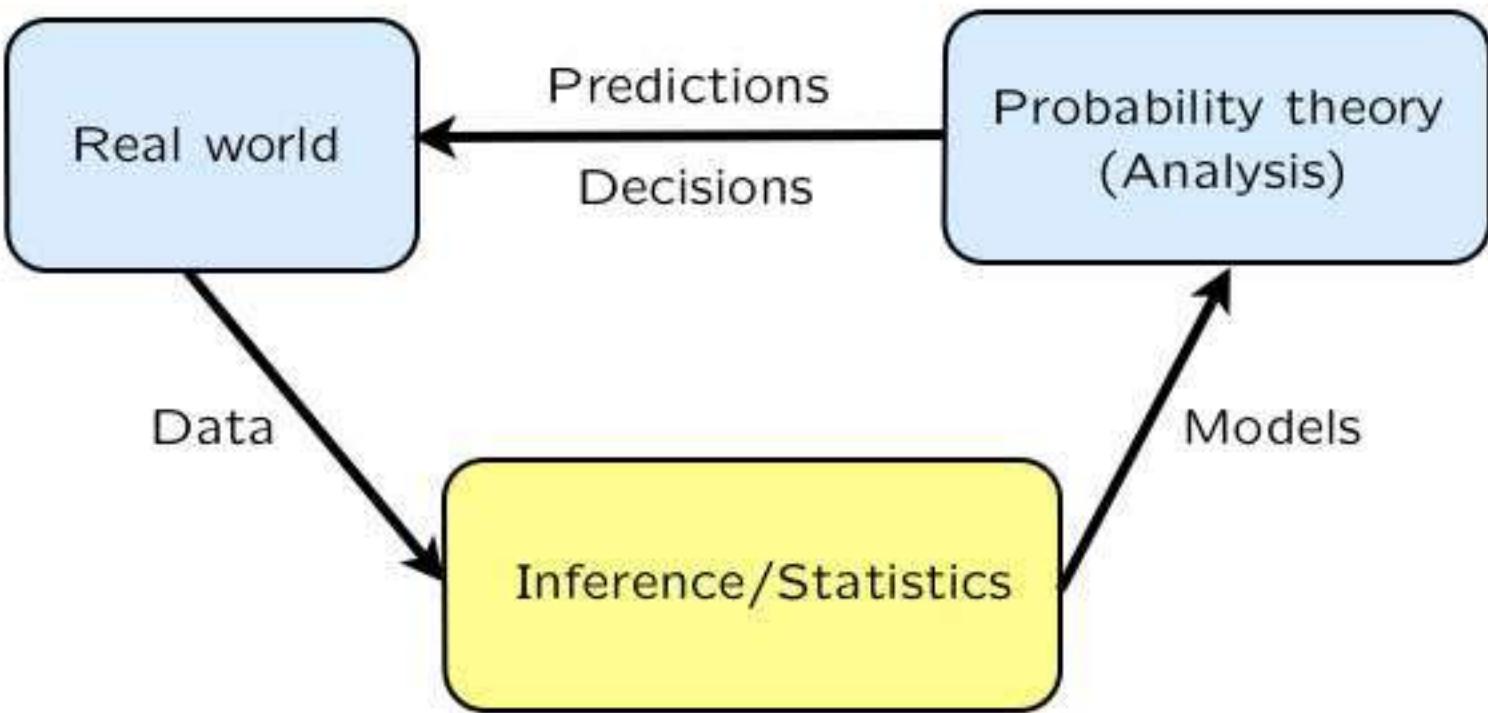
•

Model building versus inferring unobserved variables



$$X = aS + W$$

- Model building:
 - know “signal” S , observe X
 - infer a
- Variable estimation:
 - know a , observe X
 - infer S



Hypothesis testing versus estimation

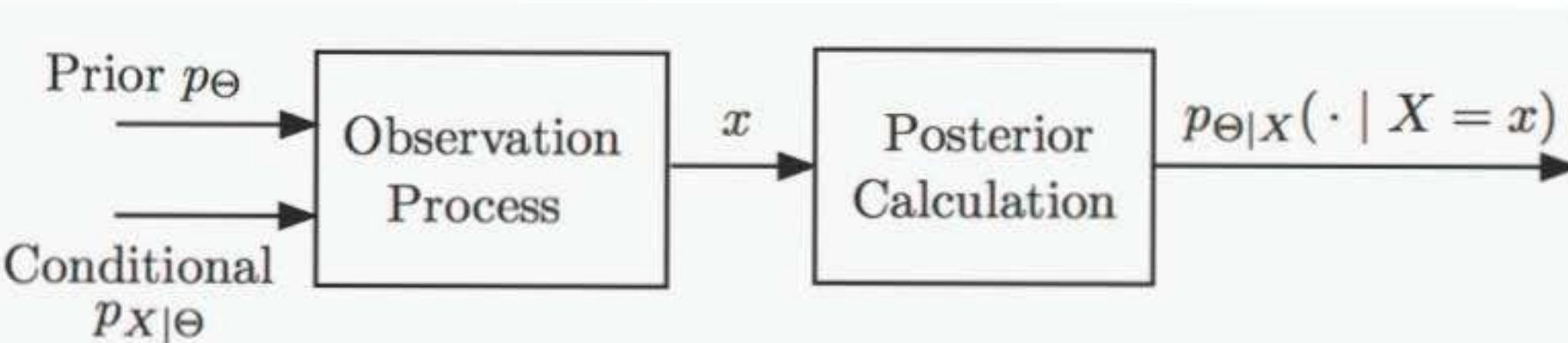
- Hypothesis testing:
 - unknown takes one of few possible values
 - aim at small probability of incorrect decision

Is it an airplane or a bird?

- Estimation:
 - numerical unknown(s)
 - aim at an estimate that is “close” to the true but unknown value

The Bayesian inference framework

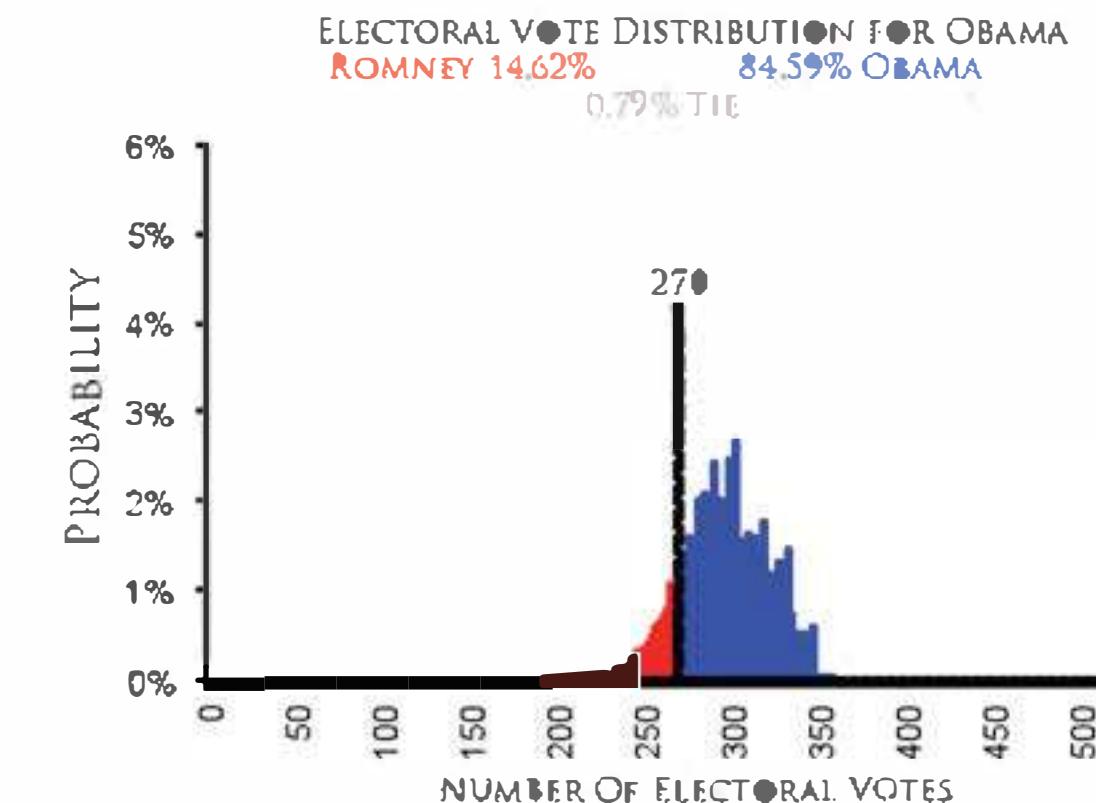
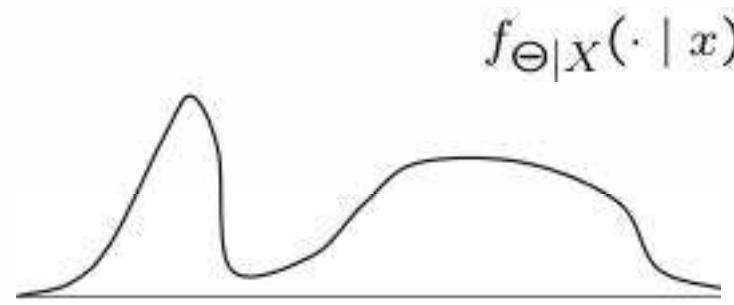
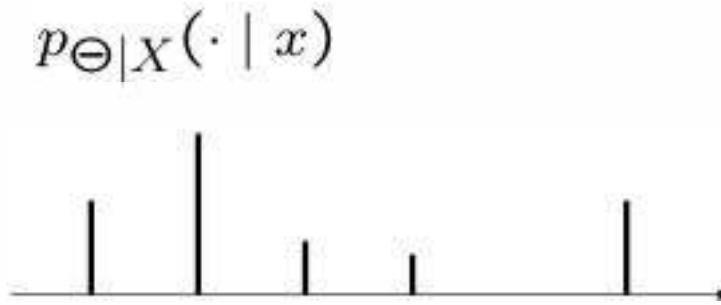
- Unknown Θ
 - treated as a random variable
 - prior distribution p_Θ or f_Θ
- Observation X
 - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot | X = x)$ or $f_{\Theta|X}(\cdot | X = x)$
- Where does the prior come from?
 - symmetry
 - known range
 - earlier studies
 - subjective or arbitrary



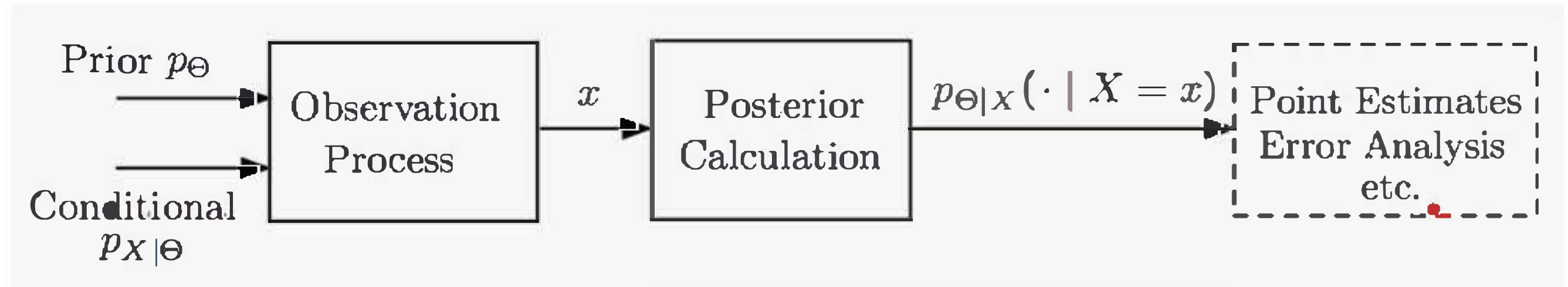
The output of Bayesian inference

The complete answer is a posterior distribution:

PMF $p_{\Theta|X}(\cdot | x)$ or PDF $f_{\Theta|X}(\cdot | x)$



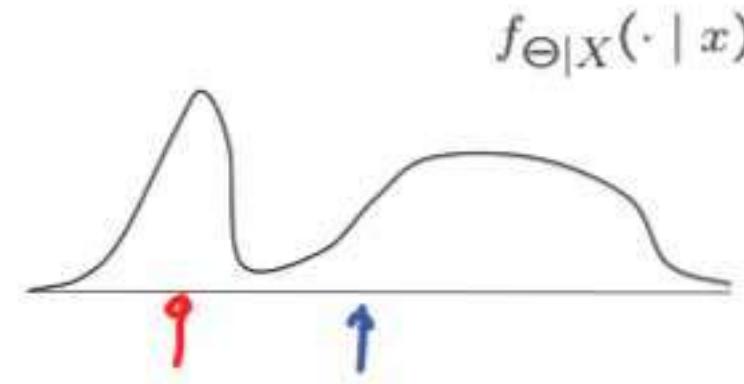
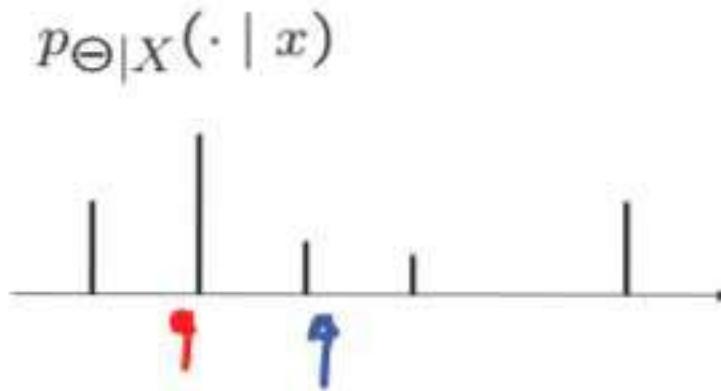
© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use>.



Point estimates in Bayesian inference

The complete answer is a posterior distribution:

PMF $p_{\Theta|X}(\cdot | x)$ or PDF $f_{\Theta|X}(\cdot | x)$



estimate: $\hat{\theta} = g(x)$
(number)

estimator: $\widehat{\Theta} = g(X)$
(random variable)

- Maximum a posteriori probability (MAP):

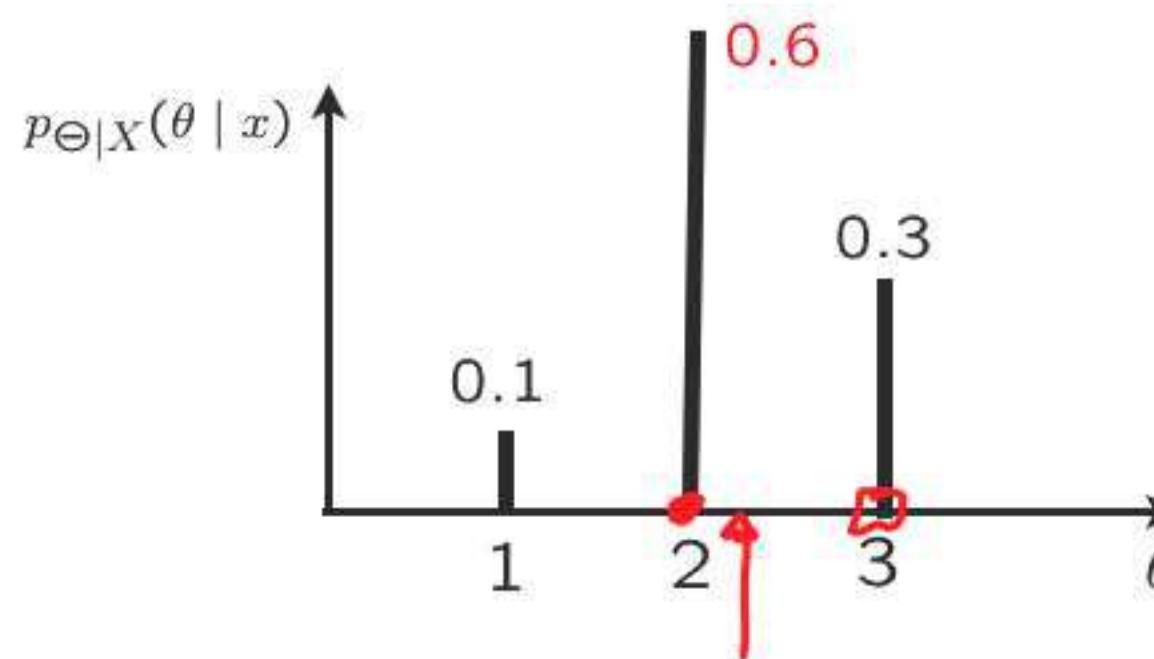
$$p_{\Theta|X}(\theta^* | x) = \max_{\theta} p_{\Theta|X}(\theta | x),$$

$$f_{\Theta|X}(\theta^* | x) = \max_{\theta} f_{\Theta|X}(\theta | x).$$

- Conditional expectation: $E[\Theta | X = x]$ (LMS: Least Mean Squares)

Discrete Θ , discrete X

- values of Θ : alternative hypotheses



- MAP rule: $\hat{\theta} = 2$

$$LMS: \hat{\theta} = E[\Theta | X=x] = 2.2$$

$$p_{\Theta|X}(\theta | x) = \frac{p_{\Theta}(\theta) p_{X|\Theta}(x | \theta)}{p_X(x)}$$

$$p_X(x) = \sum_{\theta'} p_{\Theta}(\theta') p_{X|\Theta}(x | \theta')$$

- conditional prob of error:

$$P(\hat{\theta} \neq \Theta | X=x) = 0.4$$

smallest under the MAP rule

- overall probability of error:

$$P(\hat{\Theta} \neq \Theta) = \sum_x P(\hat{\Theta} \neq \Theta | X=x) p_X(x)$$

$$= \sum_{\theta} P(\hat{\Theta} \neq \Theta | \Theta = \theta) p_{\Theta}(\theta)$$

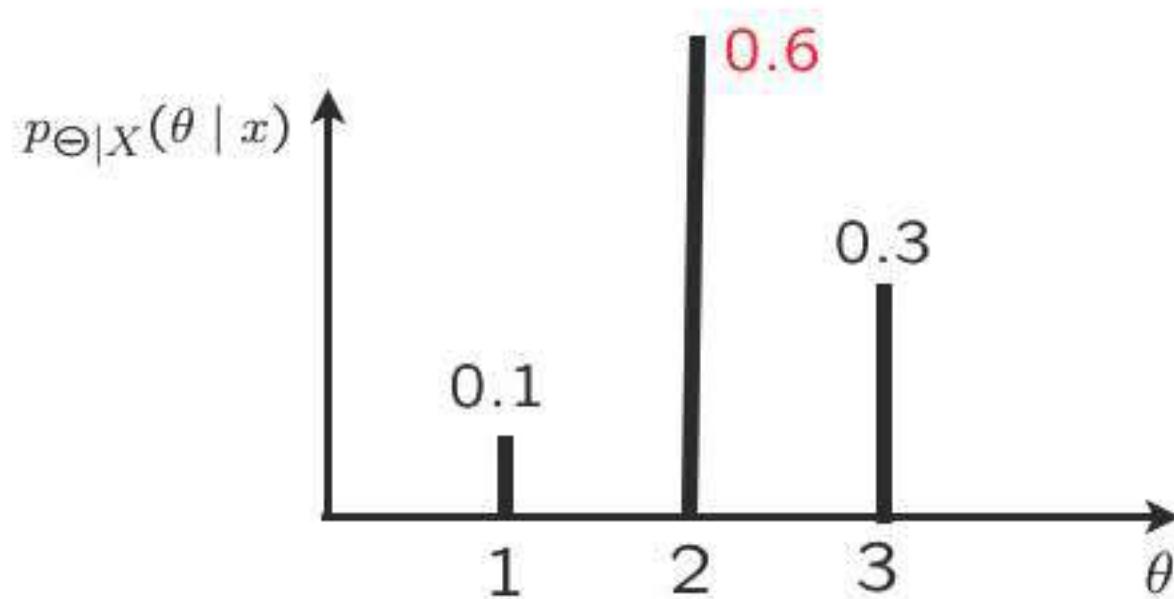
Discrete Θ , continuous X

- Standard example:
 - send signal $\Theta \in \{1, 2, 3\}$

$$X = \Theta + W$$

$W \sim N(0, \sigma^2)$, indep. of Θ

$$f_{X|\Theta}(x | \theta) = f_W(x - \theta)$$



- MAP rule: $\hat{\theta} = 2$

$$p_{\Theta|X}(\theta | x) = \frac{p_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{f_X(x)}$$

$$f_X(x) = \sum_{\theta'} p_{\Theta}(\theta') f_{X|\Theta}(x | \theta')$$

- conditional prob of error:

$$P(\hat{\theta} \neq \Theta | X = x)$$

smallest under the MAP rule

- overall probability of error:

$$\begin{aligned} P(\hat{\Theta} \neq \Theta) &= \int \underbrace{P(\hat{\Theta} \neq \Theta | X = x)}_{\text{smallest under the MAP rule}} f_X(x) dx \\ &= \sum_{\theta} P(\hat{\Theta} \neq \theta | \Theta = \theta) p_{\Theta}(\theta) \end{aligned}$$

Continuous Θ , continuous X

- linear normal models
estimation of a noisy signal

$$X = \Theta + W$$

Θ and W : independent normals

multi-dimensional versions (many normal parameters, many observations)

- estimating the parameter of a uniform

X : `uniform[0, Θ]`

Θ : `uniform [0, 1]`

$$\underline{f_{\Theta|X}(\theta | x)} = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{f_X(x)}$$

$$f_X(x) = \int f_{\Theta}(\theta') f_{X|\Theta}(x | \theta') d\theta'$$

- $\widehat{\Theta} = g(X)$ *MAP*
LMS

- interested in:

$$\left\{ \begin{array}{l} E[(\widehat{\Theta} - \Theta)^2 | X = x] \\ E[(\widehat{\Theta} - \Theta)^2] \end{array} \right.$$

Inferring the unknown bias of a coin and the Beta distribution

- Standard example:
 - coin with bias Θ ; prior $f_\Theta(\cdot)$
 - fix n ; K = number of heads
- Assume $f_\Theta(\cdot)$ is uniform in $[0, 1]$

$$f_{\Theta|K}(\theta | k) = \frac{f_\Theta(\theta) p_{K|\Theta}(k | \theta)}{p_K(k)}$$

$$p_K(k) = \int f_\Theta(\theta') p_{K|\Theta}(k | \theta') d\theta'$$

$$f_{\Theta|K}(\theta | k) = \frac{1 \cdot \binom{n}{k} \theta^k (1-\theta)^{n-k}}{p_K(k)}$$

$$= \frac{1}{d(n, k)} \theta^k (1-\theta)^{n-k}$$

“Beta distribution, with parameters $(k+1, n-k+1)$ ”

$$\theta \in [0, 1]$$

$$\bullet \text{ If prior is Beta: } f_\Theta(\theta) = \frac{1}{c} \theta^\alpha (1-\theta)^\beta \quad \alpha, \beta > 0$$

$$f_{\Theta|K}(\theta | k) = \frac{\frac{1}{c} \theta^\alpha (1-\theta)^\beta \binom{n}{k} \theta^k (1-\theta)^{n-k}}{p_K(k)} = d \theta^{\alpha+k} (1-\theta)^{\beta+n-k}$$

Inferring the unknown bias of a coin: point estimates

- Standard example:
 - coin with bias Θ ; prior $f_\Theta(\cdot)$
 - fix n ; K = number of heads
- Assume $f_\Theta(\cdot)$ is uniform in $[0, 1]$

$$f_{\Theta|K}(\theta | k) = \frac{1}{d(n, k)} \underline{\theta^k (1-\theta)^{n-k}}$$

- MAP estimate:

$$\hat{\theta}_{\text{MAP}} = \boxed{k/n}$$

$$\max_{\Theta} [k \log \Theta + (n-k) \log (1-\Theta)]$$

$$\frac{\partial}{\partial \Theta} \frac{k/\Theta - (n-k)/(1-\Theta)}{} = 0$$

$$\hat{\Theta}_{\text{MAP}} = \boxed{k/n}$$

$$\int_0^1 \theta^\alpha (1-\theta)^\beta d\theta = \frac{\alpha! \beta!}{(\alpha+\beta+1)!} \quad \begin{matrix} \alpha \geq 0 \\ \beta \geq 0 \end{matrix}$$

$$\begin{aligned} E[\Theta | K = k] &= \int_0^1 \theta f_{\Theta|K}(\theta | k) d\theta \\ &= \frac{1}{d(n, k)} \int_0^1 \theta^{k+1} (1-\theta)^{n-k} d\theta \\ &= \frac{1}{\cancel{k!} \cancel{(n-k)!} (n+1)!} \cdot \frac{(k+1)! (n-k)!}{(n+2)!} \\ &= \boxed{\frac{k+1}{n+2}} \approx \frac{k}{n} \quad (\text{large } n) \end{aligned}$$

Summary

- Problem data: $p_{\Theta}(\cdot)$, $p_{X|\Theta}(\cdot | \cdot)$
- Given the value x of X : find, e.g., $p_{\Theta|X}(\cdot | x)$
 - using appropriate version of the Bayes rule (4 choices)
- Estimator $\widehat{\Theta} = g(X)$ Estimate $\widehat{\theta} = g(x)$
 - MAP: $\widehat{\theta}_{\text{MAP}} = g_{\text{MAP}}(x)$ maximizes $p_{\Theta|X}(\theta | x)$
 - LMS: $\widehat{\theta}_{\text{LMS}} = g_{\text{LMS}}(x) = \mathbb{E}[\Theta | X = x]$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 15: Linear models with normal noise

$$X_i = \sum_{j=1}^m a_{ij} \Theta_j + W_i$$

W_i, Θ_j : independent, normal

- Very common and convenient model
- Bayes' rule: normal posteriors
- MAP and LMS estimates coincide
 - simple formulas
(linear in the observations)
- Many nice properties
- Trajectory estimation example

Recognizing normal PDFs

$$X \sim N(\mu, \sigma^2)$$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$2\alpha x + \beta = 0$$

$$c \cdot e^{-8(x-3)^2}$$

$$\mu = 3$$

$$\frac{1}{2\sigma^2} = 8 \Rightarrow \sigma^2 = \frac{1}{16}$$

$$c = \frac{1}{\frac{1}{4}\sqrt{2\pi}}$$

$$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)}$$

$$\alpha > 0$$

Normal with mean $-\beta/2\alpha$ and variance $1/2\alpha$

$$\alpha x^2 + \beta x + \gamma = \alpha \left(x^2 + \frac{\beta}{\alpha} x + \frac{\gamma}{\alpha} \right) = \alpha \left(\left(x + \frac{\beta}{2\alpha} \right)^2 - \frac{\beta^2}{4\alpha^2} + \frac{\gamma}{\alpha} \right)$$

$$f_X(x) = c \cdot e^{-\alpha \left(x + \frac{\beta}{2\alpha} \right)^2} e^{-\alpha \left(-\frac{\beta^2}{4\alpha^2} + \frac{\gamma}{\alpha} \right)}$$

$$\mu = -\frac{\beta}{2\alpha}$$

$$\frac{1}{2\sigma^2} = \alpha \Rightarrow \sigma^2 = 1/2\alpha$$

Estimating a normal random variable in the presence of additive normal noise

$$X = \Theta + W \quad \Theta, W : N(0, 1), \text{ independent}$$

$$f_{X|\Theta}(x|\theta) : X = \theta + W \quad N(\theta, 1)$$

$$f_{\Theta|X}(\theta|x) = \frac{1}{f_X(x)} c e^{-\frac{1}{2}\theta^2} c e^{-\frac{1}{2}(x-\theta)^2} = \underline{\underline{c(x)}} e^{-\text{quadratic}(\theta)}$$

$$\text{Fix } x \quad \min_{\theta} \left[\frac{1}{2}\theta^2 + \frac{1}{2}(x-\theta)^2 \right]$$

$$\text{Normals!} \\ \theta + (\theta - x) = 0$$

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{LMS}} = E[\Theta | X = x] = x/2$$

$$\widehat{\Theta}_{\text{MAP}} = E[\Theta | X] = x/2$$

Estimating a normal parameter in the presence of additive normal noise

$$X = \Theta + W \quad \Theta, W : N(0, 1), \text{ independent}$$

$$\widehat{\Theta}_{\text{MAP}} = \widehat{\Theta}_{\text{LMS}} = \mathbb{E}[\Theta | X] = \frac{X}{2}$$

- Even with general means and variances:
 - posterior is normal
 - LMS and MAP estimators coincide
 - these estimators are “linear,” of the form $\widehat{\Theta} = aX + b$

$$f_{\Theta|X}(\theta | x) = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{f_X(x)}$$

$$f_X(x) = \int f_{\Theta}(\theta) f_{X|\Theta}(x | \theta) d\theta$$

The case of multiple observations

$$\begin{aligned} X_1 &= \Theta + W_1 & \Theta \sim N(x_0, \sigma_0^2) & W_i \sim N(0, \sigma_i^2) \\ &\vdots \\ X_n &= \Theta + W_n & \Theta, W_1, \dots, W_n \text{ independent} \end{aligned}$$

$$f_{X_i|\Theta}(x_i|\theta) = c_i e^{-\frac{(x_i - \theta)^2}{2\sigma_i^2}}$$

given $\Theta = \theta$: $X_i = \theta + W_i \sim N(\theta, \sigma_i^2)$

$$f_{X|\Theta}(x|\theta) = f_{X_1, \dots, X_n|\Theta}(x_1, \dots, x_n|\theta) = \prod_{i=1}^n f_{X_i|\Theta}(x_i|\theta)$$

given $\Theta = \theta$: W_i independent $\Rightarrow X_i$ independent

$$f_{\Theta|X}(\theta|x) = \frac{1}{f_X(x)} \cdot c_0 e^{-\frac{(\theta-x_0)^2}{2\sigma_0^2}} \prod_{i=1}^n c_i e^{-\frac{(x_i - \theta)^2}{2\sigma_i^2}}$$

Normal!

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x|\theta) d\theta$$

The case of multiple observations

$$f_{\Theta|X}(\theta|x) = c \cdot \exp \left\{ -\text{quad}(\theta) \right\} \quad \text{quad}(\theta) = \frac{(\theta - x_0)^2}{2\sigma_0^2} + \frac{(\theta - x_1)^2}{2\sigma_1^2} + \cdots + \frac{(\theta - x_n)^2}{2\sigma_n^2}$$

find peak

$$\frac{d}{d\theta} \text{quad}(\theta) = 0: \sum_{i=0}^n \frac{(\theta - x_i)}{\sigma_i^2} = 0 \Rightarrow \theta \sum_{i=0}^n \frac{1}{\sigma_i^2} = \sum_{i=0}^n \frac{x_i}{\sigma_i^2}$$

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{LMS}} = \mathbb{E}[\Theta | X = \mathbf{x}] = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

The case of multiple observations

- Key conclusions:
 - posterior is normal
 - LMS and MAP estimates coincide
 - these estimates are “linear,” of the form $\hat{\theta} = a_0 + a_1x_1 + \cdots + a_nx_n$
- Interpretations:
 - estimate $\hat{\theta}$: weighted average of x_0 (prior mean) and x_i (observations)
 - weights determined by variances

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{LMS}} = E[\Theta | X = x] = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

σ_i^2 large
 x_i very noisy
 \Rightarrow small weight

The mean squared error

$$f_{\Theta|X}(\theta | x) = c \cdot \exp \left\{ -\text{quad}(\theta) \right\}$$

$$X_i = \tilde{\theta} + w_i$$

$$\text{quad}(\theta) = \frac{(\theta - x_0)^2}{2\sigma_0^2} + \frac{(\theta - x_1)^2}{2\sigma_1^2} + \dots + \frac{(\theta - x_n)^2}{2\sigma_n^2}$$

$$\hat{\theta} = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

- Performance measures:

$$\mathbb{E}[(\Theta - \hat{\Theta})^2 | X = x] = \mathbb{E}[(\Theta - \hat{\theta})^2 | X = x] = \text{var}(\Theta | X = x) = \boxed{1 / \sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

$$\mathbb{E}[(\Theta - \hat{\Theta})^2] = \int \underbrace{\mathbb{E}[(\Theta - \hat{\theta})^2 | X = x]}_{=} f_x(x) dx$$

$$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)} \quad \alpha > 0 \quad \text{Normal with mean } -\beta/2\alpha \text{ and variance } 1/2\alpha$$

$$\alpha = \frac{1}{2\sigma_0^2} + \dots + \frac{1}{2\sigma_n^2}$$

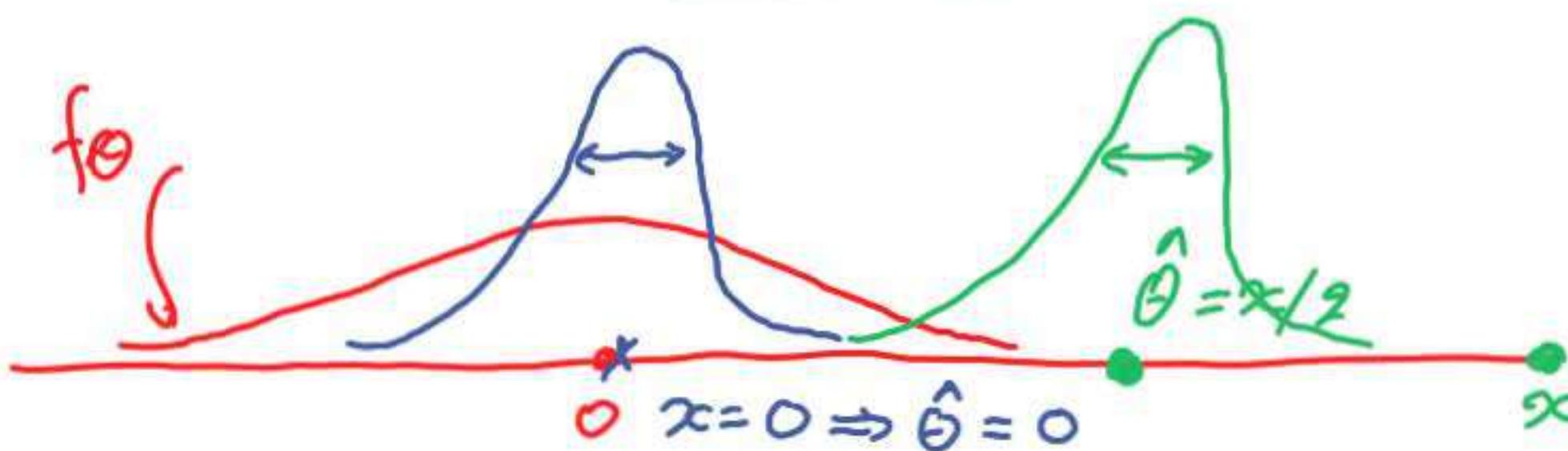
some σ_i^2 small \rightarrow MSE small
all σ_i^2 large \rightarrow MSE large

The mean squared error

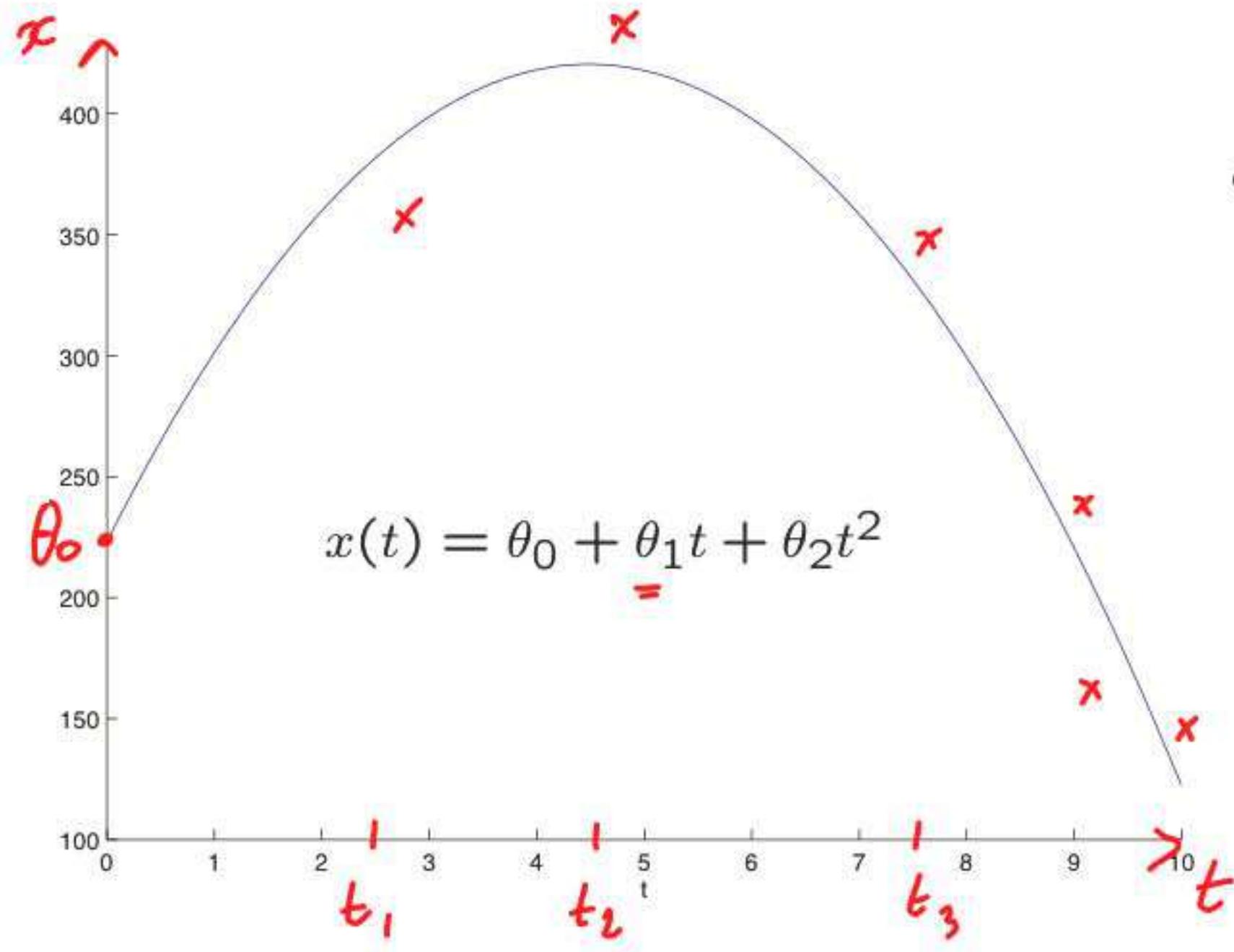
$$E[(\Theta - \hat{\Theta})^2 | X = \underline{x}] \underset{?}{=} E[(\Theta - \hat{\Theta})^2] = 1 / \sum_{i=0}^n \frac{1}{\sigma_i^2}$$

$$\hat{\theta} = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

- Example: $\sigma_0^2 = \sigma_1^2 = \dots = \sigma_n^2 = \sigma^2$ $\frac{1}{(n+1) \frac{1}{\sigma^2}} = \frac{\sigma^2}{n+1}$
- conditional mean squared error same for all x
- Example: $X = \Theta + W$ $\Theta \sim N(0, 1)$, $W \sim N(0, 1)$
independent Θ, W $\hat{\Theta} = X/2$ $E[(\Theta - \hat{\Theta})^2 | X = \underline{x}] = \underline{1/2}$



The case of multiple parameters: trajectory estimation



- Random variables $\Theta_0, \Theta_1, \Theta_2$ independent; priors f_{Θ_j}
- Measurements at times t_1, \dots, t_n
$$X_i = \underline{\Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2} + W_i$$
noise model: f_{W_i} independent W_i ; independent from Θ_j

A model with normality assumptions

$$X_i = \underline{\Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2} + W_i \quad i = 1, \dots, n$$

$$f_{\Theta|X}(\underline{\theta} | \underline{x}) = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{f_X(x)}$$

$$f_X(x) = \int f_{\Theta}(\theta) f_{X|\Theta}(x | \theta) d\theta$$

- assume $\Theta_j \sim N(0, \sigma_j^2)$, $W_i \sim N(0, \sigma^2)$; independent

- Given $\Theta = \theta = (\theta_0, \theta_1, \theta_2)$, X_i is: $N(\theta_0 + \theta_1 t_i + \theta_2 t_i^2, \sigma^2)$

$$f_{X_i|\Theta}(x_i | \theta) = c \cdot \exp \left\{ - \underbrace{(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2 / 2\sigma^2}_{\text{red}} \right\}$$

- posterior: $f_{\Theta|X}(\theta | x) = \frac{1}{f_X(x)} \prod_{j=0}^2 f_{\Theta_j}(\theta_j) \prod_{i=1}^n f_{X_i|\Theta}(x_i | \theta)$

$$c(x) \exp \left\{ - \frac{1}{2} \left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^n \underbrace{(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2}_{\text{red}} \right\}$$

A model with normality assumptions

$$\underline{f_{\Theta|X}(\theta|x)} = c(x) \exp \left\{ -\frac{1}{2} \left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2 \right\}$$

- MAP estimate: maximize over $(\theta_0, \theta_1, \theta_2)$;
(minimize quadratic function)

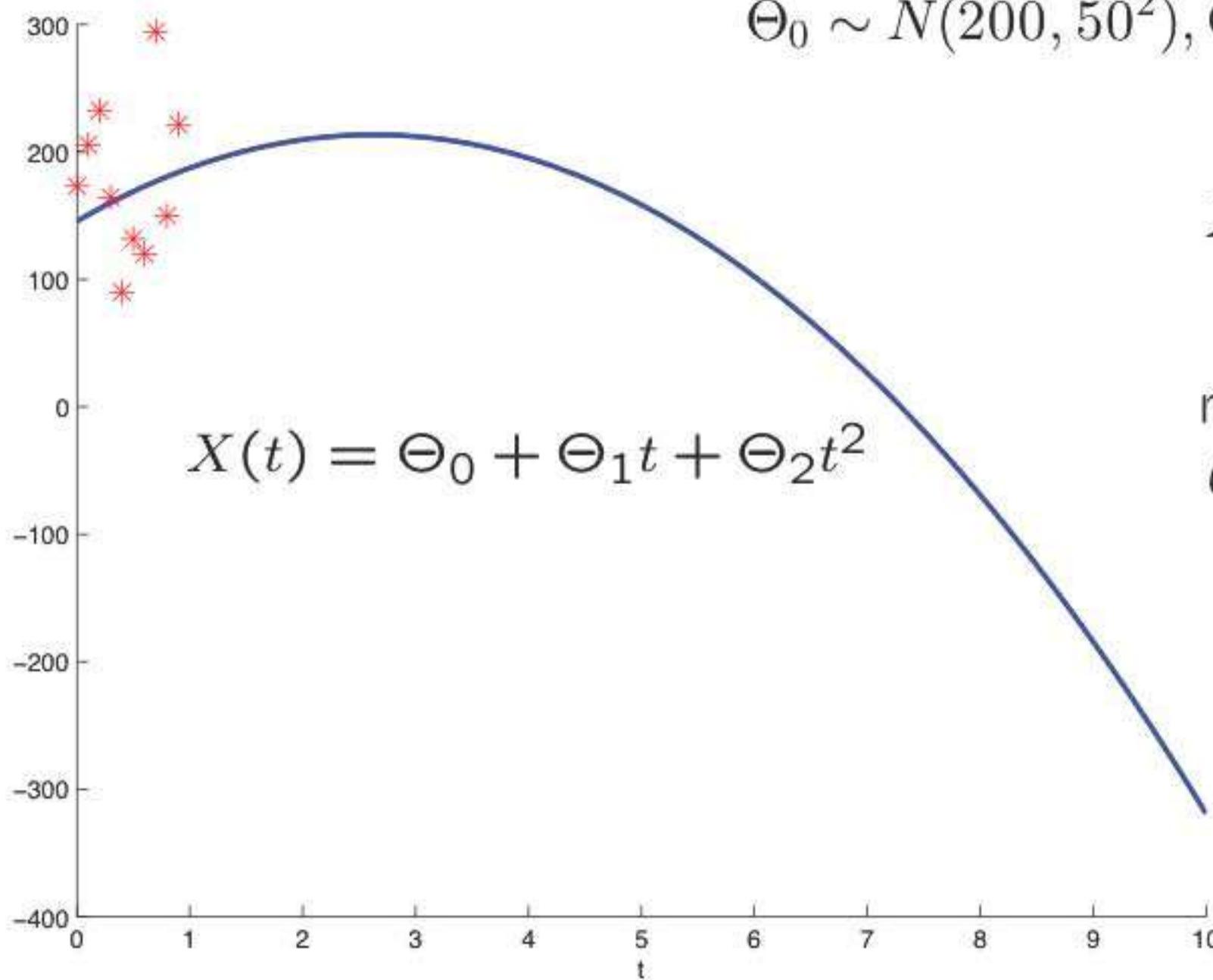
$$\frac{\partial}{\partial \theta_j} (\text{quad}(\theta)) = 0 \quad \begin{matrix} 3 \text{ equations, 3 unknowns} \\ \uparrow \text{linear} \end{matrix} .$$

Linear normal models •

- Θ_j and X_i are linear functions of independent normal random variables
- $f_{\Theta|X}(\theta | x) = c(x) \exp \left\{ -\text{quadratic}(\theta_1, \dots, \theta_m) \right\}$ *linear regression*
- MAP estimate: maximize over $(\theta_1, \dots, \theta_m)$:
(minimize quadratic function) *linear equations*
- $\widehat{\Theta}_{\text{MAP},j}$: linear function of $X = (X_1, \dots, X_n)$
- Facts:
 - $\widehat{\Theta}_{\text{MAP},j} = \mathbf{E}[\Theta_j | X]$
 - marginal posterior PDF of Θ_j : $f_{\Theta_j|X}(\theta_j | x)$, is normal
 - MAP estimate based on the joint posterior PDF:
same as MAP estimate based on the marginal posterior PDF
 - $\mathbf{E}[(\widehat{\Theta}_{i,\text{MAP}} - \Theta_i)^2 | X = x]$: same for all x

An illustration

Estimating the trajectory of a free-falling object



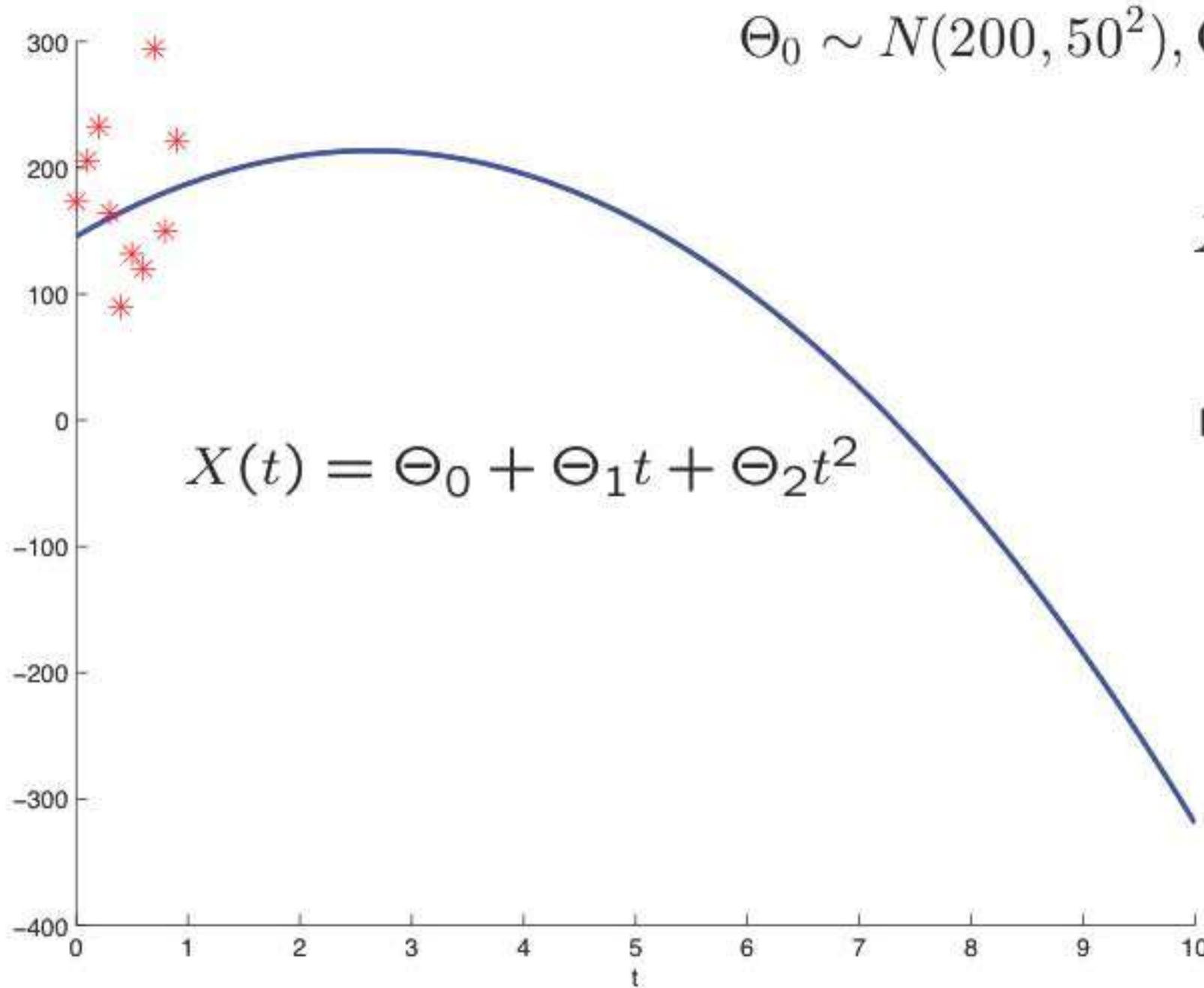
$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

$$\begin{aligned} & \text{minimize} \\ & \theta_0, \theta_1, \theta_2 \end{aligned}$$

$$\begin{aligned} & \frac{1}{2} \left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2} \right) \\ & + \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2 \end{aligned}$$

An illustration

Estimating the trajectory of a free-falling object



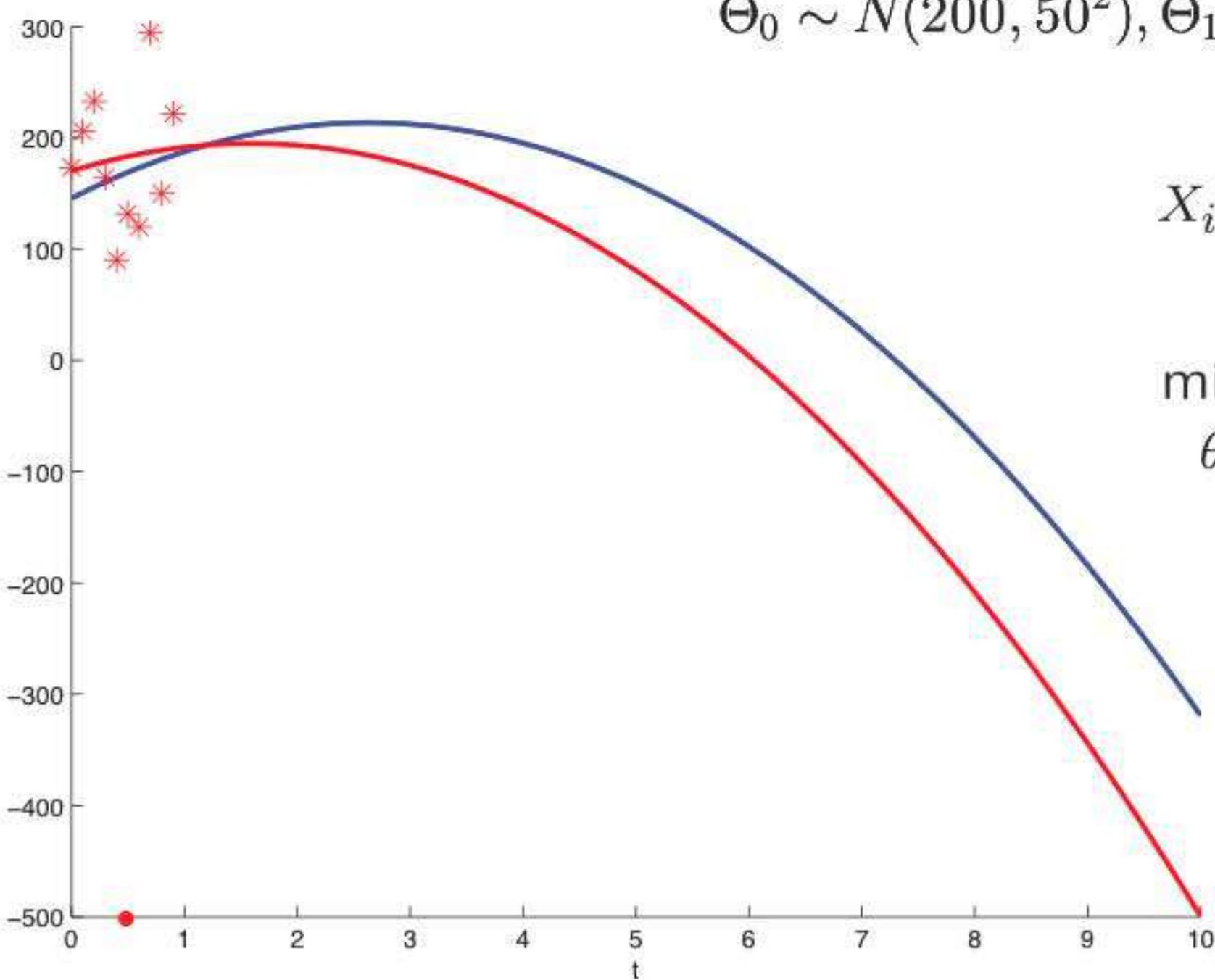
$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

$$\begin{aligned} & \text{minimize}_{\theta_0, \theta_1} && (\theta_0 - 200)^2 + (\theta_1 - 50)^2 \\ & && + \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2 \end{aligned}$$

An illustration

Estimating the trajectory of a free-falling object



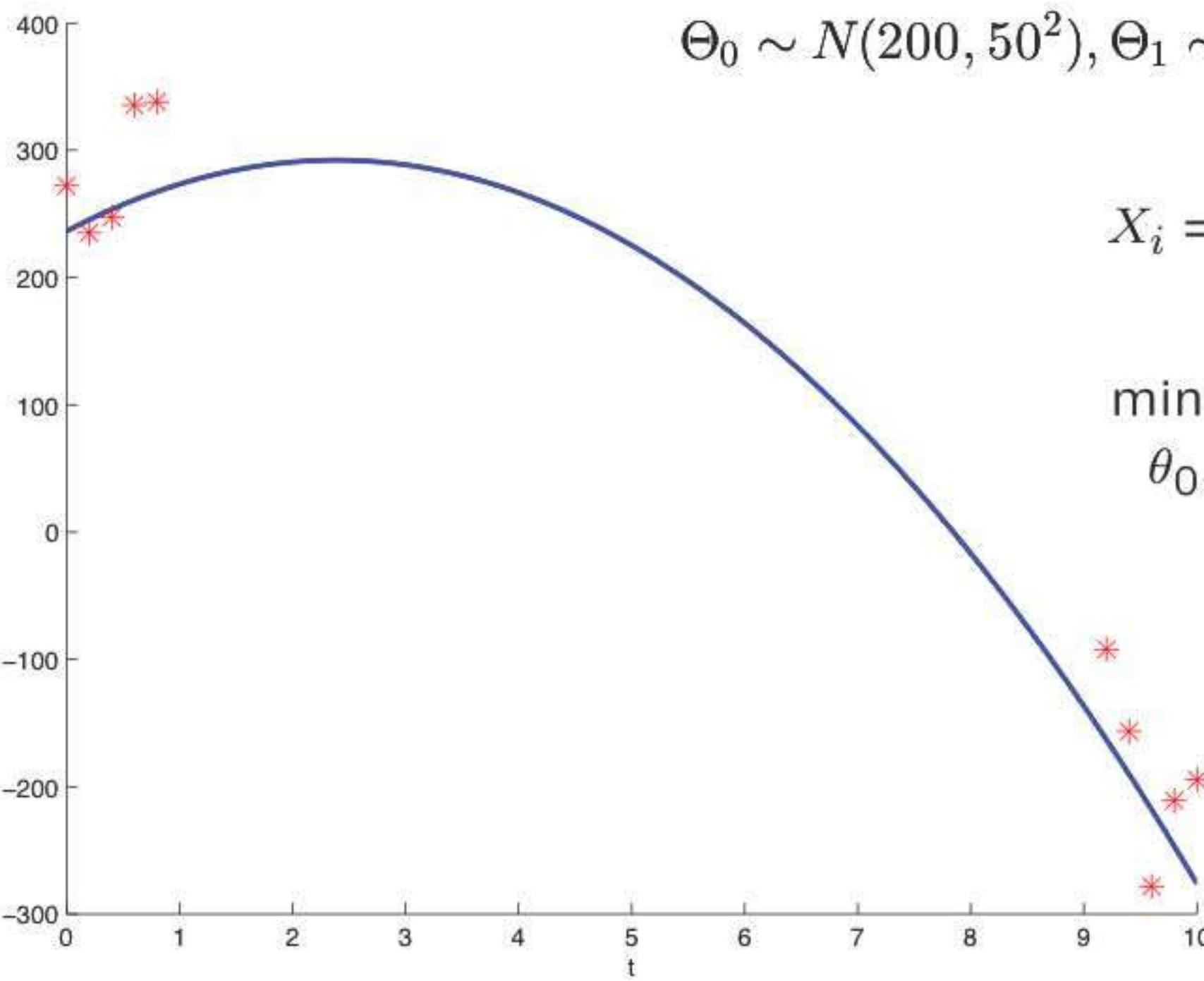
$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

$$\begin{aligned} & \text{minimize}_{\theta_0, \theta_1} && (\theta_0 - 200)^2 + (\theta_1 - 50)^2 \\ & && + \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2 \end{aligned}$$

An illustration

Estimating the trajectory of a free-falling object



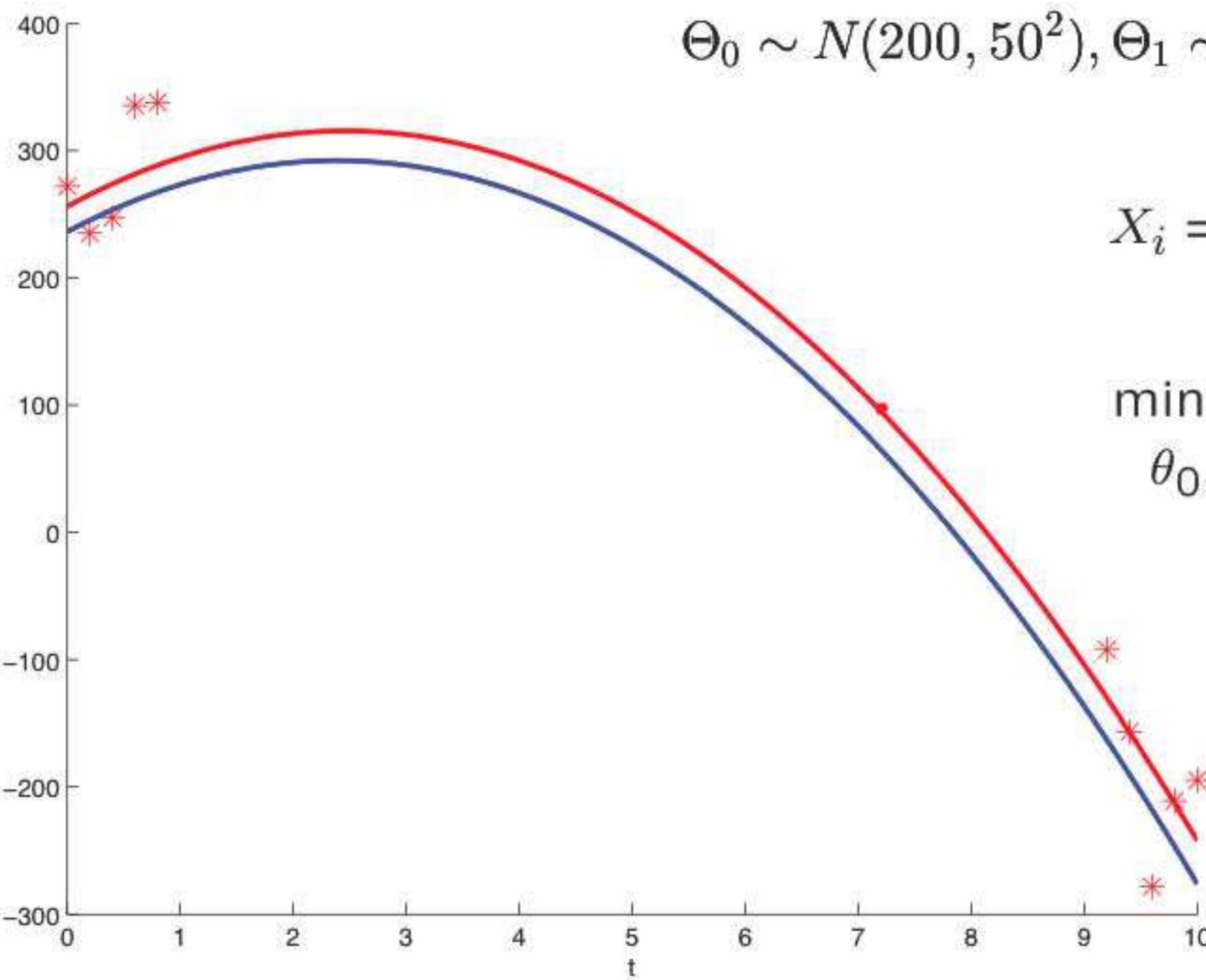
$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

$$\begin{aligned} & \text{minimize}_{\theta_0, \theta_1} && (\theta_0 - 200)^2 + (\theta_1 - 50)^2 \\ & && + \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2 \end{aligned}$$

An illustration

Estimating the trajectory of a free-falling object



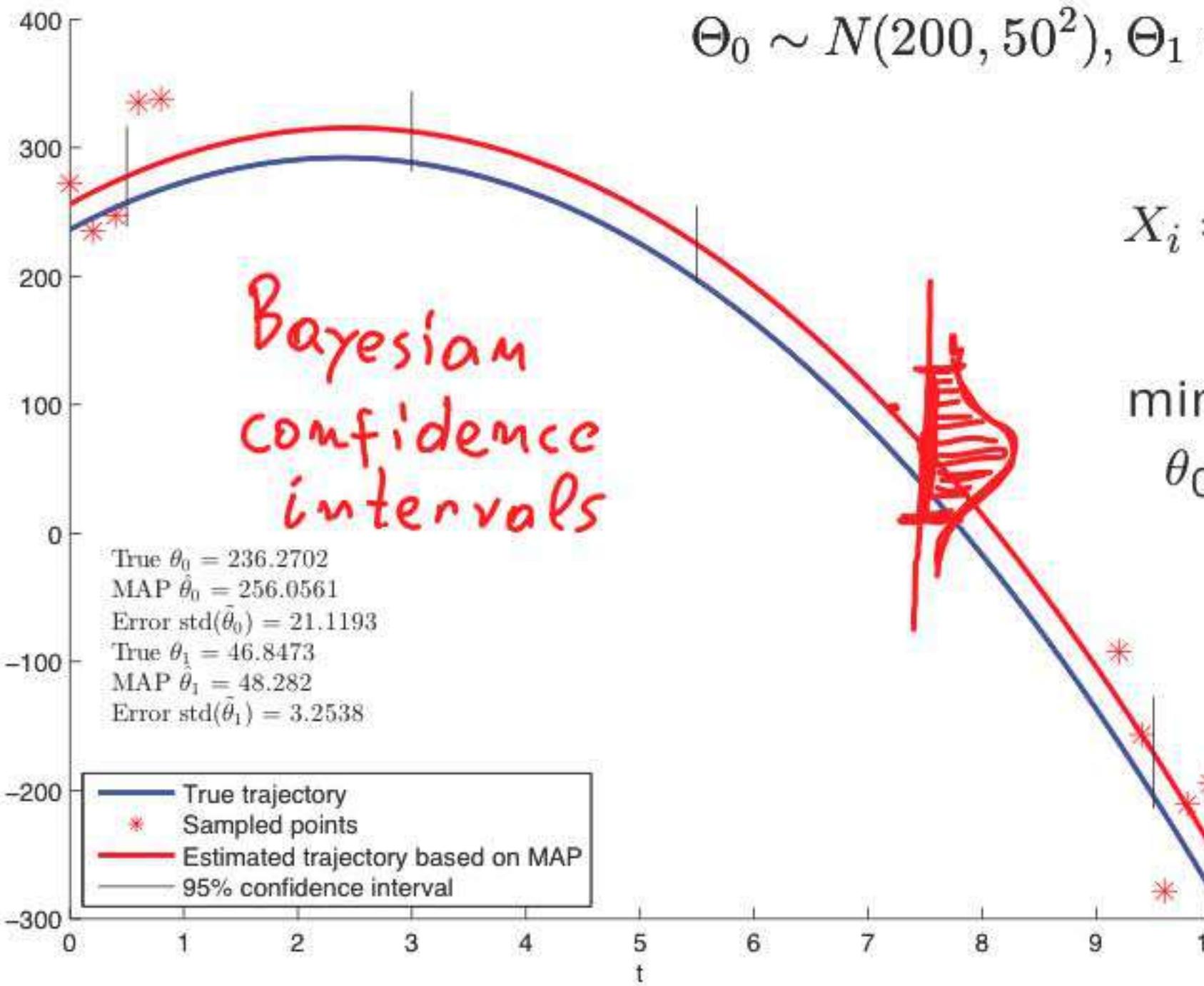
$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$x(t) \\ X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

$$\begin{aligned} & \text{minimize}_{\theta_0, \theta_1} && (\theta_0 - 200)^2 + (\theta_1 - 50)^2 \\ & && + \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2 \end{aligned}$$

An illustration

Estimating the trajectory of a free-falling object



$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$x(t) \\ X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

$$\begin{aligned} & \text{minimize}_{\theta_0, \theta_1} && (\theta_0 - 200)^2 + (\theta_1 - 50)^2 \\ & &+ \sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2 \end{aligned}$$

$$P(x(t) \in \text{interval} | \text{data}) = 0.95$$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

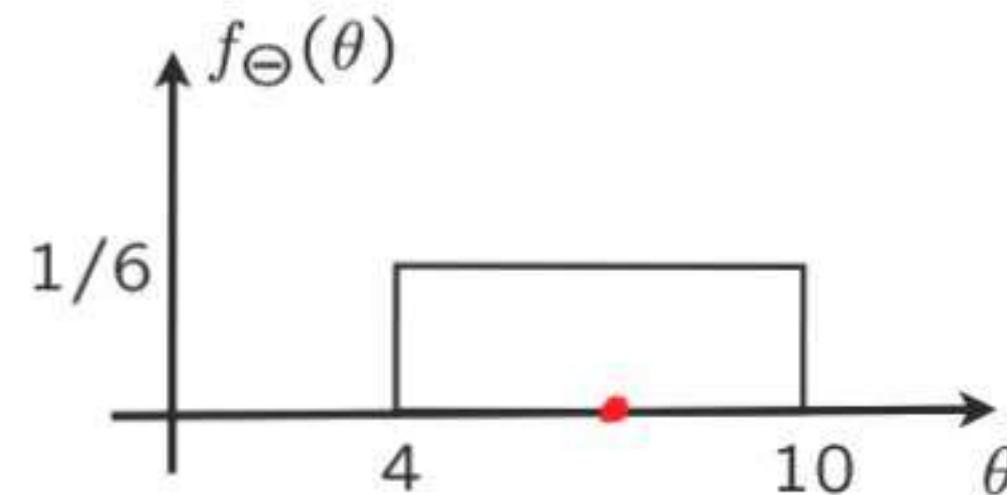
For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 16: Least mean squares (LMS) estimation

- minimize (conditional) mean squared error $E[(\Theta - \hat{\theta})^2 | X = x]$
 - solution: $\hat{\theta} = E[\Theta | X = x]$
 - general estimation method
- Mathematical properties
- Example

LMS estimation in the absence of observations

- unknown Θ ; prior $p_\Theta(\theta)$
 - interested in a point estimate $\hat{\theta}$
 - no observations available
 - MAP rule: $\text{any } \hat{\theta} \in [4, 10]$
 - (Conditional) expectation: $\hat{\theta} = 7$
- Criterion: Mean Squared Error (MSE): $E[(\Theta - \hat{\theta})^2]$
minimize mean squared error



LMS estimation in the absence of observations

- Least mean squares formulation:

minimize mean squared error (MSE), $E[(\Theta - \hat{\theta})^2]$: $\hat{\theta} = E[\Theta]$.

$$E[\Theta^2] - 2E[\Theta]\hat{\theta} + \hat{\theta}^2 \quad \frac{d}{d\hat{\theta}} = 0 : -2E[\Theta] + 2\hat{\theta} = 0$$
$$\hat{\theta} = E[\Theta]$$

$$\text{Var}(\Theta - \hat{\theta}) + (E[\Theta - \hat{\theta}])^2$$

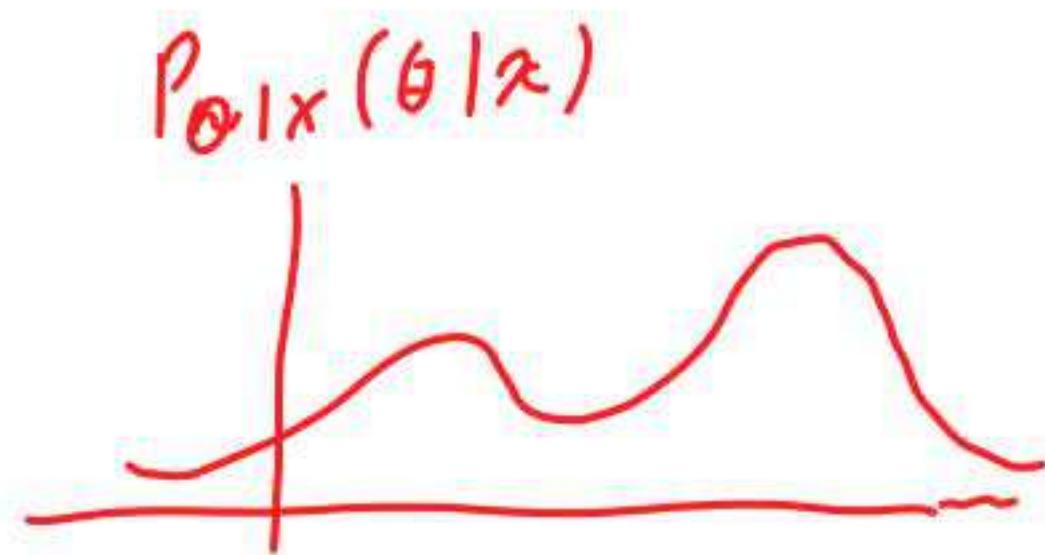
minimized
when $\hat{\theta} = E[\Theta]$

$\text{Var}''(\Theta)$

- Optimal mean squared error: $E[(\Theta - E[\Theta])^2] = \text{var}(\Theta)$

LMS estimation of Θ based on X

- unknown Θ ; prior $p_\Theta(\theta)$
 - interested in a point estimate $\hat{\theta}$
- observation X ; model $p_{X|\Theta}(x|\theta)$
 - observe that $X = x$



minimize mean squared error (MSE), $E[(\Theta - \hat{\theta})^2]$: $\hat{\theta} = E[\Theta]$

minimize conditional mean squared error, $E[(\Theta - \hat{\theta})^2 | X = x]$: $\hat{\theta} = E[\Theta | X = x]$

- LMS estimate: $\hat{\theta} = E[\Theta | X = x]$

estimator: $\widehat{\Theta} = E[\Theta | X_\bullet]$

LMS estimation of Θ based on X

- $E[\Theta]$ minimizes $E[(\Theta - \hat{\theta})^2]$

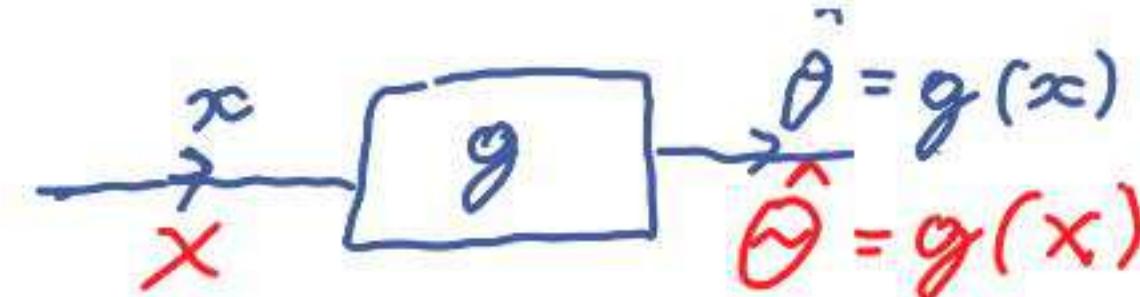
$$E[(\Theta - E[\Theta])^2] \leq E[(\Theta - c)^2], \text{ for all } c$$

- $E[\Theta | X = x]$ minimizes $E[(\Theta - \hat{\theta})^2 | X = x]$

$$E[(\Theta - E[\Theta | X = x])^2 | X = x] \leq E[(\Theta - g(x))^2 | X = x] \text{ for all } x$$

$$E[(\Theta - E[\Theta | X])^2 | X] \leq E[(\Theta - g(X))^2 | X]$$

$$E[(\Theta - \underline{E[\Theta | X]})^2] \leq E[(\Theta - g(X))^2]$$



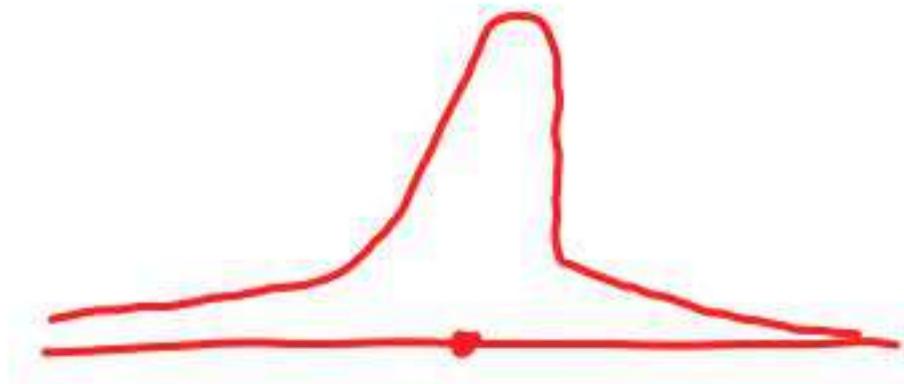
$\widehat{\Theta}_{\text{LMS}} = E[\Theta | X]$ minimizes $E[(\Theta - g(X))^2]$, over all estimators $\widehat{\Theta} = g(X)$

LMS performance evaluation

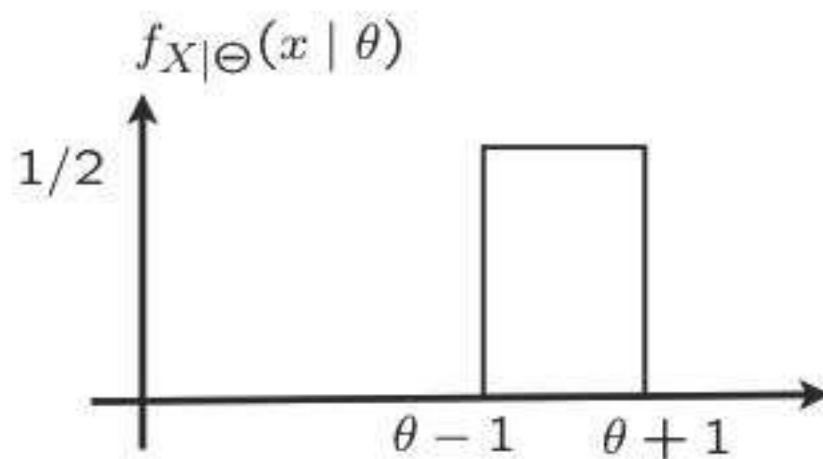
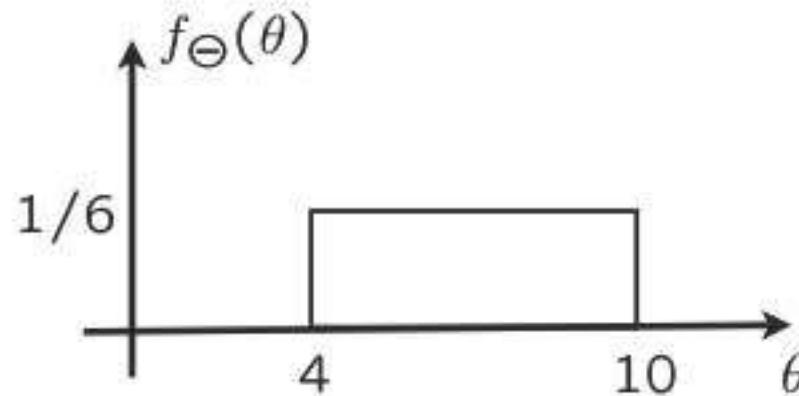
- LMS estimate: $\hat{\theta} = \mathbb{E}[\Theta | X = x]$
estimator: $\widehat{\Theta} = \mathbb{E}[\Theta | X]$
- Expected performance, once we have a measurement:
$$\text{MSE} = \mathbb{E}\left[\left(\Theta - \mathbb{E}[\Theta | X = x]\right)^2 | X = x\right] = \underline{\text{var}(\Theta | X = x)}$$
- Expected performance of the design:
$$\text{MSE} = \mathbb{E}\left[\left(\Theta - \mathbb{E}[\Theta | X]\right)^2\right] = \mathbb{E}\left[\underline{\text{var}(\Theta | X)}\right]$$

LMS estimation of Θ based on X

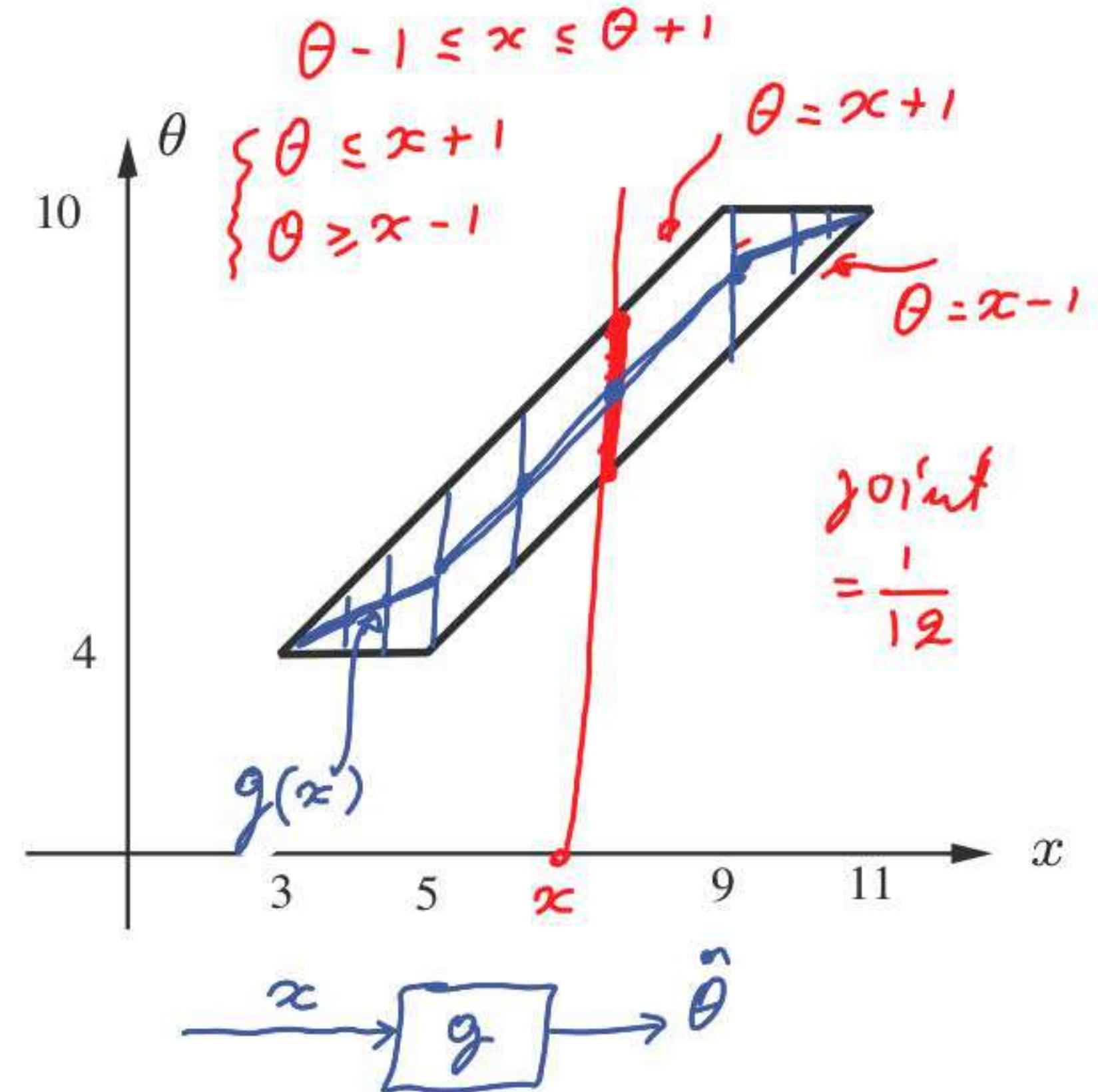
- LMS relevant to estimation (not hypothesis testing)
- Same as MAP if the posterior is unimodal and symmetric around the mean
 - e.g., when posterior is normal (the case in “linear–normal” models)



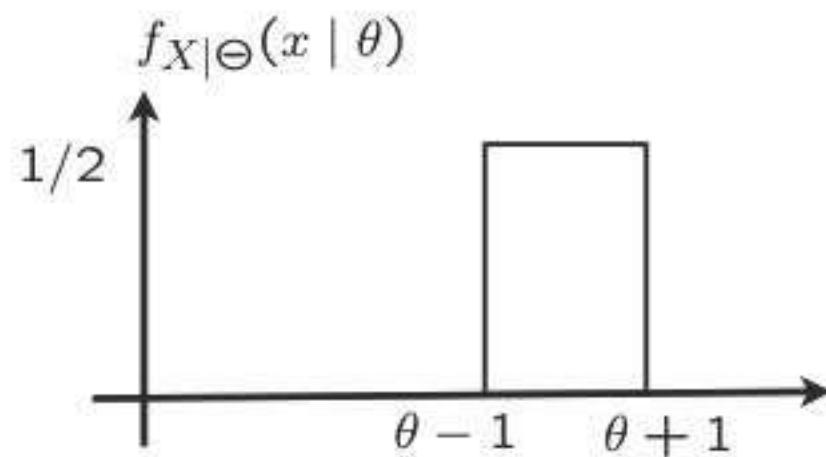
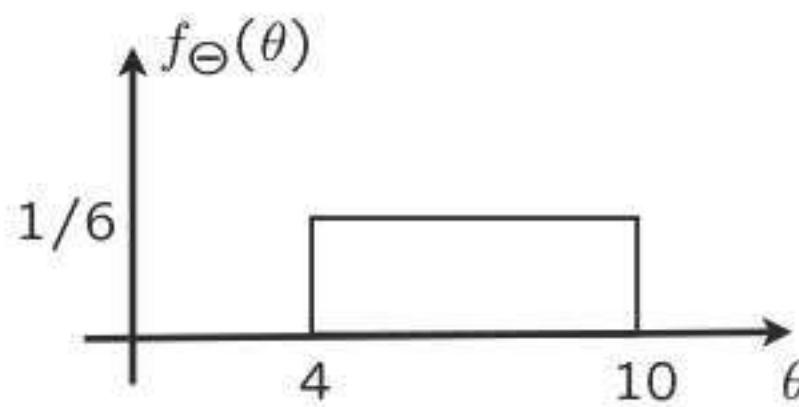
Example



$$\lambda = \theta + U \quad U \sim \text{unif}(-1, 1)$$

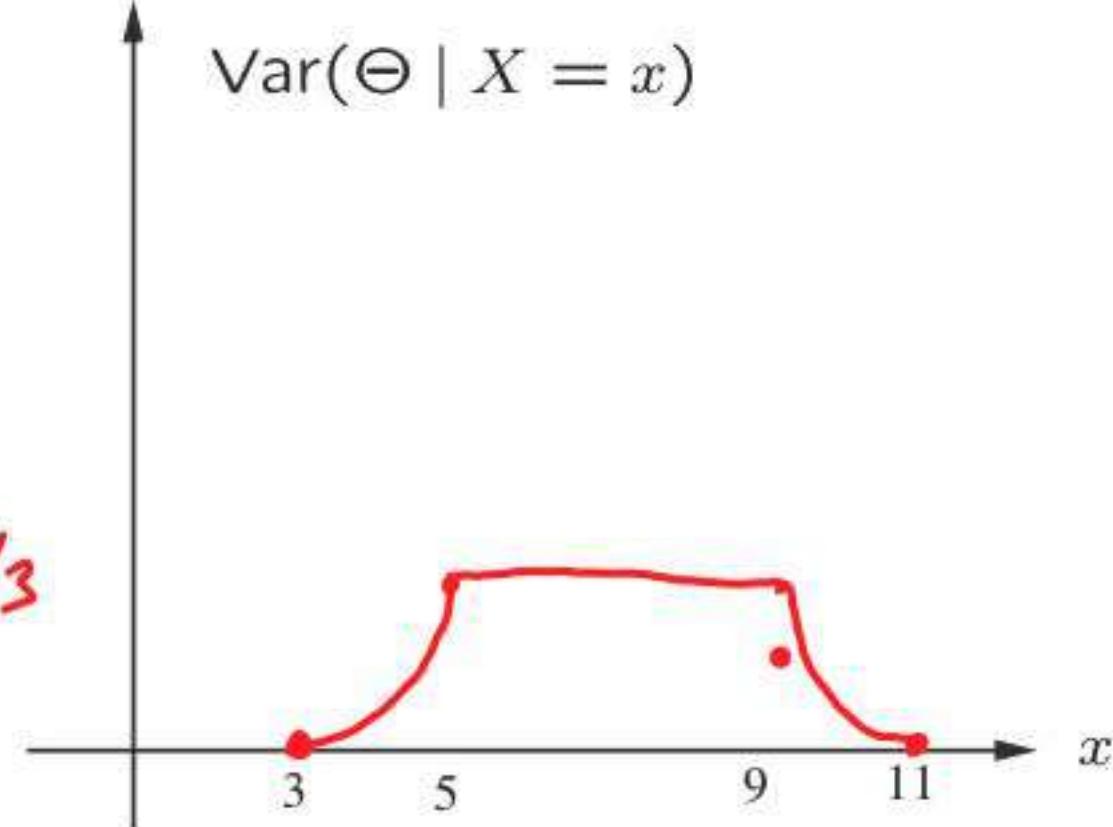
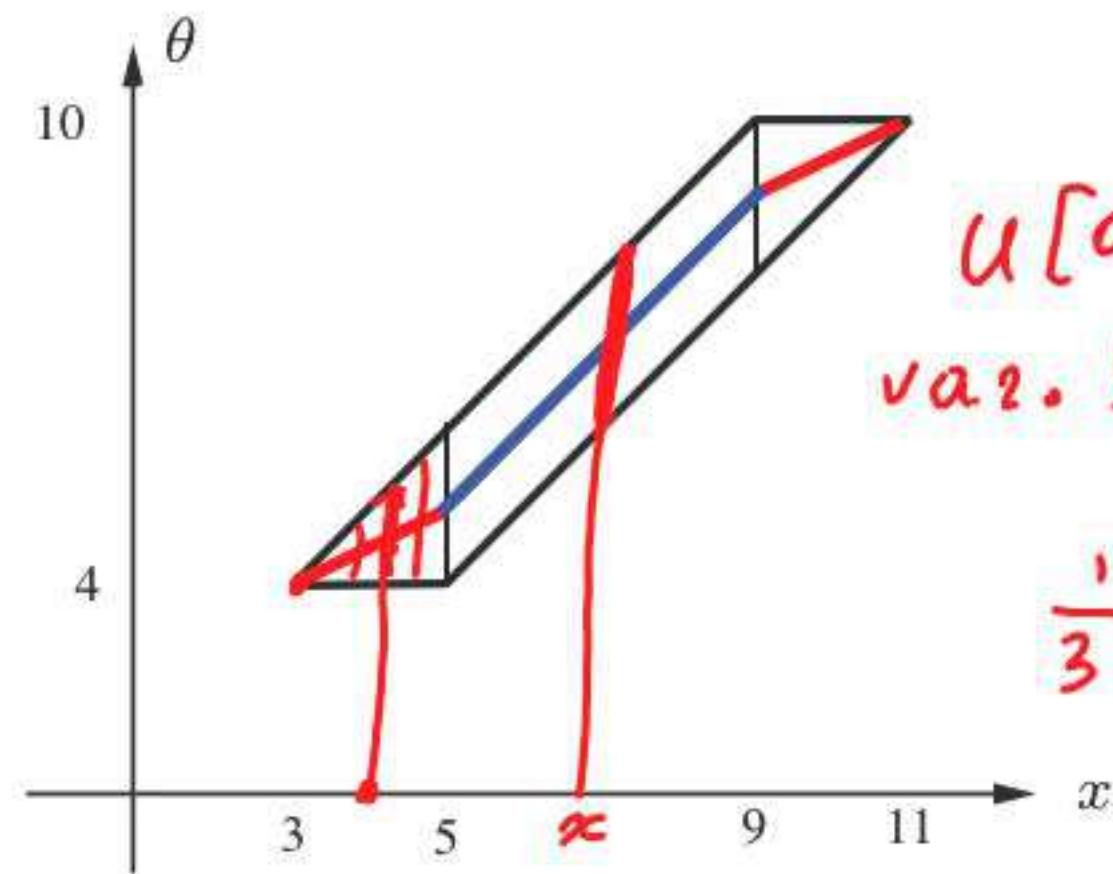


Conditional mean squared error



- $E[(\Theta - E[\Theta | X = x])^2 | X = x]$
 - same as $\text{Var}(\Theta | X = x)$: variance of conditional distribution of Θ

$$E[\text{Var}(\Theta | x)] = \int f_x(x) \text{Var}(\Theta | x=x) dx$$



LMS estimation with multiple observations or unknowns

- unknown Θ ; prior $p_\Theta(\theta)$
 - interested in a point estimate $\hat{\theta}$
- observations $X = (X_1, X_2, \dots, X_n)$; model $p_{X|\Theta}(x | \theta)$
 - observe that $X = x$
 - new universe: condition on $X = x$
- LMS estimate: $E[\Theta | X_1 = x_1, \dots, X_n = x_n]$
- If Θ is a vector, apply to each component separately

$$\Theta = (\Theta_1, \dots, \Theta_m) \quad \hat{\Theta}_j = E[\Theta_j | X_1 = x_1, \dots, X_n = x_n]$$

Some challenges in LMS estimation

$$f_{\Theta|X}(\theta | x) = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x | \theta)}{f_X(x)}$$

$$f_X(x) = \int f_{\Theta}(\theta') f_{X|\Theta}(x | \theta') d\theta'$$

- Full correct model, $f_{X|\Theta}(x | \theta)$, may not be available •
- Can be hard to compute/implement/analyze

$$E[\theta_j | x=x] = \iiint \theta_j \cdot f_{\Theta|X}(\theta | x) d\theta_1 \dots d\theta_m$$

Properties of the estimation error in LMS estimation

- Estimator: $\hat{\theta} = E[\theta | x]$

$$E[\tilde{\theta} | x = x] = 0$$

- Error: $\tilde{\theta} = \hat{\theta} - \theta$

$$E[\hat{\theta}] = E[\theta]$$

$$E[\tilde{\theta}] = 0$$

$$E[\hat{\theta} - \theta | x = x] = \hat{\theta} - E[\theta | x = x] = 0$$

$$\text{cov}(\tilde{\theta}, \hat{\theta}) = 0$$

$$\underbrace{E[\tilde{\theta} \hat{\theta}]}_{= 0} - \cancel{E[\tilde{\theta}] E[\hat{\theta}]} = 0$$

$$E[\tilde{\theta} \hat{\theta} | x = x] = \hat{\theta} E[\tilde{\theta} | x = x] = 0$$

$$\text{var}(\theta) = \text{var}(\hat{\theta}) + \text{var}(\tilde{\theta})$$

•

$$\theta = \hat{\theta} - \tilde{\theta}$$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

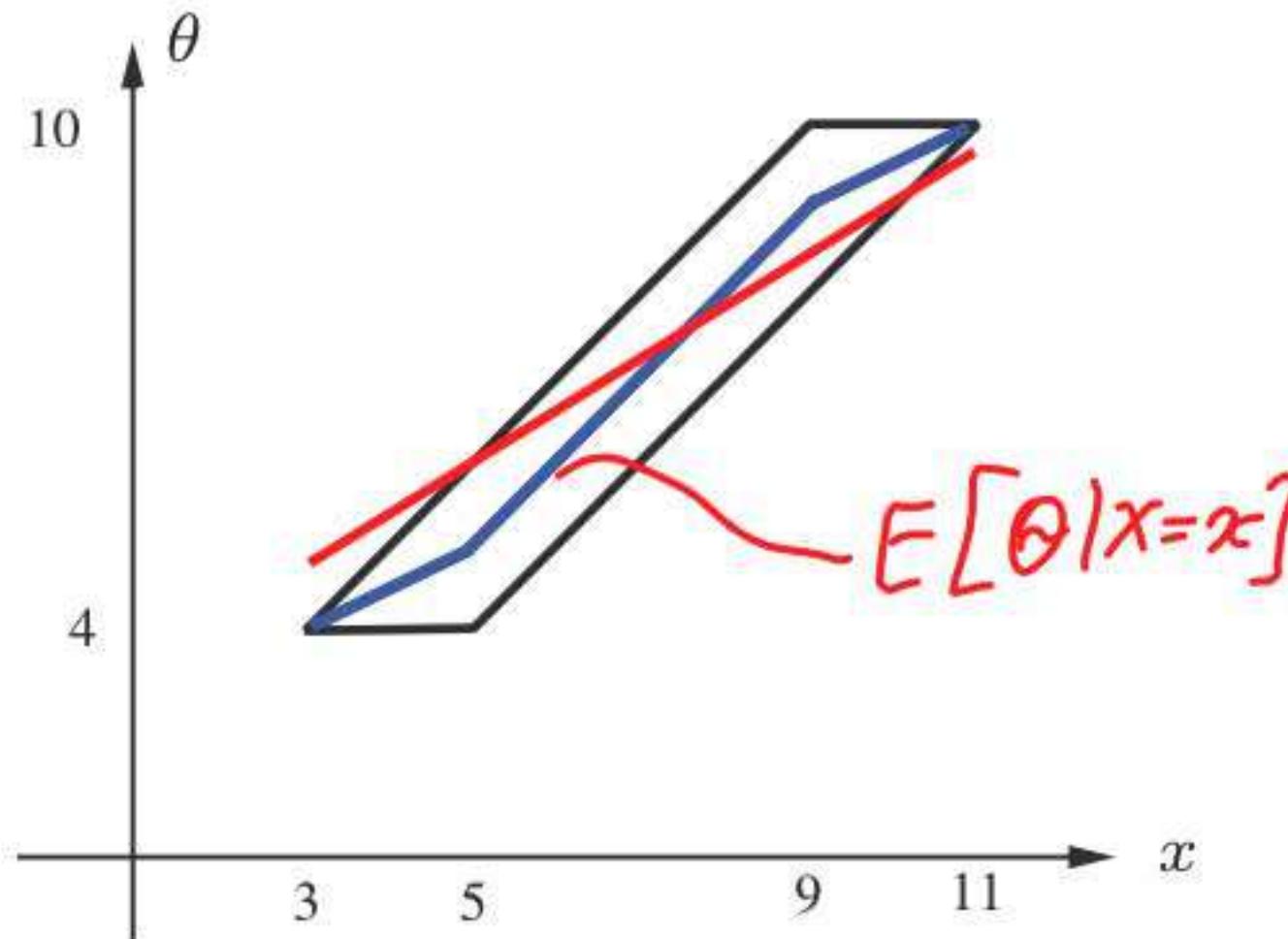
For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 17: Linear least mean squares (LLMS) estimation

- Conditional expectation $E[\Theta | X]$ may be hard to compute/implement
- Restrict to estimators $\widehat{\Theta} = aX + b$
 - minimize mean squared error
- Simple solution
- Mathematical properties
- Example

LLMS formulation

- Unknown Θ ; observation X
- Minimize $E[(\widehat{\Theta} - \Theta)^2]$
- Estimators $\widehat{\Theta} = g(X) \rightarrow \widehat{\Theta}_{LLMS} = E[\Theta | X]$
- Consider estimators of Θ ,
of the form $\widehat{\Theta} = aX + b$
- Minimize $E[(\Theta - aX - b)^2]$, w.r.t. a, b
- If $E[\Theta | X]$ is linear in X , then $\widehat{\Theta}_{LLMS} = \widehat{\Theta}_{LMS}$



Solution to the LLMS problem

- Minimize $E[(\Theta - aX - b)^2]$, w.r.t. a, b

– suppose a has already been found:

$$\begin{aligned} \min \quad & E[(\Theta - aX - E[\Theta - aX])^2] = \text{var}(\Theta - aX) \\ & = \text{var}(\Theta) + a^2 \text{var}(X) - 2a \text{cov}(\Theta, X) \\ \frac{d}{da} \quad & 0 : 2a \text{var}(X) - 2 \text{cov}(\Theta, X) = 0 \quad \left| \begin{array}{l} \rho = \frac{\text{cov}(\Theta, X)}{\sigma_\Theta \sigma_X} \\ a = \frac{\rho \sigma_\Theta \sigma_X}{\sigma_X^2} \end{array} \right. \\ \text{da} \quad & a = \frac{\text{cov}(\Theta, X) / \text{var}(X)}{\sigma_X^2} \end{aligned}$$

$$\widehat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X]) = E[\Theta] + \rho \frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

Remarks on the solution and on the error variance

$$\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X]) = E[\Theta] + \rho \frac{\sigma_\Theta}{\sigma_X} (X - E[X])$$

- Only means, variances, covariances matter

- $\rho > 0: x > E[x] \Rightarrow \hat{\Theta}_L > E[\Theta]$

$$|\rho| = 1$$

- $\rho = 0: \hat{\Theta}_L = E[\Theta]$

$$\hat{\Theta}_L = \Theta$$

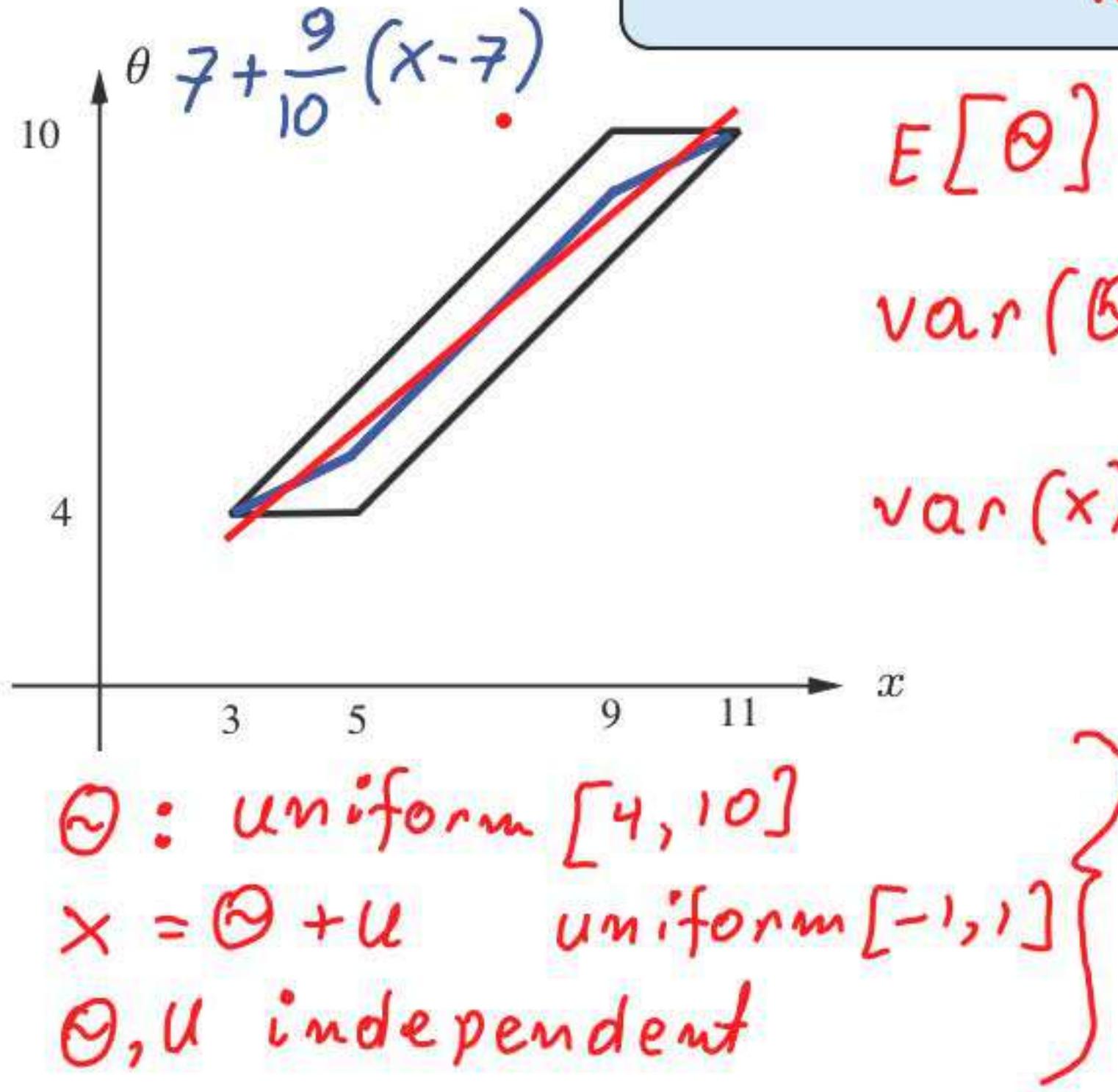
$E[(\hat{\Theta}_L - \Theta)^2] = (1 - \rho^2) \text{var}(\Theta)$

assume $E[\Theta] = E[x] = 0$

$$E[(\Theta - \rho \frac{\sigma_\Theta}{\sigma_x} x)^2] = \sigma_\Theta^2 - 2\rho \frac{\sigma_\Theta}{\sigma_x} \cancel{\rho \sigma_\Theta \sigma_x} + \rho^2 \frac{\sigma_\Theta^2}{\cancel{\sigma_x^2}} \cancel{\sigma_x^2}$$

Example

$$\widehat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X]) = E[\Theta] + \rho \frac{\sigma_\Theta}{\sigma_X} (X - E[X])$$



$$E[\Theta] = 7 \quad E[U] = 0 \quad E[X] = 7$$

$$\text{var}(\Theta) = \frac{6^2}{12} = 3 \quad \text{var}(U) = \frac{2^2}{12} = \frac{1}{3}$$

$$\text{var}(X) = 3 + \frac{1}{3} = \frac{10}{3}$$

$$\begin{aligned} \text{cov}(\Theta, \Theta + U) &= \\ &= \text{cov}(\Theta, \Theta) + \cancel{\text{cov}(\Theta, U)} = 3 \end{aligned}$$

LLMS for inferring the parameter of a coin

- Standard example:
 - coin with bias Θ ; prior $f_\Theta(\cdot)$
 - fix n ; X = number of heads
- Assume $f_\Theta(\cdot)$ is uniform in $[0, 1]$

$$\widehat{\Theta}_{\text{LMS}} = \frac{X + 1}{n + 2} = \widehat{\Theta}_{\text{LLMS}}$$

$$\widehat{\Theta}_{\text{LLMS}} = \mathbf{E}[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)} \left(X - \mathbf{E}[X] \right)$$

LLMS for inferring the parameter of a coin

- Θ : uniform on $[0, 1]$ $E[\Theta] = \frac{1}{2}$ $\text{var}(\Theta) = \frac{1}{12}$ $E[\Theta^2] = \frac{1}{12} + \frac{1}{2^2} = \frac{1}{3}$

- $p_{X|\Theta}$: $\text{Bin}(n, \Theta)$ $E[X | \Theta] = n\Theta$ $\text{var}(X | \Theta) = n\Theta(1 - \Theta)$

$$E[X] = E[n\Theta] = n/2 \quad E[X^2 | \Theta] = n\Theta(1 - \Theta) + n^2\Theta^2$$

- $E[X^2] = E[E[X^2 | \Theta]] = E[n\Theta + (n^2 - n)\Theta^2] = \frac{n}{2} + \frac{n^2 - n}{3} = \frac{n}{6} + \frac{n^2}{3}$

$$\text{var}(X) = E[X^2] - (E[X])^2 = \frac{n}{6} + \frac{n^2}{3} - \frac{n^2}{4} = \frac{n}{6} + \frac{n^2}{12} = \frac{n(n+2)}{12}$$

$$E[\Theta X | \Theta] = \Theta E[X | \Theta] = n\Theta^2$$

$$E[\Theta X] = E[E[\Theta X | \Theta]] = E[n\Theta^2] = n/3$$

$$\text{cov}(\Theta, X) = E[\Theta X] - E[\Theta]E[X] = \frac{n}{3} - \frac{n}{4} = \frac{n}{12}$$

LLMS for inferring the parameter of a coin

$$\widehat{\Theta}_{LLMS} = \mathbf{E}[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)} (X - \mathbf{E}[X])$$

$$\text{cov}(\Theta, X) = \frac{n}{12} \quad \text{var}(X) = \frac{n(n+2)}{12} \quad \mathbf{E}[X] = \frac{n}{2}$$

$$\widehat{\Theta}_{LLMS} = \frac{X+1}{n+2} = \widehat{\Theta}_{LMS}$$

LLMS with multiple observations

- Unknown Θ ; observations $X = (X_1, \dots, X_n)$
- Consider estimators of the form: $\widehat{\Theta} = a_1X_1 + \dots + a_nX_n + b$
- Find best choices of a_1, \dots, a_n, b
minimize: $E[(a_1X_1 + \dots + a_nX_n + b - \Theta)^2] = a_1^2 E[X^2] + 2a_1 a_2 E[X, X_2] + \dots + a_n^2 E[X, \Theta] + \dots$
- If $E[\Theta | X]$ is linear in X , then $\widehat{\Theta}_{\text{LMS}} = \widehat{\Theta}_{\text{LLMS}}$
- Solve linear system in b and the a_i •
- Only means, variances, covariances matter
- If multiple unknown Θ_j , apply to each one, separately

The simplest LLMS example with multiple observations

$$X_1 = \Theta + W_1 \quad \Theta \sim x_0, \sigma_0^2 \quad W_i \sim 0, \sigma_i^2$$

⋮

$$X_n = \Theta + W_n \quad \Theta, W_1, \dots, W_n \text{ uncorrelated}$$

- Suppose Θ, W_1, \dots, W_n are independent normal

$$\hat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta | X = x] = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$
$$\widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta | X] = \frac{\frac{x_0}{\sigma_0^2} + \sum_{i=1}^n \frac{X_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}} = \widehat{\Theta}_{\text{LLMS}}$$

- Suppose general (not normal) distributions, but same means, variances, as in normal example
 - all covariances also the same
 - solution must be the same

The representation of the data matters in LLMS

- Estimation based on X versus X^3
 - LMS: $\underline{E[\Theta | X]}$ is the same as $\underline{E[\Theta | X^3]}$
 - LLMS is different: estimator $\widehat{\Theta} = \underline{aX + b}$ versus $\widehat{\Theta} = \underline{aX^3 + b}$
 $\text{cov}(\Theta, X^3) \quad \text{var}(x^3)$
 - can also consider $\widehat{\Theta} = \underline{a_1}\widehat{X} + \underline{a_2}\widehat{X^2} + \underline{a_3}\widehat{X^3} + b$
 - can also consider $\widehat{\Theta} = \underline{a_1}X + \underline{a_2}e^X + \underline{a_3}\log X + b$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 18: Inequalities, convergence, and the Weak Law of Large Numbers

- Inequalities
 - bound $\mathbf{P}(X \geq a)$ based on limited information about a distribution
 - Markov inequality (based on the mean)
 - Chebyshev inequality (based on the mean and variance)
- WLLN: X, X_1, \dots, X_n i.i.d.

$$\frac{X_1 + \dots + X_n}{n} \longrightarrow \mathbf{E}[X]$$

- application to polling
- Precise defn. of convergence
 - convergence “in probability”

The Markov inequality

- Use a bit of information about a distribution to learn something about probabilities of “extreme events”
- “If $X \geq 0$ and $E[X]$ is small, then X is unlikely to be very large”

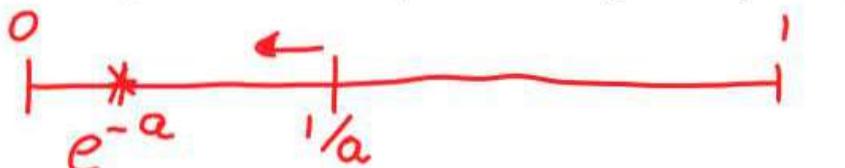
Markov inequality: If $X \geq 0$ and $a > 0$, then $P(X \geq a) \leq \frac{E[X]}{a}$.

$$Y = \begin{cases} 0, & \text{if } X < a \\ a, & \text{if } X \geq a \end{cases} \quad \text{and } P(X \geq a) = E[Y] \leq E[X]$$

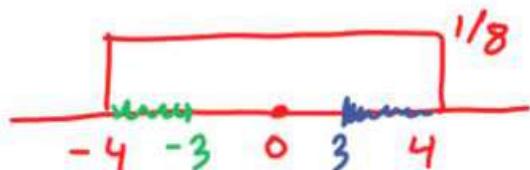
The Markov inequality

Markov inequality: If $X \geq 0$ and $a > 0$, then $P(X \geq a) \leq \frac{E[X]}{a}$

- **Example:** X is Exponential($\lambda = 1$): $P(X \geq a) \leq \frac{1}{a}$

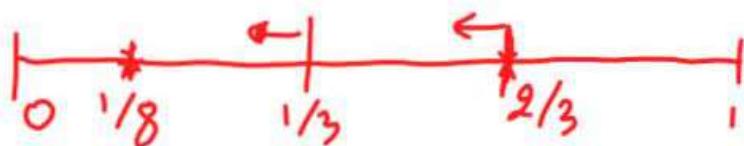


- **Example:** X is Uniform[-4, 4]: $P(|X| \geq 3) \leq \frac{E[|X|]}{3} = \frac{2}{3}$



$$P(|X| \geq 3) \leq \frac{E[|X|]}{3} = \frac{2}{3}$$

$= \frac{1}{2} P(|X| \geq 3) \leq \frac{1}{3}$



The Chebyshev inequality

- Random variable X , with finite mean μ and variance σ^2
- “If the variance is small, then X is unlikely to be too far from the mean”

Chebyshev inequality: $P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$

Markov inequality: If $X \geq 0$ and $a > 0$, then $P(X \geq a) \leq \frac{E[X]}{a}$

$$P(|x-\mu| \geq c) = P(\underbrace{(x-\mu)^2}_{\geq c^2} \geq c^2) \leq \frac{E[(x-\mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$$

The Chebyshev inequality

Chebyshev inequality: $P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$

$$P(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2} \quad k=3 \quad \leq \frac{1}{9}$$

- Example: X is Exponential($\lambda = 1$): $P(X \geq a) \leq \frac{1}{a}$ (Markov)



$$P(X \geq a) = P(X-1 \geq a-1) \leq P(|X-1| \geq a-1) \leq \frac{1}{(a-1)^2} \sim \frac{1}{a^2}$$

The Weak Law of Large Numbers (WLLN)

- X_1, X_2, \dots i.i.d.; finite mean μ and variance σ^2

Sample mean: $M_n = \frac{X_1 + \dots + X_n}{n}$ $\mu = E[X_i]$

- $E[M_n] = \frac{E[X_1 + \dots + X_n]}{n} = \frac{n\mu}{n} = \mu$

- $\text{Var}(M_n) = \frac{\text{Var}(X_1 + \dots + X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

$$P(|M_n - \mu| \geq \epsilon) \leq \frac{\text{Var}(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \xrightarrow[n \rightarrow \infty]{} 0 \quad (\text{fixed } \epsilon > 0)$$

WLLN: For $\epsilon > 0$, $P(|M_n - \mu| \geq \epsilon) = P\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| \geq \epsilon\right) \rightarrow 0$, as $n \rightarrow \infty$

Interpreting the WLLN

$$M_n = (X_1 + \dots + X_n)/n$$

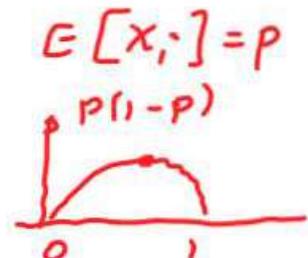
WLLN: For $\epsilon > 0$, $P(|M_n - \mu| \geq \epsilon) = P\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| \geq \epsilon\right) \rightarrow 0, \text{ as } n \rightarrow \infty$

- One experiment
 - many measurements $X_i = \mu + W_i$
 - W_i : measurement noise; $E[W_i] = 0$; independent W_i
 - **sample mean** M_n is unlikely to be far off from **true mean** μ
- Many independent repetitions of the same experiment
 - event A , with $p = P(A)$
 - X_i : indicator of event A
 - the sample mean M_n is the **empirical frequency** of event A

$$X_i = 1, \text{ if } A \text{ occurs} \\ 0, \text{ o.w.}$$
$$E[X_i] = p$$

The pollster's problem

- p : fraction of population that will vote "yes" in a referendum
- i th (randomly selected) person polled: $X_i = \begin{cases} 1, & \text{if yes,} \\ 0, & \text{if no.} \end{cases}$
- $M_n = (X_1 + \dots + X_n)/n$: fraction of "yes" in our sample
- Would like "small error," e.g.: $|M_n - p| < 0.01$
- Try $n = 10,000$
- $P(|M_{10,000} - p| \geq 0.01) \leq \frac{\sigma^2}{n \varepsilon^2} = \frac{p(1-p)}{10^4 \cdot 10^{-4}} \leq \frac{1}{4}$ ← want $\leq 5\%$
- $$\frac{1/4}{n \cdot 10^{-4}} \leq \frac{5}{10^2} \Leftrightarrow n \geq \frac{10^6}{20} = 50,000$$
 ← will suffice



Convergence “in probability”

WLLN: For any $\epsilon > 0$, $\mathbf{P}(|M_n - \mu| \geq \epsilon) \rightarrow 0$, as $n \rightarrow \infty$

$$M_n \xrightarrow[n \rightarrow \infty]{i.p} \mu$$

- Would like to say that “ M_n converges to μ ”
- Need to define the word “converges”
- Sequence of random variables Y_n ; not necessarily independent

Definition: A sequence Y_n converges in probability to a number a if:

for any $\epsilon > 0$, $\lim_{n \rightarrow \infty} \mathbf{P}(|Y_n - a| \geq \epsilon) = 0$

Understanding convergence “in probability”

- Ordinary convergence
 - Sequence a_n ; number a

$a_n \rightarrow a$

“ a_n eventually gets and stays (arbitrarily) close to a ”
 - Convergence in probability
 - Sequence Y_n ; number a

$Y_n \rightarrow a$

 - for any $\epsilon > 0$, $\mathbf{P}(|Y_n - a| \geq \epsilon) \rightarrow 0$
- For every $\epsilon > 0$, there exists n_0 , such that for every $n \geq n_0$, we have $|a_n - a| \leq \epsilon$
- “(almost all) of the PMF/PDF of Y_n eventually gets concentrated (arbitrarily) close to a ”

Some properties

- Suppose that $X_n \rightarrow a$, $Y_n \rightarrow b$, in probability

- If g is continuous, then $g(X_n) \rightarrow g(a)$

$$X_n^2 \rightarrow a^2$$

- $X_n + Y_n \rightarrow a + b$

- **But:** $E[X_n]$ need not converge to a

Convergence in probability examples



$$Y_n \xrightarrow[n \rightarrow \infty]{i.p.} 0.$$

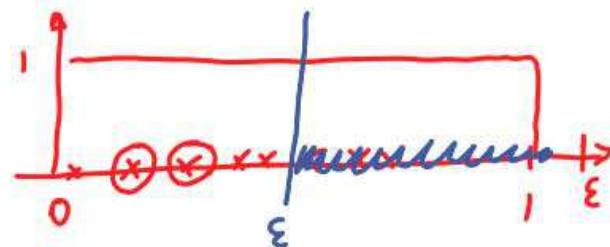
$$\varepsilon > 0 \quad P(|Y_n - 0| \geq \varepsilon) = 1/n \xrightarrow[n \rightarrow \infty]{} 0$$

$$E[Y_n] = n^2 \cdot \frac{1}{n} = n \xrightarrow[n \rightarrow \infty]{} \infty$$

- convergence in probability does **not** imply convergence of expectations

Convergence in probability examples

- X_i : i.i.d., uniform on $[0, 1]$
- $Y_n = \min\{X_1, \dots, X_n\}$



$$Y_{n+1} \leq Y_n$$

$$P(|Y_n - 0| \geq \epsilon) = P(Y_n \geq \epsilon).$$

$$\epsilon > 0$$

$$= P(X_1 \geq \epsilon, \dots, X_n \geq \epsilon)$$

$$Y_n \xrightarrow[n \rightarrow \infty]{i.p.} 0$$

$$\epsilon > 1$$

$$= P(X_1 \geq \epsilon) \cdots P(X_n \geq \epsilon)$$

$$\epsilon \leq 1$$

$$= (1 - \epsilon)^n \xrightarrow{n \rightarrow \infty} 0$$

Related topics

- Better bounds/approximations on tail probabilities

- Markov and Chebyshev inequalities

- Chernoff bound

$$P(|M_n - \mu| \geq a) \leq e^{-n \frac{\lambda(a)}{a^2}}$$

- Central limit theorem " $M_n \sim N(\mu, \sigma^2/n)$ "

- Different types of convergence

- Convergence in probability

- Convergence "with probability 1"

$$P\left(\left\{\omega : Y_n(\omega) \xrightarrow[n \rightarrow \infty]{} Y(\omega)\right\}\right) = 1$$

- Strong law of large numbers $M_n \xrightarrow[n \rightarrow \infty]{\text{wpt}} \mu$

- Convergence of a sequence of distributions (CDFs) to a limiting CDF

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability

John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 19: The Central Limit Theorem (CLT)

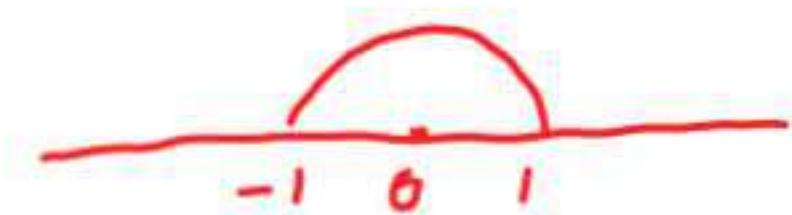
- WLLN: $\frac{X_1 + \cdots + X_n}{n} \rightarrow E[X]$

- CLT: $X_1 + \cdots + X_n \approx \text{normal}$

- precise statement
- universality, usefulness
- many examples
- refinement for discrete r.v.s
- application to polling

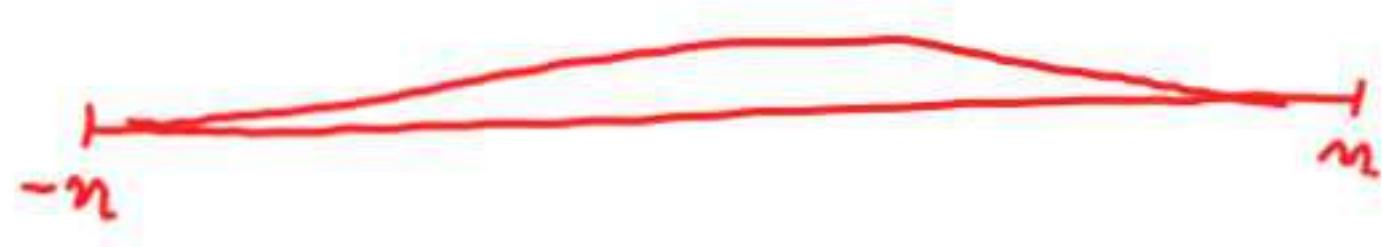
Different scalings of the sum of i.i.d. random variables

- X_1, \dots, X_n i.i.d., finite mean μ and variance σ^2



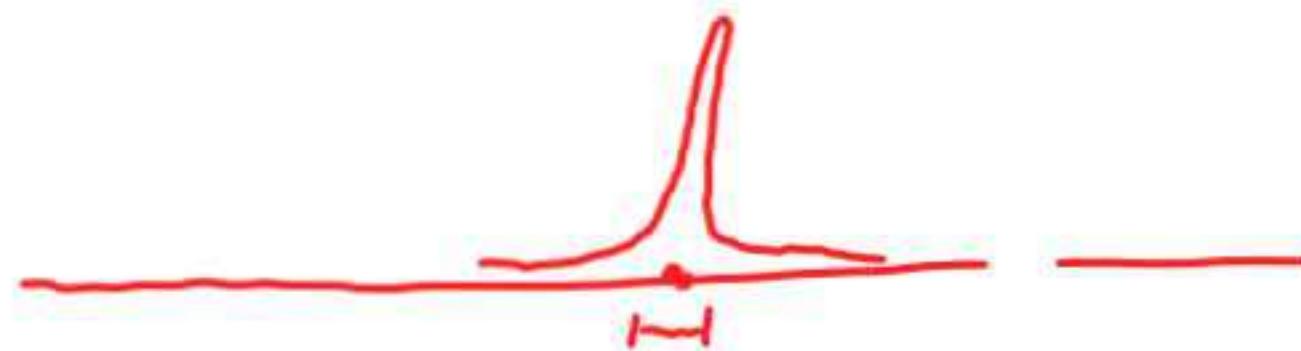
- $S_n = X_1 + \dots + X_n$

variance: $n\sigma^2$



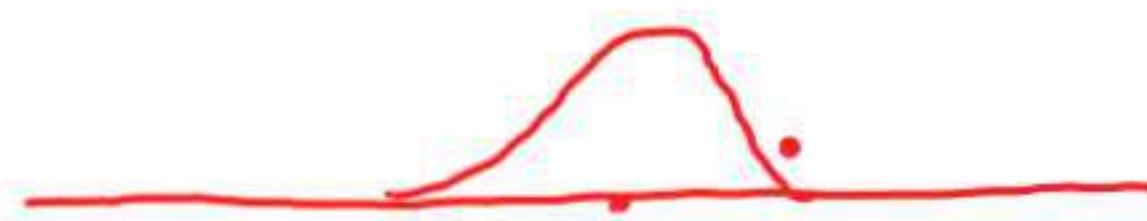
- $M_n = \frac{S_n}{n} = \frac{X_1 + \dots + X_n}{n}$

variance: $\frac{\sigma^2}{n} \xrightarrow{n \rightarrow \infty} 0$



- $\frac{S_n}{\sqrt{n}} = \frac{X_1 + \dots + X_n}{\sqrt{n}}$

variance: $\sigma^2 = \frac{n\sigma^2}{n}$



The Central Limit Theorem (CLT)

- X_1, \dots, X_n i.i.d., finite mean μ and variance σ^2
- $S_n = X_1 + \dots + X_n$ variance: $n\sigma^2$
- $\frac{S_n}{\sqrt{n}} = \frac{X_1 + \dots + X_n}{\sqrt{n}}$ variance: σ^2

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

$$\mathbf{E}[Z_n] = 0$$

$$\text{var}(Z_n) = 1$$

- Let Z be a standard normal r.v. (zero mean, unit variance)

Central Limit Theorem: For every z : $\lim_{n \rightarrow \infty} \mathbf{P}(Z_n \leq z) = \mathbf{P}(Z \leq z)$

- $\mathbf{P}(Z \leq z)$ is the standard normal CDF, $\Phi(z)$, available from the normal tables

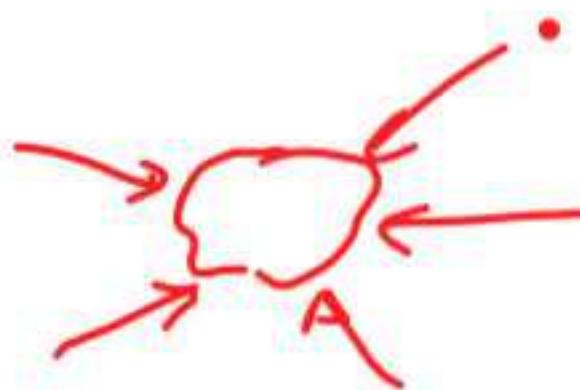
Usefulness of the CLT

$$S_n = X_1 + \dots + X_n$$

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} \quad Z \sim N(0, 1)$$

Central Limit Theorem: For every z : $\lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z)$

- universal and easy to apply; only means, variances matter
- fairly accurate computational shortcut
- justification of normal models



What exactly does the CLT say? — Theory

$$S_n = X_1 + \dots + X_n \quad Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} \quad Z \sim N(0, 1)$$

Central Limit Theorem: For every z : $\lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z)$

- CDF of Z_n converges to normal CDF
- results for convergence of PDFs or PMFs (with more assumptions)
- results without assuming that the X_i are identically distributed
- results under “weak dependence”
- proof: uses “transforms”: $E[e^{sZ_n}] \rightarrow E[e^{sZ_\bullet}]$, for all s

What exactly does the CLT say? — Practice

$$S_n = X_1 + \dots + X_n$$

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} \quad Z \sim N(0, 1)$$

Central Limit Theorem: For every z : $\lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z)$

- The **practice** of normal approximations:

- treat Z_n as if it were normal

$$S_n = \sqrt{n}\sigma Z_n + n\mu$$

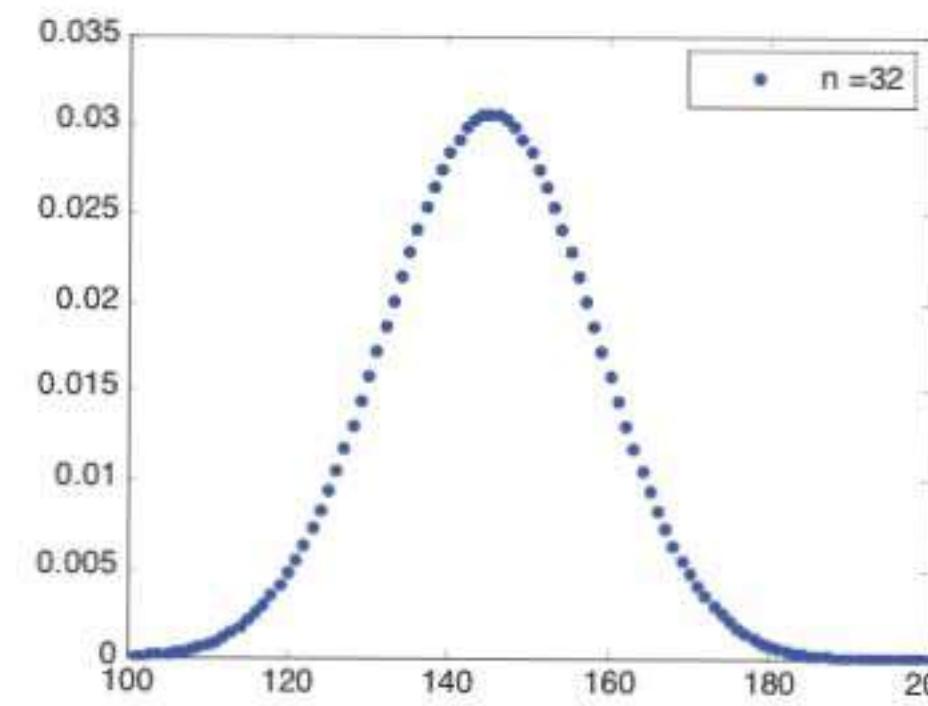
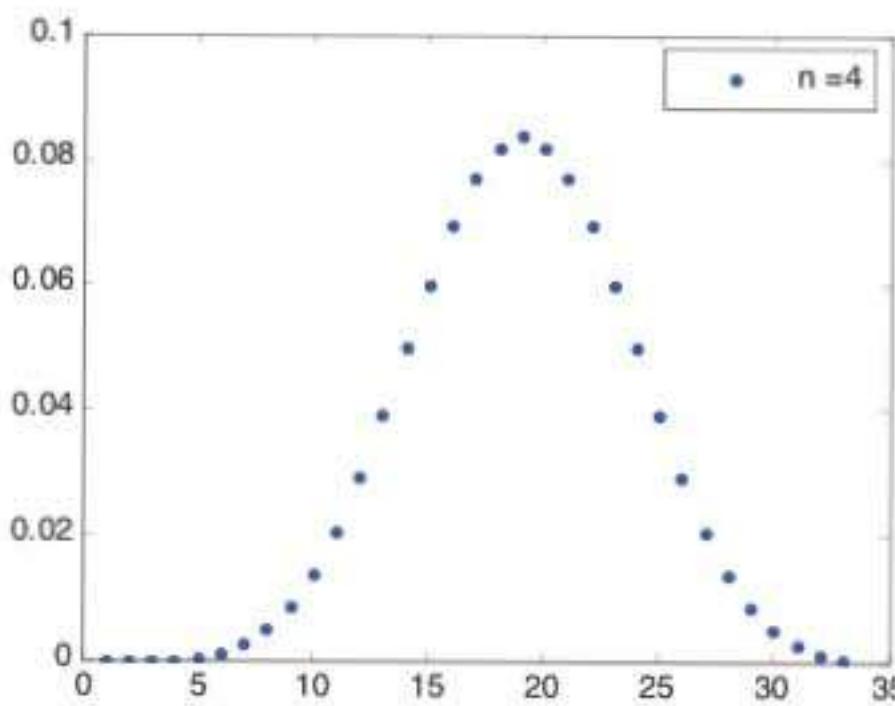
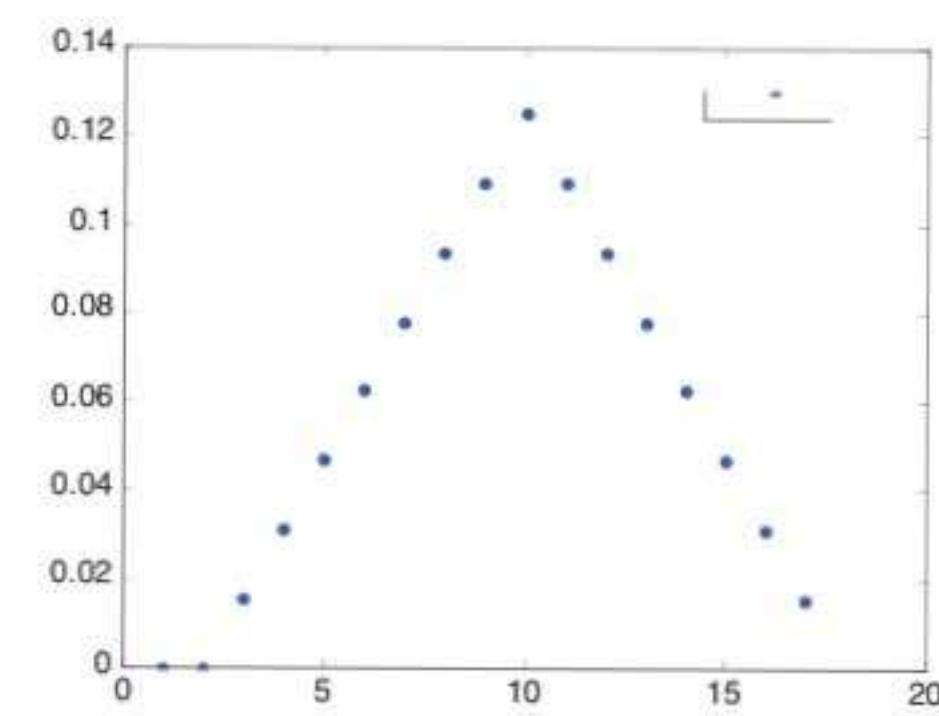
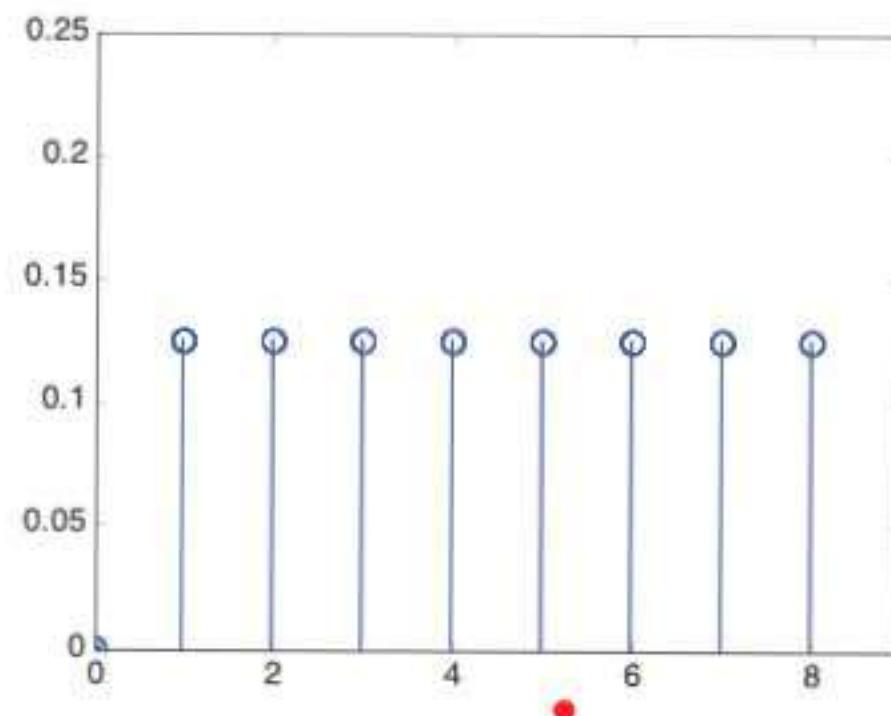
- hence treat S_n as if normal: $N(n\mu, n\sigma^2)$

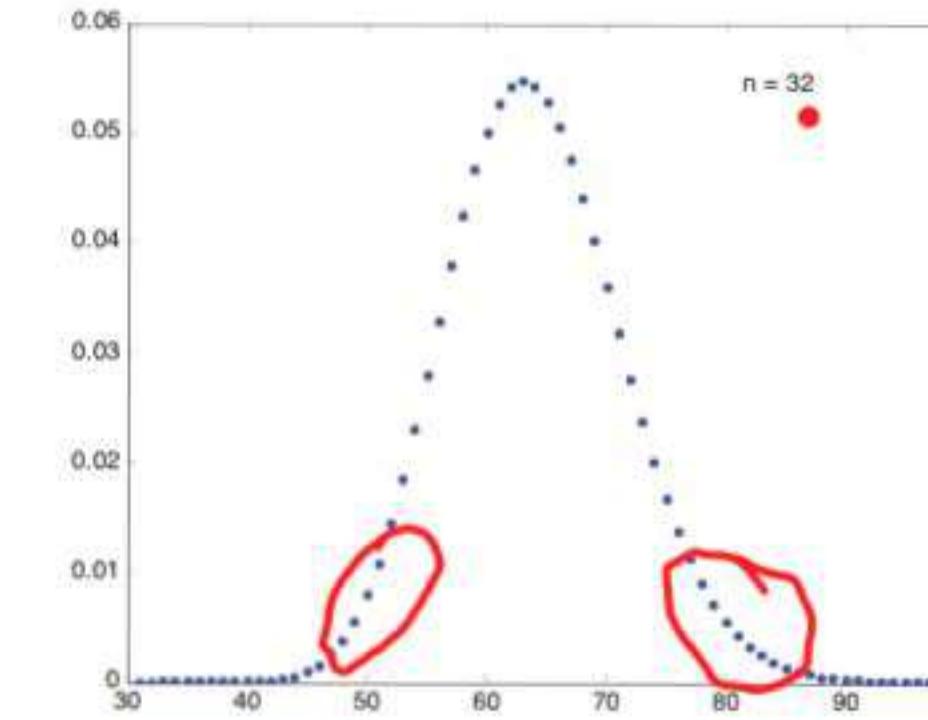
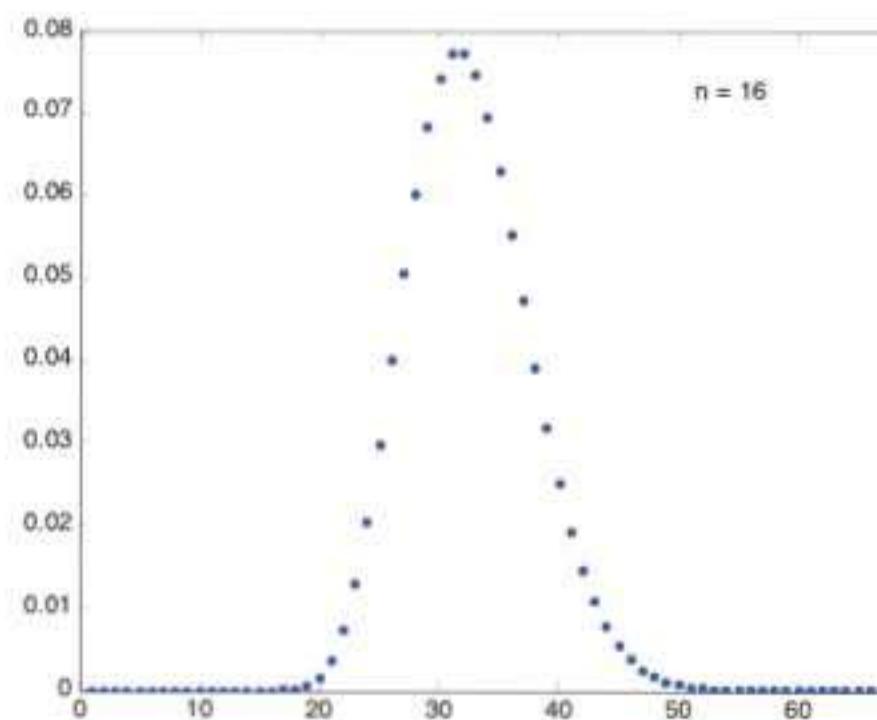
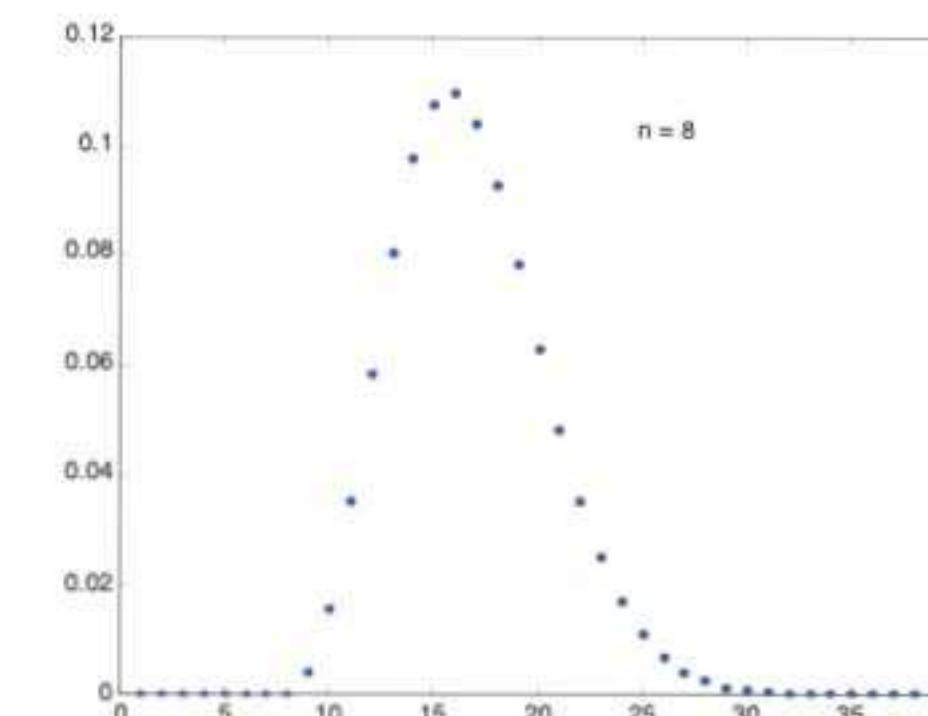
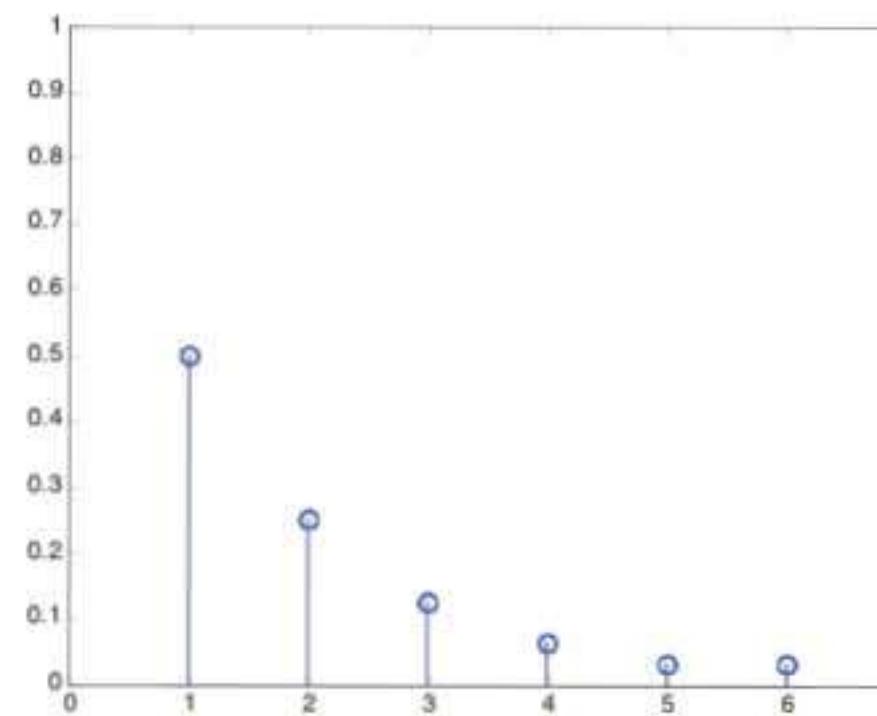
- Can we use the CLT when n is “moderate”?

$n = 30$?

- usually, yes

- symmetry and unimodality help





Example 1

- $P(S_n \leq a) \approx b$ given two parameters, find the third
- Package weights X_i , i.i.d. exponential, $\lambda = 1/2$;
- Load container with $n = 100$ packages

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

$$\mu = \sigma = 2$$

$$P(S_n \geq 210)$$

$$= P\left(\frac{S_n - 200}{20} > \frac{210 - 200}{20}\right)$$

$$= P(Z_n > 0.5) \approx P(Z > 0.5)$$

$$= 1 - P(Z < 0.5) = 1 - \Phi(0.5)$$

$$= 1 - 0.6915 = 0.3085$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	0.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Example 2

- $P(S_n \leq a) \approx b$ given two parameters, find the third
- Package weights X_i , i.i.d. exponential, $\lambda = 1/2$;
- Let $n = 100$. Choose the “capacity” a , so that $P(S_n \geq a) \approx 0.05$.

$$0.05 \approx P\left(\frac{S_n - 200}{20} > \frac{a - 200}{20}\right)$$

$$\approx 1 - \Phi\left(\frac{a - 200}{20}\right)$$

0.95

$$\frac{a - 200}{20} = 1.645 \quad a = 232.9$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Example 3

- $P(S_n \leq a) \approx b$ given two parameters, find the third
- Package weights X_i , i.i.d. exponential, $\lambda = 1/2$;
- How large can n be,
so that $P(S_n \geq 210) \approx 0.05$?

$$P\left(\frac{S_n - n\mu}{\sqrt{n}} \geq \frac{210 - n\mu}{\sqrt{n}}\right)$$

$$\approx 1 - \Phi\left(\frac{210 - n\mu}{\sqrt{n}}\right) \approx 0.05$$

0.95

$$\frac{210 - n\mu}{\sqrt{n}} = 1.645$$

•
 $n = 89$

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

$$\mu = \sigma = 2$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Example 4

- $P(S_n \leq a) \approx b$ given two parameters, find the third

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

$$\mu = \sigma = 2$$

- Package weights X_i , i.i.d. exponential, $\lambda = 1/2$;

- Load container until weight exceeds 210

N : number of packages loaded

- $P(N > 100)$

$$= P\left(\sum_{i=1}^{100} X_i \leq 210\right)$$

$$\approx \Phi\left(\frac{210 - 200}{20}\right) = \Phi(0.5)$$

$$= 0.6915$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Normal approximation to the binomial

- X_i : independent, Bernoulli(p); $0 < p < 1$
- $S_n = X_1 + \dots + X_n$: Binomial(n, p)
 - mean np , variance $np(1 - p)$
- $n = 36, p = 0.5$; find $P(S_n \leq 21)$

$$np = 18 \quad \sqrt{np(1 - p)} = 3$$

$$P\left(\frac{S_n - 18}{3} \leq \frac{21 - 18}{3}\right)$$

$$= P(Z_n \leq 1) \approx \Phi(1) = .8413$$

- CDF of $\frac{S_n - np}{\sqrt{np(1 - p)}}$ → standard normal

$$\sum_{k=0}^{21} \binom{36}{k} \left(\frac{1}{2}\right)^{36} = 0.8785$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

The $1/2$ correction for integer random variables

- $0.8413 \approx P(S_n \leq 21) = P(S_n < 22)$, because S_n is integer

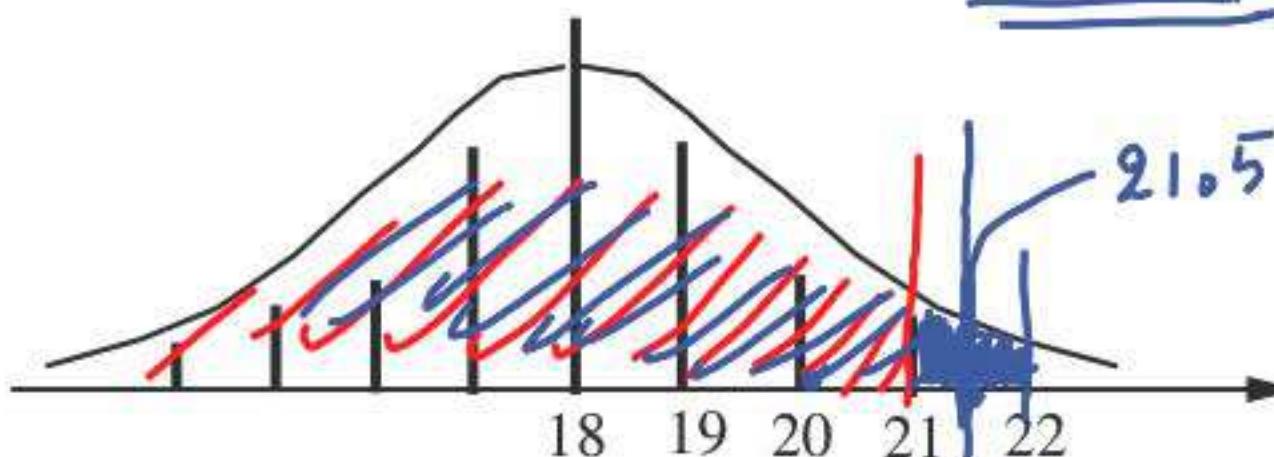
$$= P\left(\frac{S_n - 18}{3} < \frac{22 - 18}{3}\right)$$

$$= P(Z_n < 1.33) \approx \Phi(1.33) = 0.9082$$

true value 0.8785

$$P(S_n \leq 21.5) = P\left(Z_n \leq \frac{21.5 - 18}{3}\right)$$

$$\approx \Phi(1.17) = 0.8790$$



	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

De Moivre–Laplace CLT to the binomial

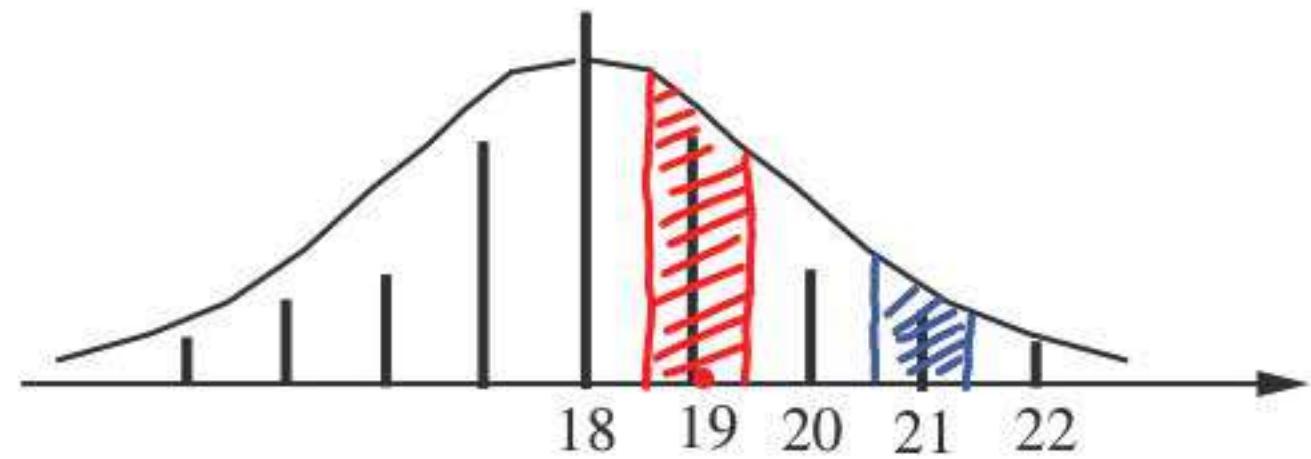
$$P(S_n = 19) = P(18.5 \leq S_n \leq 19.5)$$

$$= P\left(\frac{18.5 - 18}{3} \leq Z_n \leq \frac{19.5 - 18}{3}\right)$$

$$= P(0.17 \leq Z_n \leq 0.5)$$

$$\approx \Phi(0.5) - \Phi(0.17)$$

$$= 0.6915 - 0.5675 = 0.124$$



- Exact answer:

$$\binom{36}{19} \left(\frac{1}{2}\right)^{36} = 0.1251$$

- When the 1/2 correction is used, the CLT can also approximate the binomial PMF (not just the binomial CDF)

The pollster's problem revisited

- p : fraction of population that will vote "yes" in a referendum
- i th (randomly selected) person polled: $X_i = \begin{cases} 1, & \text{if yes,} \\ 0, & \text{if no.} \end{cases}$
- $M_n = (X_1 + \dots + X_n)/n$: fraction of "yes" in our sample
- Would like "small error," e.g.: $|M_n - p| < 0.01$

$$E[X_i] = p = \mu$$

$$\sigma = \sqrt{p(1-p)}$$

$$P(|M_n - p| \geq 0.01) = P\left(|Z_n| \geq \frac{0.01\sqrt{n}}{\sigma}\right) \approx P\left(|Z| \geq \frac{0.01\sqrt{n}}{\sigma}\right)$$

$\overset{N(0,1)}{\overbrace{Z_n}}$

$$Z_n = \frac{S_n - np}{\sqrt{n}\sigma}$$

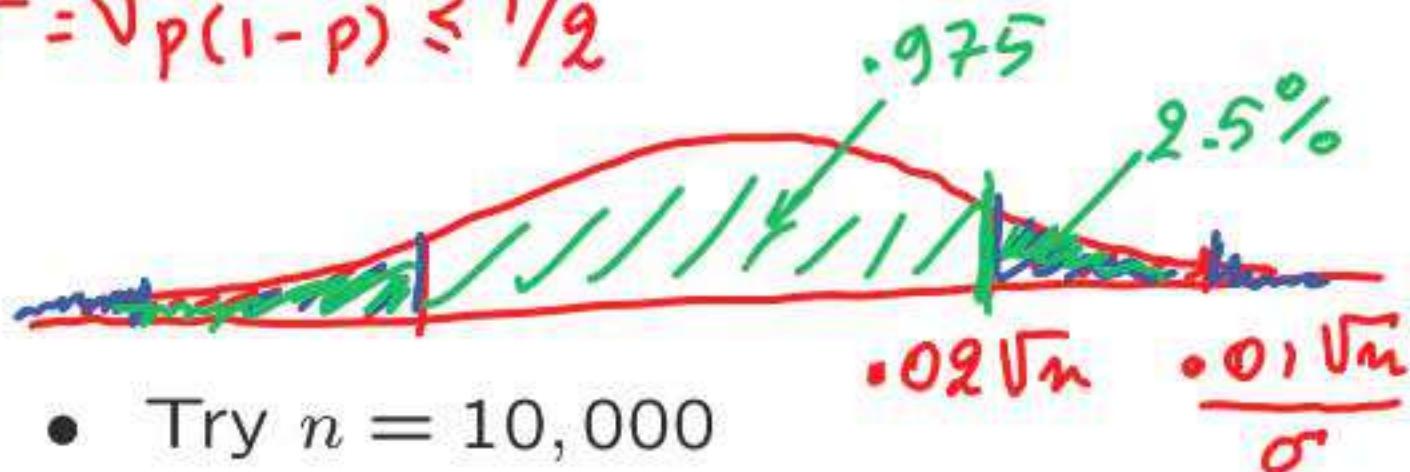
$$\left| \frac{S_n - np}{n} \right| \geq 0.01$$

$$\left| \frac{S_n - np}{\sqrt{n}\sigma} \right| \geq \frac{0.01\sqrt{n}}{\sigma}$$

The pollster's problem revisited

$$P(|M_n - p| \geq .01) \approx P\left(|Z| \geq \frac{.01\sqrt{n}}{\sigma}\right) \leq P(|Z| \geq .02\sqrt{n}) = 2(1 - \Phi(.02\sqrt{n})) = 0.05$$

$$\sigma = \sqrt{p(1-p)} \leq 1/2$$



- Try $n = 10,000$

$$\text{prob} \leq 2(1 - \Phi(2)) =$$

$$= 2(1 - 0.9772) = 0.046$$

- Specs: $P(|M_n - p| \geq .01) \leq .05$

$$\Phi(.02\sqrt{n}) = 0.975$$

$$.02\sqrt{n} = 1.96 \Rightarrow n = 9604$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 20: An introduction to classical statistics

- Unknown constant θ (not a r.v.)
- if $\theta = \mathbf{E}[X]$: estimate using the sample mean $(X_1 + \dots + X_n)/n$
 - terminology and properties
- Confidence intervals (CIs)
 - CIs using the CLT
 - CIs when the variance is unknown
- Other uses of sample means
- Maximum Likelihood estimation

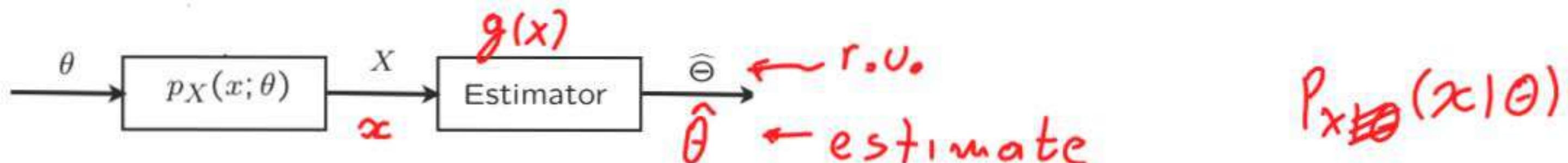
Classical statistics

- Inference using the Bayes rule:

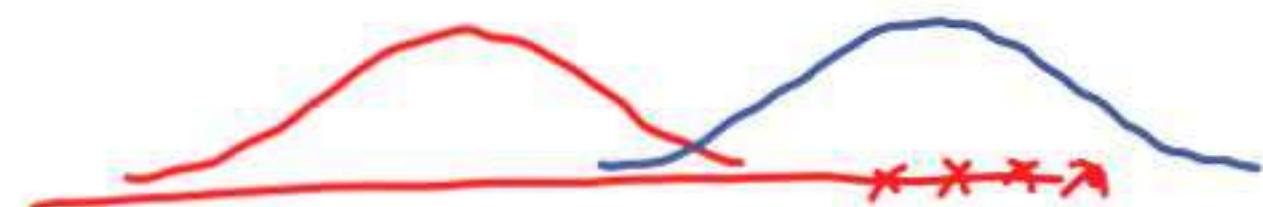
- unknown Θ and observation X are both random variables

- Find $p_{\Theta|X}$

- Classical statistics: unknown constant θ

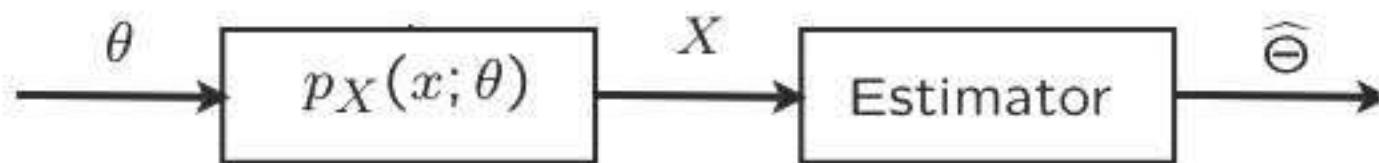


- also for vectors X and θ : $p_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta_1, \dots, \theta_m)$
- $p_X(x; \theta)$ are NOT conditional probabilities; θ is NOT random
- mathematically: many models, one for each possible value of θ



Problem types in classical statistics

- Classical statistics: unknown constant θ



- Hypothesis testing: $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$
- Composite hypotheses: $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$
- Estimation: design an **estimator** $\widehat{\Theta}$, to “keep estimation **error** $\widehat{\Theta} - \theta$ small”



Art! •

Estimating a mean

- X_1, \dots, X_n : i.i.d., mean θ , variance σ^2

$$\widehat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \dots + X_n}{n}$$

$\widehat{\Theta}_n$: estimator (a random variable)

Properties and terminology:

- $E[\widehat{\Theta}_n] = \theta$ (unbiased)
for all θ
- WLLN: $\widehat{\Theta}_n \xrightarrow{i.p.} \theta$ (consistency)
for all θ
- mean squared error (MSE): $E[(\widehat{\Theta}_n - \theta)^2] = \text{var}(\widehat{\Theta}_n) = \frac{\sigma^2}{n}$

$$\begin{aligned}\widehat{\Theta} &= g(x) \\ E[\widehat{\Theta}] &= \sum_x g(x) P_x(x; \theta)\end{aligned}$$

On the mean squared error of an estimator

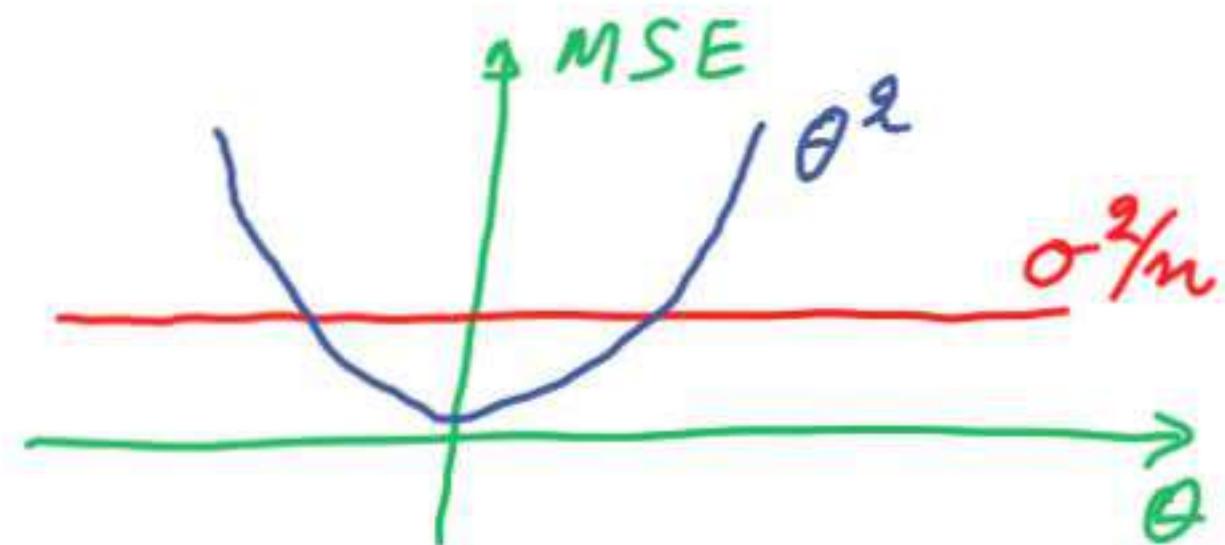
- For any estimator, using $E[Z^2] = \text{var}(Z) + (E[Z])^2$:

$$Z = \hat{\Theta} - \theta$$

$$E[(\hat{\Theta} - \theta)^2] = \text{var}(\hat{\Theta} - \theta) + (E[\hat{\Theta} - \theta])^2 = \text{var}(\hat{\Theta}) + (\text{bias})^2$$

$$\hat{\Theta}_n = M_n : \text{MSE} = \sigma^2/n + 0$$

$$\hat{\Theta} = 0 : \text{MSE} = 0 + \theta^2$$



- $\sqrt{\text{var}(\hat{\Theta})}$ is called the standard error

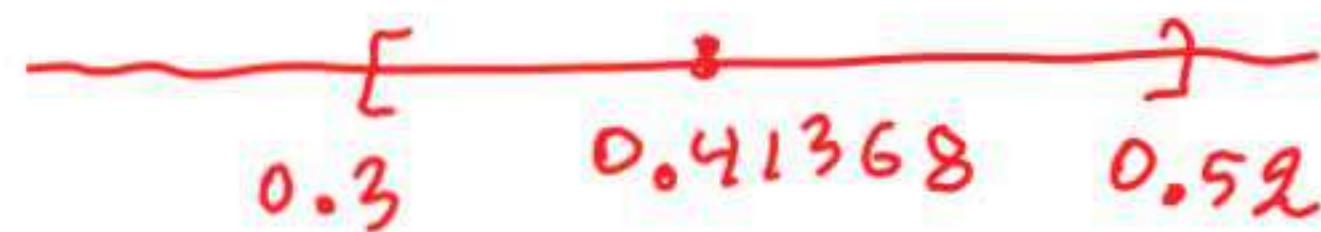
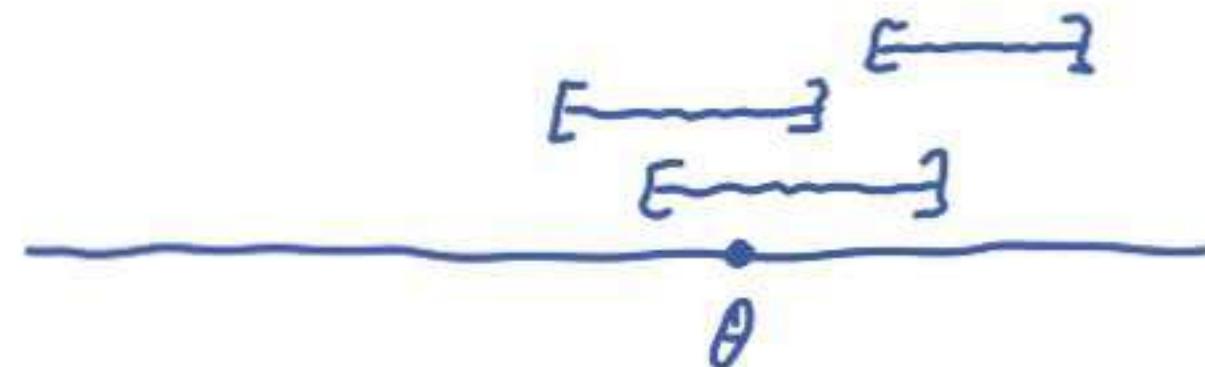


Confidence intervals (CIs)

- The value of an estimator $\widehat{\Theta}$ may not be informative enough
95%
- An $1 - \alpha$ **confidence interval** is an interval $[\widehat{\Theta}^-, \widehat{\Theta}^+]$,

s.t. $P(\widehat{\Theta}^- \leq \theta \leq \widehat{\Theta}^+) \geq 1 - \alpha$, for all θ

- often $\alpha = 0.05$, or 0.025 , or 0.01
- interpretation is subtle



$$P(0.3 < \theta < 0.52) \geq .95$$

CI for the estimation of the mean

$$\widehat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

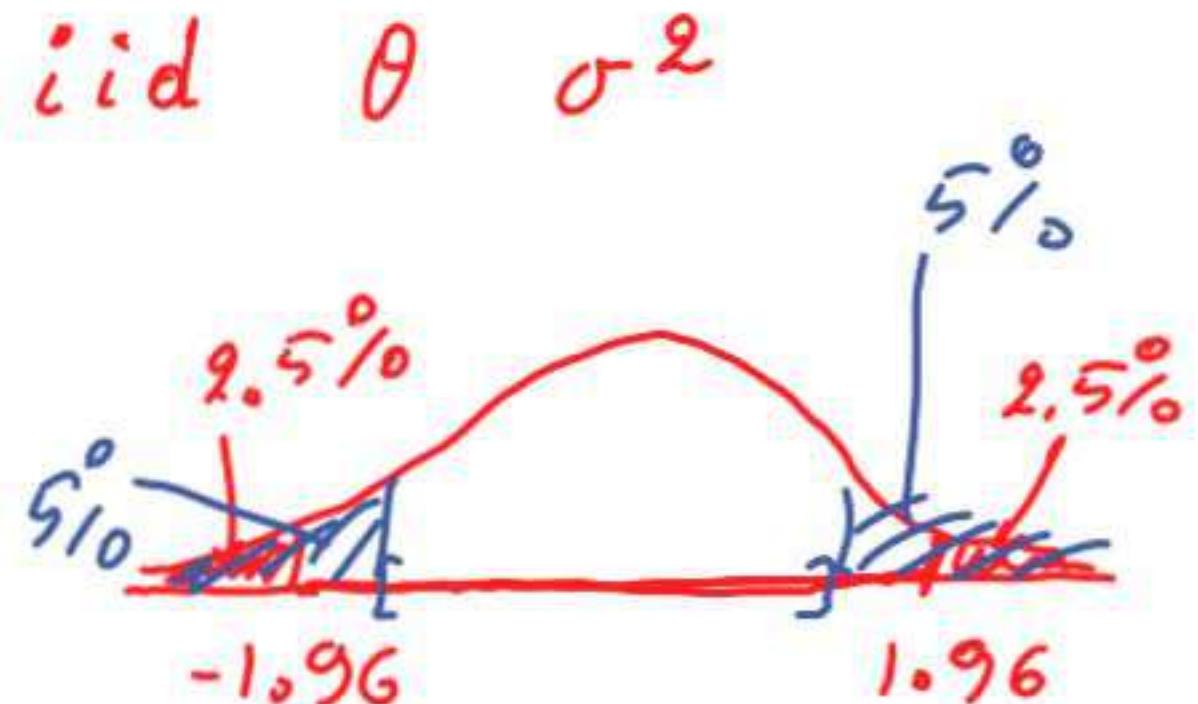
95%

normal tables: $\Phi(1.96) = 0.975 = 1 - 0.025$

90%

$$\Phi(1.645) = 0.90$$

$$P\left(\frac{|\widehat{\Theta}_n - \theta|}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95 \quad (\text{CLT})$$



$$P\left(\widehat{\Theta}_n - \frac{1.96 \sigma}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{1.96 \sigma}{\sqrt{n}}\right) \approx 0.95$$

$$\widehat{\Theta}^-$$

$$\widehat{\Theta}^+$$

Confidence intervals for the mean when σ is unknown

$$\widehat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

$$P\left(\widehat{\Theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{1.96\sigma}{\sqrt{n}}\right) \approx 0.95$$

- **Option 1:** use upper bound on σ
 - if X_i Bernoulli: $\sigma \leq 1/2$
- **Option 2:** use ad hoc estimate of σ
 - if X_i Bernoulli: $\hat{\sigma} = \sqrt{\widehat{\Theta}_n(1 - \widehat{\Theta}_n)}$

$$\sigma = \sqrt{\theta(1-\theta)}$$

Confidence intervals for the mean when σ is unknown

$$P\left(\widehat{\Theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{1.96\sigma}{\sqrt{n}}\right) \approx 0.95$$

Start from $\sigma^2 = E[(X_i - \theta)^2]$

- **Option 3:** Use sample mean estimate of the variance

$$\frac{1}{n} \sum_{i=1}^n (X_i - \theta)^2 \rightarrow \sigma^2$$

- Two approximations involved here:
 - CLT: approximately normal
 - using estimate of σ
- correction for second approximation (t -tables)
used when n is small

(but do not know θ)

$$\frac{1}{n-1} \sum_{i=1}^n (X_i - \widehat{\Theta}_n)^2 \rightarrow \sigma^2$$

Other natural estimators

- $\theta_X = \mathbb{E}[X]$ $\widehat{\Theta}_X = \frac{1}{n} \sum_{i=1}^n X_i$ • $\theta = \mathbb{E}[g(X)]$ $\widehat{\Theta} = \frac{1}{n} \sum_{i=1}^n g(X_i)$
- $v_X = \text{var}(X) = \mathbb{E}[(X - \theta_X)^2]$ $\widehat{v}_X = \frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\Theta}_X)^2$
- $\text{cov}(X, Y) = \mathbb{E}[(X - \theta_X)(Y - \theta_Y)]$
 (x_i, y_i) $\widehat{\text{cov}}(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\Theta}_X)(Y_i - \widehat{\Theta}_Y)$
- $\rho = \frac{\text{cov}(X, Y)}{\sqrt{v_X} \cdot \sqrt{v_Y}}$ $\widehat{\rho} = \frac{\widehat{\text{cov}}(X, Y)}{\sqrt{\widehat{v}_X} \cdot \sqrt{\widehat{v}_Y}}$
- next steps: find the distribution of $\widehat{\Theta}$, MSE, confidence intervals,...

Maximum Likelihood (ML) estimation

- Pick θ that “makes data most likely”

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} p_X(x; \theta)$$

– also applies when x, θ are vectors or x is continuous

- compare to Bayesian posterior: $p_{\Theta|X}(\theta | x) = \frac{p_{X|\Theta}(x | \theta) p_{\Theta}(\theta)}{p_X(x)}$ *constant*
- interpretation is very different

Comments on ML

- maximize $p_X(x; \theta)$
- maximization is usually done numerically
- if have n i.i.d. data drawn from model $p_X(x; \theta)$, then, under mild assumptions:
 - consistent: $\widehat{\Theta}_n \rightarrow \theta$
 - asymptotically normal: $\frac{\widehat{\Theta}_n - \theta}{\sigma(\widehat{\Theta}_n)} \rightarrow N(0, 1)$ (CDF convergence)
- analytical and simulation methods for calculating $\widehat{\sigma} \approx \sigma(\widehat{\Theta}_n)$
 - hence confidence intervals $P(\widehat{\Theta}_n - 1.96 \widehat{\sigma} \leq \theta \leq \widehat{\Theta}_n + 1.96 \widehat{\sigma}) \approx 0.95$
 - asymptotically “efficient” (“best”)

•

ML estimation example: parameter of binomial

- K : binomial with parameters n (known), and θ (unknown)

k

$$p_K(k; \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

$$\log \left[\binom{n}{k} \right] + k \log \theta + (n - k) \log (1 - \theta)$$

$$0 + \frac{k}{\theta} - \frac{n - k}{1 - \theta} = 0 \Rightarrow k - k\theta = n\theta - k\theta$$

$$\hat{\theta}_{\text{ML}} = \frac{k}{n} \quad \widehat{\Theta}_{\text{ML}} = \frac{K}{n}$$

- same as MAP estimator with uniform prior on θ

ML estimation example — normal mean and variance

- X_1, \dots, X_n : i.i.d., $N(\mu, v)$ $f_X(x; \mu, v) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi v}} \exp\left\{-\frac{(x_i - \mu)^2}{2v}\right\}$

minimize $\frac{n}{2} \log v + \sum_{i=1}^n \frac{(x_i - \mu)^2}{2v}$

– minimize w.r.t. μ : $\hat{\mu} = \frac{x_1 + \dots + x_n}{n}$

$$\frac{1}{v} \sum_{i=1}^n (x_i - \mu) = 0 \Rightarrow \sum x_i = n\mu$$

– minimize w.r.t. v : $\hat{v} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$

$$\cancel{\frac{n}{2} \cdot \frac{1}{v}} \rightarrow \sum_{i=1}^n \frac{(x_i - \mu)^2}{\cancel{2v^2}} = 0$$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 21: The Bernoulli process

- Definition of Bernoulli process
- Stochastic processes
- Basic properties (memorylessness)
- The time of the k th success/arrival
- Distribution of interarrival times
- Merging and splitting
- Poisson approximation

The Bernoulli process

- A sequence of independent Bernoulli trials, X_i

- At each trial, i :

$$P(X_i = 1) = P(\text{success at the } i\text{th trial}) = p$$

$$P(X_i = 0) = P(\text{failure at the } i\text{th trial}) = 1 - p$$

- Key assumptions:

- Independence
 - Time-homogeneity

- Model of:

- Sequence of lottery wins/losses
 - Arrivals (each second) to a bank
 - Arrivals (at each time slot) to server
 - ...

$$0 < p < 1$$



- Jacob Bernoulli
(1655–1705)

(Image is in the public domain.
Source: [Wikipedia](#))

Stochastic processes

infinite

- First view: sequence of random variables X_1, X_2, \dots

{ Interested in: $E[X_i] = p$ $\text{var}(X_i) = p(1-p)$ $p_{X_i}(x) = \begin{cases} p & x=1 \\ 1-p & x=0 \end{cases}$

$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = P_{X_1}(x_1) \dots P_{X_n}(x_n)$

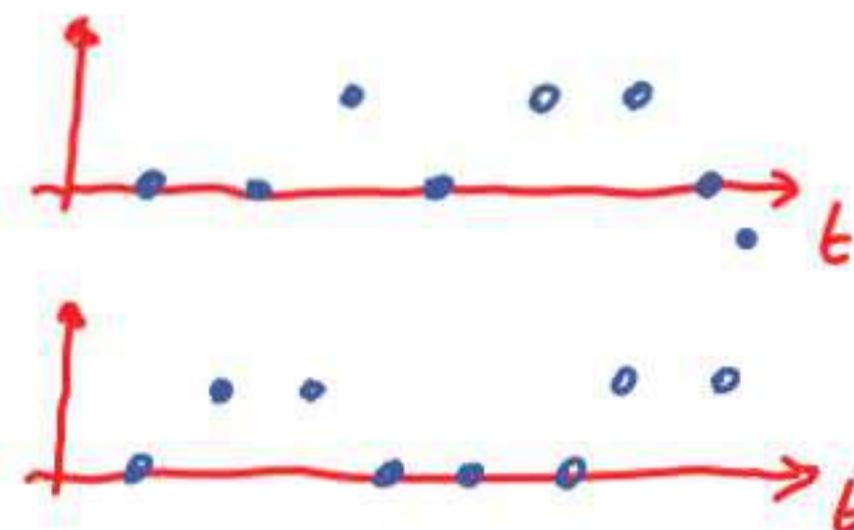
for all n

- Second view – sample space:

{ $\Omega = \text{set of infinite sequences of 0's and 1's}$

- Example (for Bernoulli process):

$$P(X_i = 1 \text{ for all } i) = 0 \quad (p < 1)$$



$$\leq P(X_1 = 1, \dots, X_n = 1) = p^n, \text{ for all } n$$

Number of successes/arrivals S in n time slots

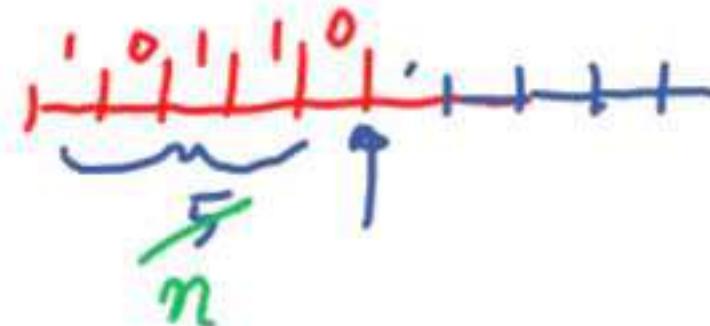
- $S = X_1 + \dots + X_n$
- $P(S = k) = \binom{n}{k} p^k (1-p)^{n-k}$ $k = 0, \dots, n$
- $E[S] = np$
- $\text{var}(S) = np(1-p)$

Time until the first success/arrival

- $T_1 = \min\{i : X_i = 1\}$
- $P(T_1 = k) = P(\underbrace{00\cdots 0}_{k-1} 1) = (1-p)^{k-1} p$
 $k = 1, 2, \dots$
- $E[T_1] = \frac{1}{p}$
- $\text{var}(T_1) = \frac{1-p}{p^2}$
•

Independence, memorylessness, and fresh-start properties

$$\{X_i\} \sim \text{Ber}(p)$$



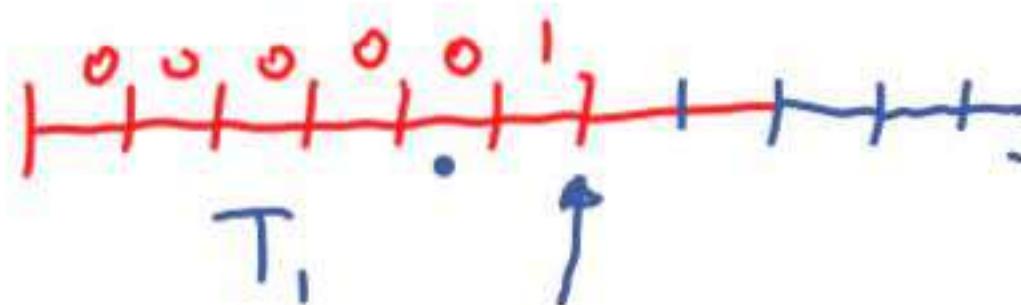
$$Y_1 = X_6^{X_{n+1}} \quad \{Y_i\}$$

$$Y_2 = X_7^{X_{n+2}} \quad i=1, 2, \dots$$

⋮

- ① $\{Y_i\}$ independent of X_1, \dots, X_{5n}
- ② $\text{Ber}(p)$

- Fresh-start after time n



$$Y_1 = X_{T_1+1} \quad \{Y_i\}$$

$$Y_2 = X_{T_1+2} \quad \text{independent}$$

⋮

of X_1, \dots, X_{T_1}

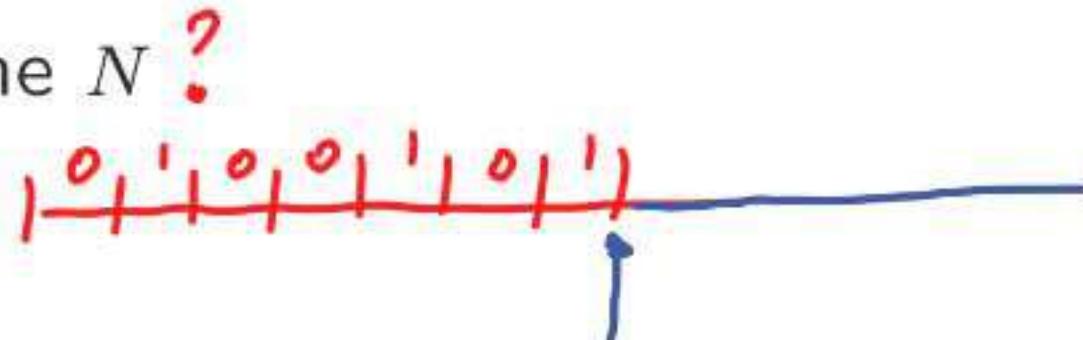
② $\text{Ber}(p)$

- Fresh-start after time T_1

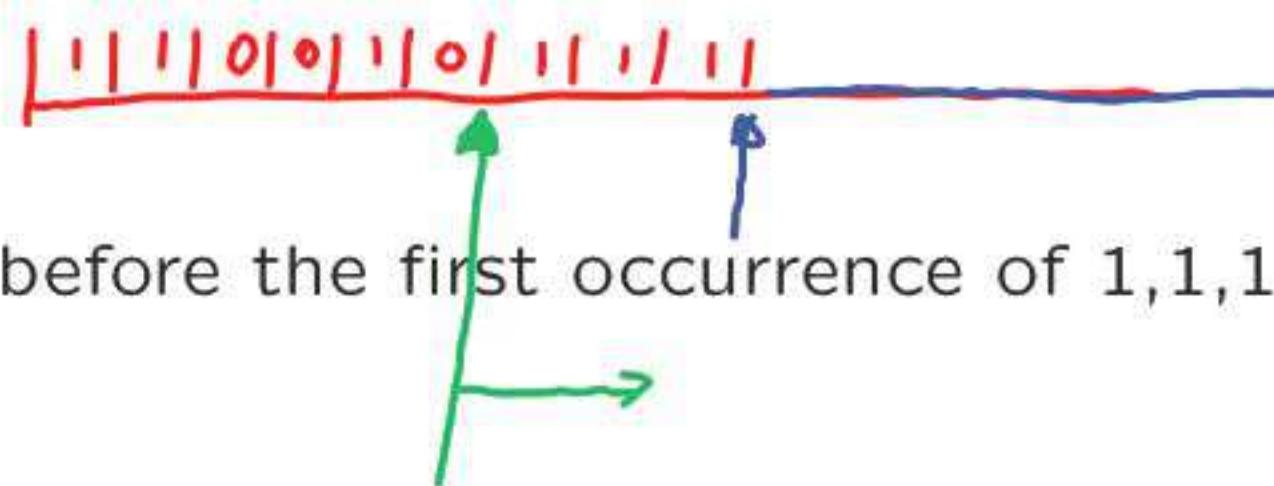
Independence, memorylessness, and fresh-start properties

- Fresh-start after a random time N ?

N = time of 3rd success



N = first time that 3 successes in a row have been observed



N = the time just before the first occurrence of 1,1,1
=

$\} N$ is causally determined

$\} N$ not causally determined

The process X_{N+1}, X_{N+2}, \dots is:

- a Bernoulli process
- independent of N, X_1, \dots, X_N

(as long as N is determined "causally")

The distribution of busy periods

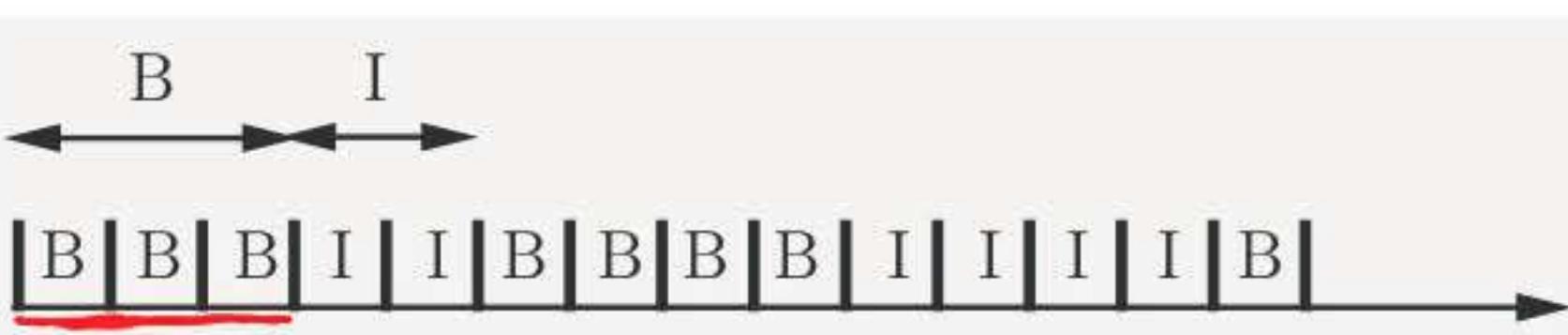
- At each slot, a server is busy or idle (Bernoulli process)

P

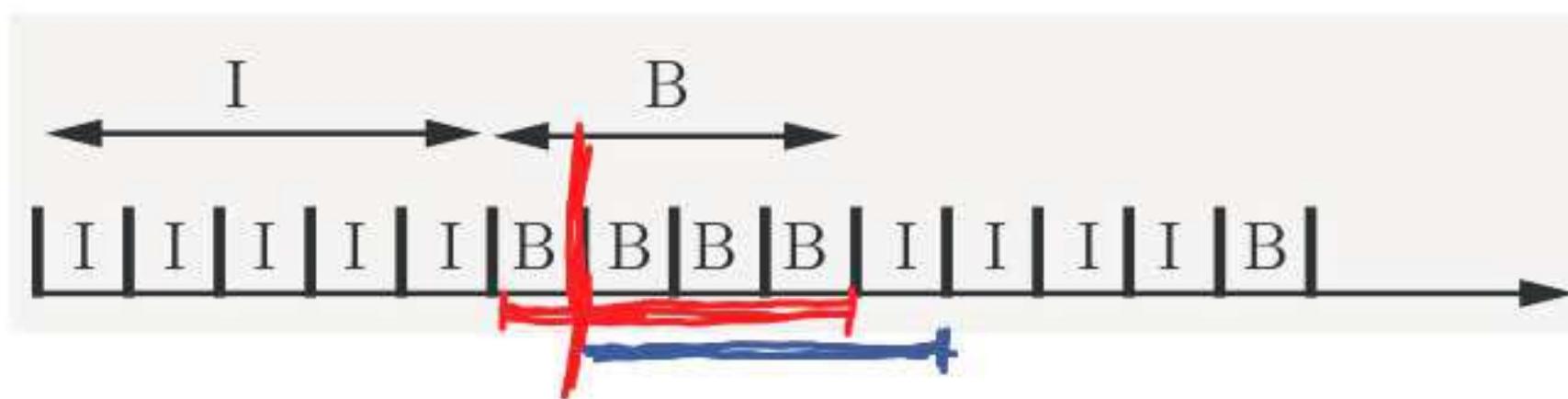
- First busy period:

$\text{Geo}(1-p)$

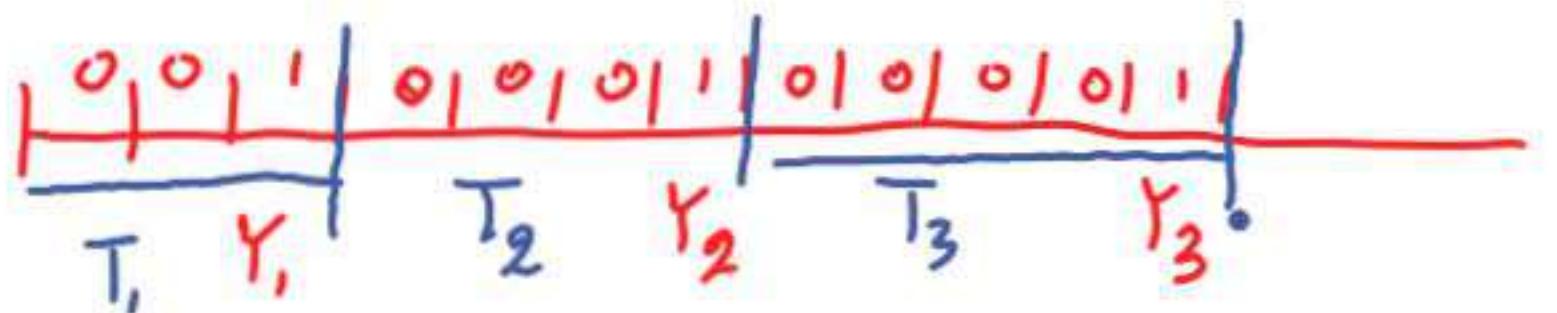
- starts with first busy slot
- ends just before the first subsequent idle slot



$\text{Geo}(1-p)$



Time of the k th success/arrival



- Y_k = time of k th arrival $Y_k = T_1 + \dots + T_k$
- T_k = k th inter-arrival time = $Y_k - Y_{k-1}$ ($k \geq 2$)
- The process starts fresh after time T_1
- T_2 is independent of T_1 ; Geometric(p); etc.

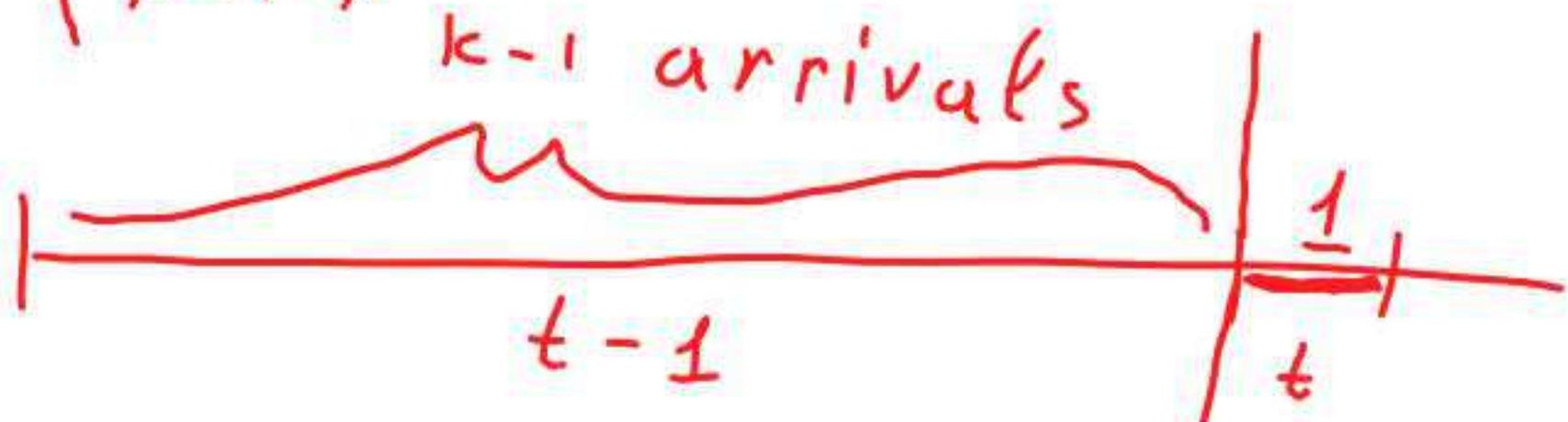
Time of the k th success/arrival

$$P(Y_k = t)$$

$= P(k-1 \text{ arrivals in time } t-1)$

$\cdot P(\text{arrival at time } t)$

$$= \binom{t-1}{k-1} p^{k-1} (1-p)^{t-k} \cdot p$$



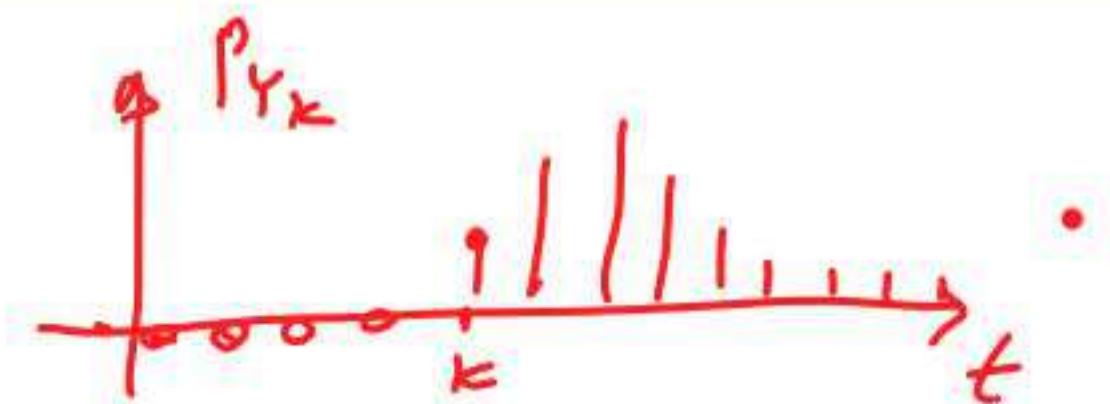
$$Y_k = T_1 + \cdots + T_k$$

the T_i are i.i.d., Geometric(p)

$$E[Y_k] = \frac{k}{p} \quad \text{var}(Y_k) = \frac{k(1-p)}{p^2}$$

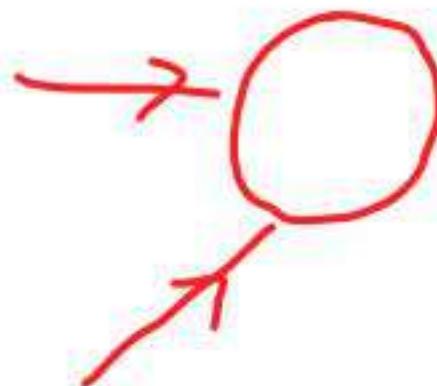
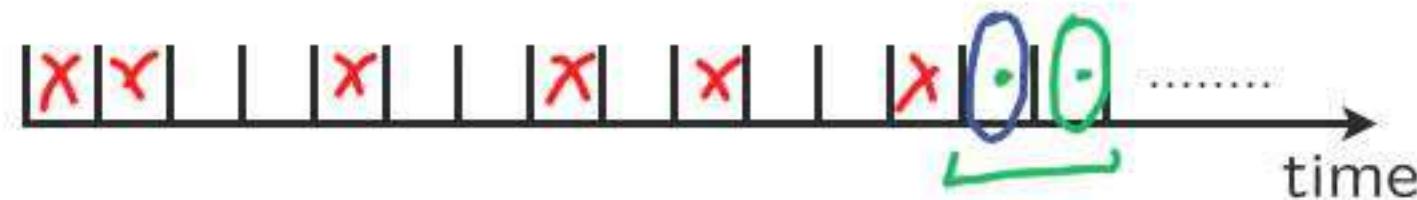
$$p_{Y_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k},$$

$$\underline{t = k, k+1, \dots}$$



Merging of independent Bernoulli processes

X_t Bernoulli(p)

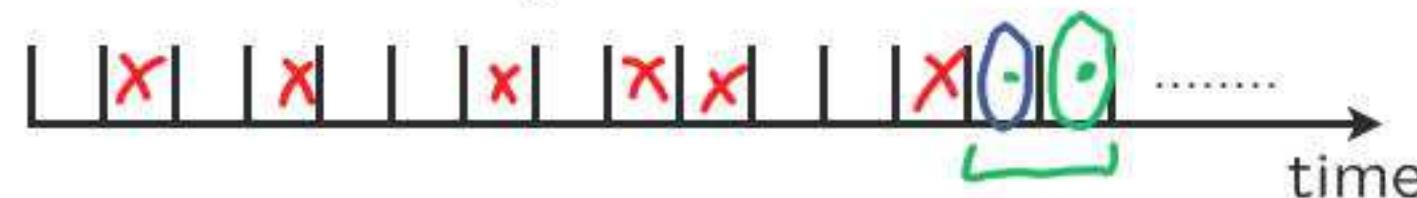


Z_t merged process

Bernoulli($p + q - pq$)

(collisions are counted as one arrival)

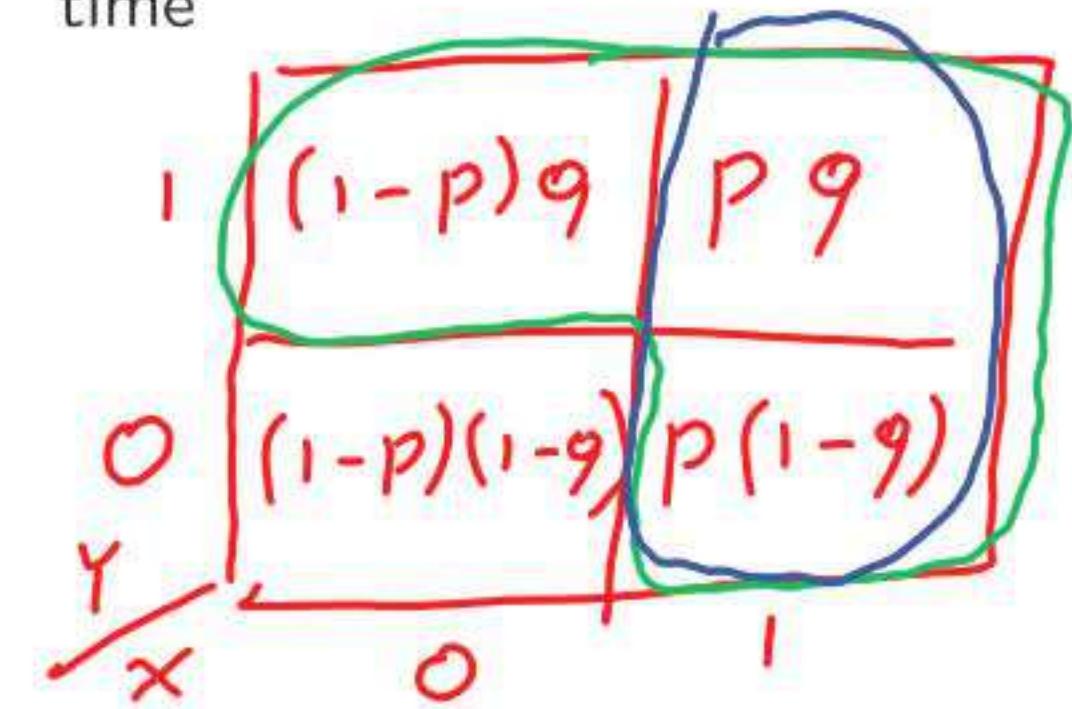
Y_t Bernoulli(q)



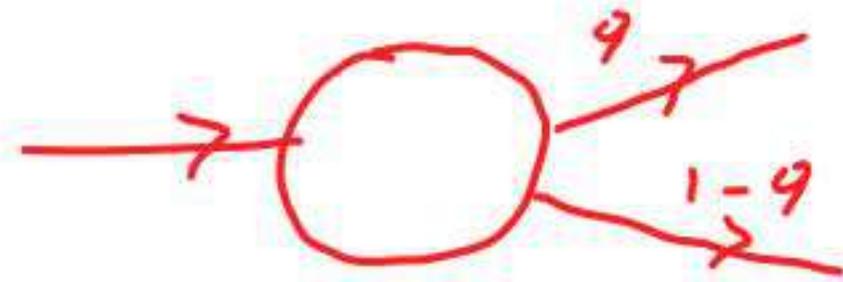
$$Z_t = g(X_t, Y_t) \quad (Z_1, \dots, Z_t)$$

$$Z_{t+1} = g(X_{t+1}, Y_{t+1}) \quad 1 - (1-p)(1-q)$$

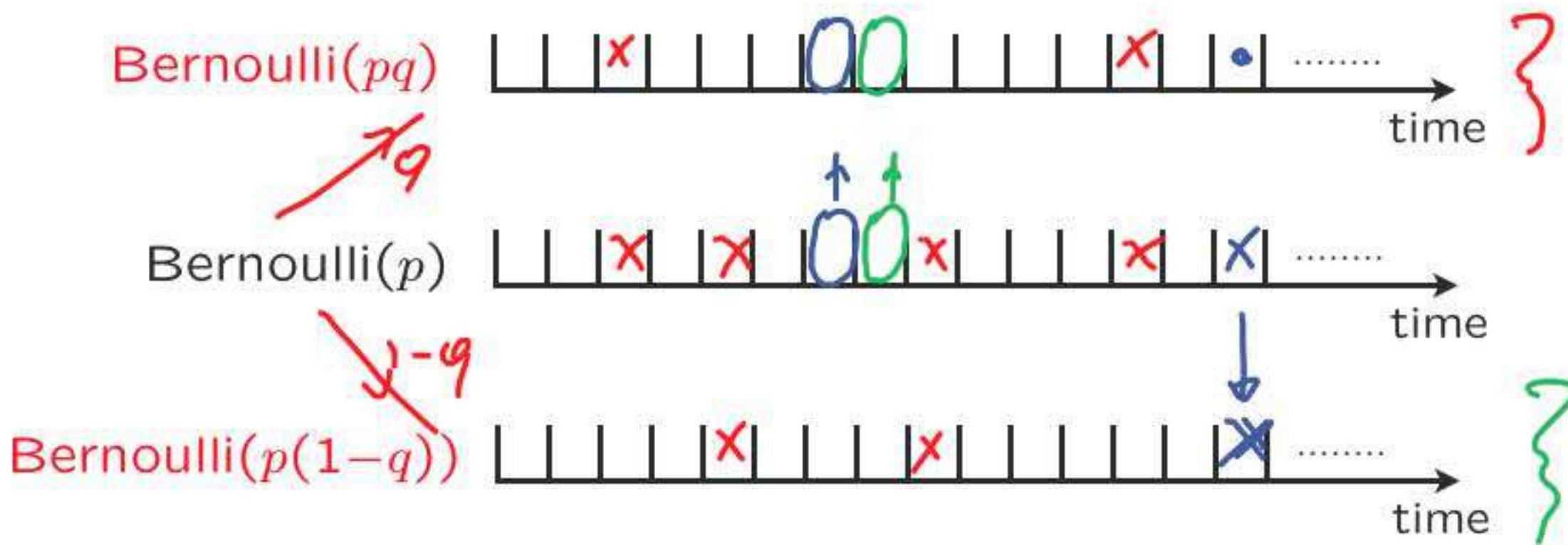
$$P(\text{arrival in first process} \mid \text{arrival}) = \frac{p}{p+q-pq}$$



Splitting of a Bernoulli process



- Split successes into two streams, using independent flips of a coin with bias q
 - assume that coin flips are independent from the original Bernoulli process



- Are the two resulting streams independent? **No**

Poisson approximation to binomial

- Interesting regime: large n , small p , moderate $\lambda = np$

$$\begin{array}{l} \cdot n \rightarrow \infty \\ \cdot p \rightarrow 0 \quad p = \frac{\lambda}{n} \end{array}$$

- Number of arrivals S in n slots: $p_S(k) = \frac{n!}{(n-k)!k!} \cdot p^k (1-p)^{n-k}, \quad k = 0, \dots, n$

For fixed $k = 0, 1, \dots$,

$$p_S(k) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda},$$

$$= \frac{n \cdot (n-1) \cdots (n-k+1)}{k!} \cdot \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} \cdot \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}$$

$$\xrightarrow[n \rightarrow \infty]{} 1 \cdot 1 \cdots 1 \cdot \underbrace{\frac{\lambda^k}{k!} e^{-\lambda}}_1 \cdot 1$$

- Fact: $\lim_{n \rightarrow \infty} (1 - \lambda/n)^n = e^{-\lambda}$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 22: The Poisson process

- Definition of the Poisson process
 - applications
- Distribution of number of arrivals
- The time of the k th arrival
- Memorylessness
- Distribution of interarrival times

Definition of the Poisson process

Poisson



$$\sum_{k=0}^{\infty} P(k, \tau) = 1$$

- Numbers of arrivals in disjoint time intervals are **independent**

$P(k, \tau)$ = Prob. of k arrivals in interval of duration τ

- Small interval probabilities:**

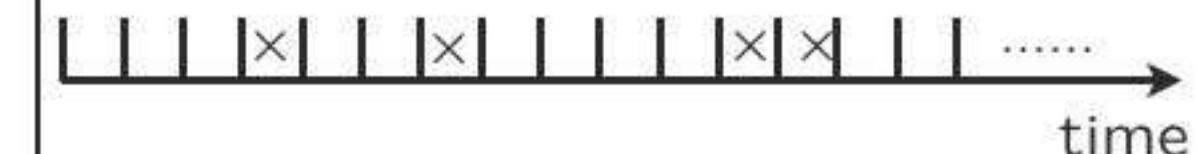
For VERY small δ :

$$P(k, \delta) \approx \begin{cases} 1 - \lambda\delta & \text{if } k = 0 \\ \lambda\delta & \text{if } k = 1 \\ 0 & \text{if } k > 1 \end{cases}$$

$$P(k, \delta) = \begin{cases} 1 - \lambda\delta + O(\delta^2) & \text{if } k = 0 \\ \lambda\delta + O(\delta^2) & \text{if } k = 1 \\ 0 + O(\delta^2) & \text{if } k > 1 \end{cases}$$

$$\frac{O(\delta^2)}{\delta} \xrightarrow{\delta \rightarrow 0} 0$$

Bernoulli



- Independence
- Time homogeneity:**
Constant p at each slot

λ : "arrival rate"

Applications of the Poisson process



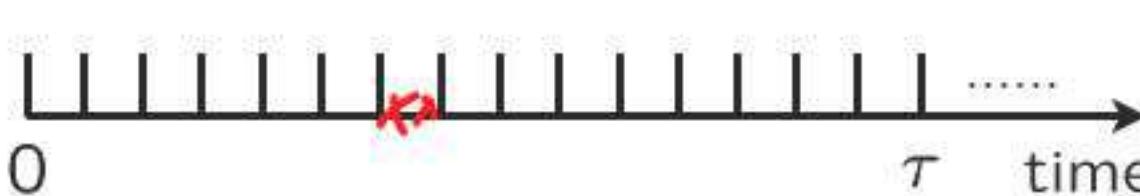
- Deaths from horse kicks in the Prussian army (1898)
- Particle emissions and radioactive decay
- Photon arrivals from a weak source
- Financial market shocks
- Placement of phone calls, service requests, etc.



Siméon Denis Poisson
(1781-1840)

(This image is in the public domain.
Source: [Wikipedia](#))

The Poisson PMF for the number of arrivals



$$P(k, \delta)$$

- N_τ : arrivals in $[0, \tau]$ $P(k, \tau) = P(N_\tau = k)$

$n = \tau/\delta$ intervals/slots of length δ ~~is small~~

$P(\text{some slot contains two or more arrivals})$

$$\begin{aligned} &\leq \sum_i P(\text{slot } i \text{ has } \geq 2 \text{ arrivals}) \\ &= \frac{\tau}{\delta} O(\delta^2) \xrightarrow{\delta \rightarrow 0} 0 \end{aligned}$$

$P(k \text{ arrivals in Poisson}) \approx P(k \text{ slots have arrival})$

$N_\tau \approx \text{binomial}$

$$p = \lambda\delta + O(\delta^2)$$

$$np = \lambda\tau + O(\delta) \approx \lambda\tau$$

Bernoulli

$$p_S(k) = \frac{n!}{(n-k)! k!} \cdot p^k (1-p)^{n-k}, \quad k = 0, \dots, n$$

$$\lambda = np \quad n \rightarrow \infty \quad p \rightarrow 0$$

For fixed $k = 0, 1, \dots$,

$$p_S(k) \xrightarrow{k!} \frac{\lambda^k}{k!} e^{-\lambda},$$

$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

Mean and variance of the number of arrivals

$$P(k, \tau) = \mathbf{P}(N_\tau = k) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

$$\mathbf{E}[N_\tau] = \lambda\tau$$

$$\text{var}(N_\tau) = \lambda\tau$$

$$\mathbf{E}[N_\tau] = \sum_{k=0}^{\infty} k \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!} = \dots = \lambda\tau$$

$$\lambda = \frac{\mathbf{E}[N_\tau]}{\tau}$$

$N_\tau \approx \text{Binomial}(n, p)$

$$n = \tau/\delta, \quad p = \lambda\delta + O(\delta^2)$$

$$\mathbf{E}[N_\tau] \approx np \approx \lambda\tau$$

$$\text{var}(N_\tau) \approx np(1-p) \approx \lambda\tau$$

Example

- You get email according to a Poisson process, at a rate of $\lambda = 5$ messages per hour.

$$\mathbf{E}[N_\tau] = \lambda\tau$$

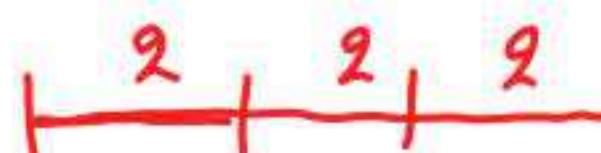
$$\text{var}(N_\tau) = \lambda\tau$$

- Mean and variance of mails received during a day = $5 \cdot 24$

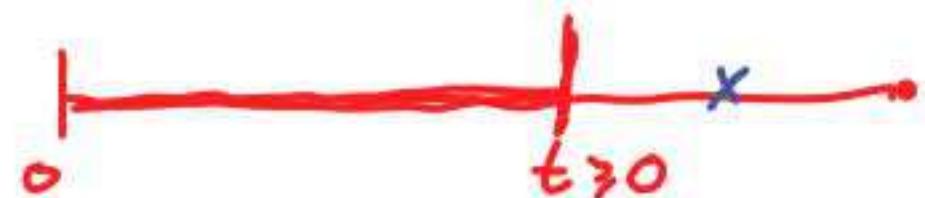
- $P(\text{one new message in the next hour}) = P(1,1) = 5e^{-5}$

$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

- $P(\text{exactly two messages during each of the next three hours}) =$


$$(P(2,1))^3 = \left(\frac{5^2 e^{-5}}{2}\right)^3$$

The time T_1 until the first arrival



$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

- Find the CDF: $P(T_1 \leq t) =$

$$= 1 - P(T_1 > t) = 1 - P(0, t) = 1 - e^{-\lambda t}$$

$$f_{T_1}(t) = \lambda e^{-\lambda t}, \quad \text{for } t \geq 0$$

Exponential(λ)

Memorylessness: conditioned on $T_1 > t$,
the PDF of $\underline{T_1 - t}$ is again exponential

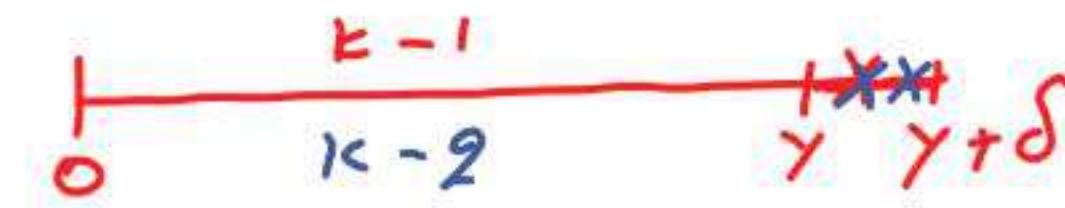
The time Y_k of the k th arrival

$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

- Can derive its PDF by first finding the CDF
- More intuitive argument:

$$\begin{aligned} f_{Y_k}(y) \delta &\approx P(y \leq Y_k \leq y + \delta) = \\ &\approx [P(k-1, y) \lambda \delta] - \\ &+ P(k-2, y) O(\delta^2) \\ &+ P(k-3, y) O(\delta^3) \end{aligned}$$

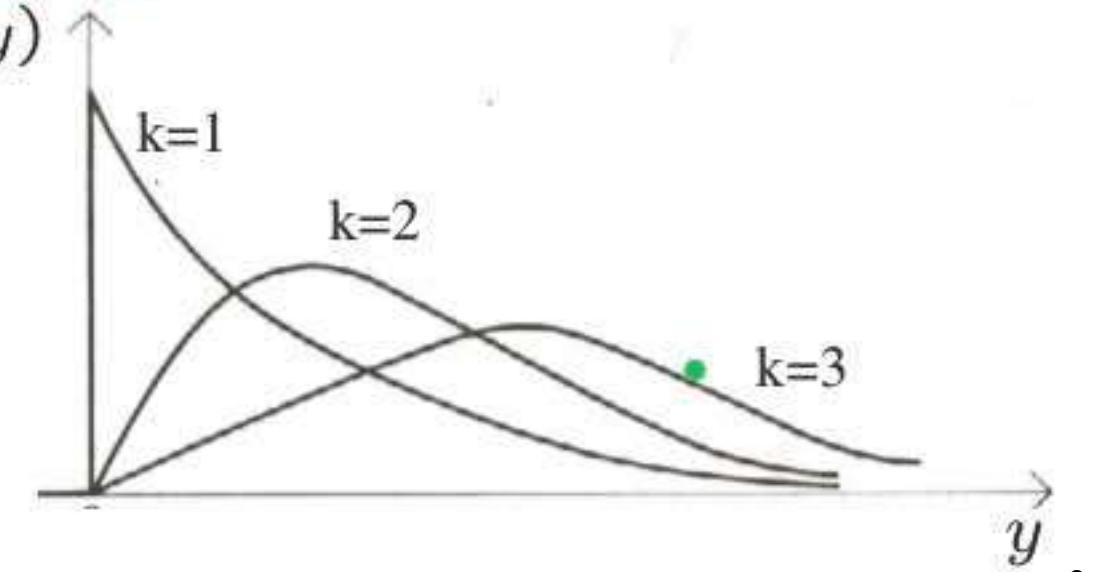
$$P(Y_k \leq y) = \sum_{n=k}^{\infty} P(n, y)$$



$$\frac{(dy)^{k-1} e^{-\lambda y}}{(k-1)! f_{Y_k}(y)}$$

Erlang distribution: $f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0$

order k

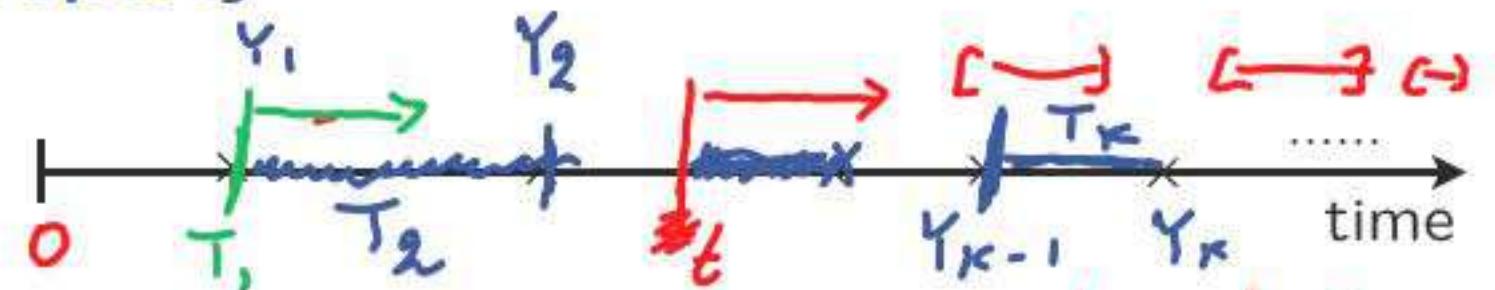


Memorylessness and the fresh-start property

- Analogous to the properties for the Bernoulli process
 - plausible, given the relation between the two processes
 - use intuitive reasoning
 - can be proved rigorously

Memorylessness and the fresh-start property

- If we start watching at time t ,



we see Poisson process, independent of the history until time t *start fresh*

- time until next arrival: $\text{Exp}(\lambda)$, *independent of past*

- If we start watching at time T_1 , $T_1 = 3$

we see Poisson process, independent of the history until time T_1

- hence: time between first and second arrival, $T_2 = Y_2 - Y_1$ is: $\text{Exp}(\lambda)$

- similarly for all $T_k = Y_k - Y_{k-1}$, $k \geq 2$

inol. of T_1

$Y_k = T_1 + \dots + T_k$ is sum of i.i.d. exponentials

$$\mathbf{E}[Y_k] = k/\lambda \quad \mathbf{var}(Y_k) = k/\lambda^2$$

- An equivalent definition
- A simulation method

Bernoulli/Poisson relation



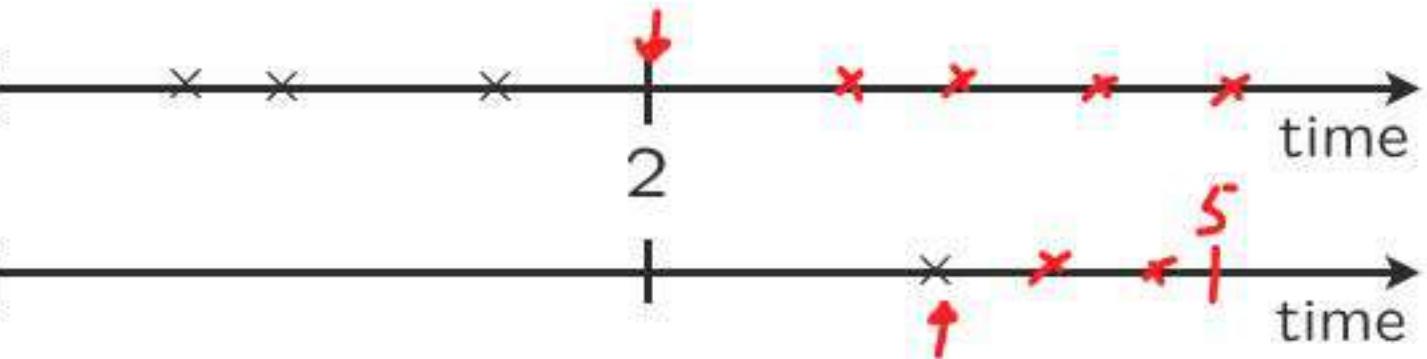
$$n = \tau/\delta,$$
$$p = \lambda\delta$$

$$np = \lambda\tau$$

	POISSON	BERNOULLI
Times of Arrival	Continuous	Discrete
Arrival Rate	$\lambda/\text{unit time}$	$p/\text{per trial}$
PMF of # of Arrivals	Poisson	Binomial
Interarrival Time Distr.	Exponential	Geometric
Time to k -th arrival	Erlang	Pascal

Example: Poisson fishing

- Fish are caught as a Poisson process, $\lambda = 0.6/\text{hour}$
 - fish for two hours;
 - if you caught at least one fish, stop
 - else continue until first fish is caught



$P(\text{fish for more than two hours}) = P(0, 2)$

$$P(T_1 > 2) = \int_2^{\infty} f_{T_1}(t) dt$$

$P(\text{fish for more than two and less than five hours}) =$

$$P(0, 2) (1 - P(0, 3))$$

$$P(2 < T_1 \leq 5) = \int_2^5 f_{T_1}(t) dt$$

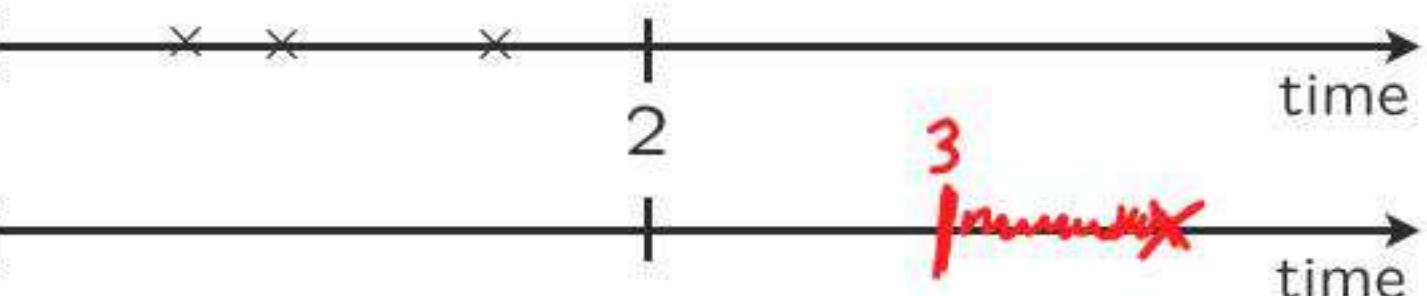
$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}$$

$$E[N_\tau] = \lambda\tau$$

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}$$

Example: Poisson fishing

- Fish are caught as a Poisson process, $\lambda = 0.6/\text{hour}$
 - fish for two hours;
 - if you caught at least one fish, stop
 - else continue until first fish is caught



$P(\text{catch at least two fish}) =$

$$\sum_{k=2}^{\infty} P(k, 2) = 1 - P(0, 2) - P(1, 2)$$

$$P(Y_2 \leq 2) = \int_0^2 f_{Y_2}(y) dy$$

$E[\text{future fishing time} \mid \text{already fished for three hours}] = \frac{1}{\lambda}$

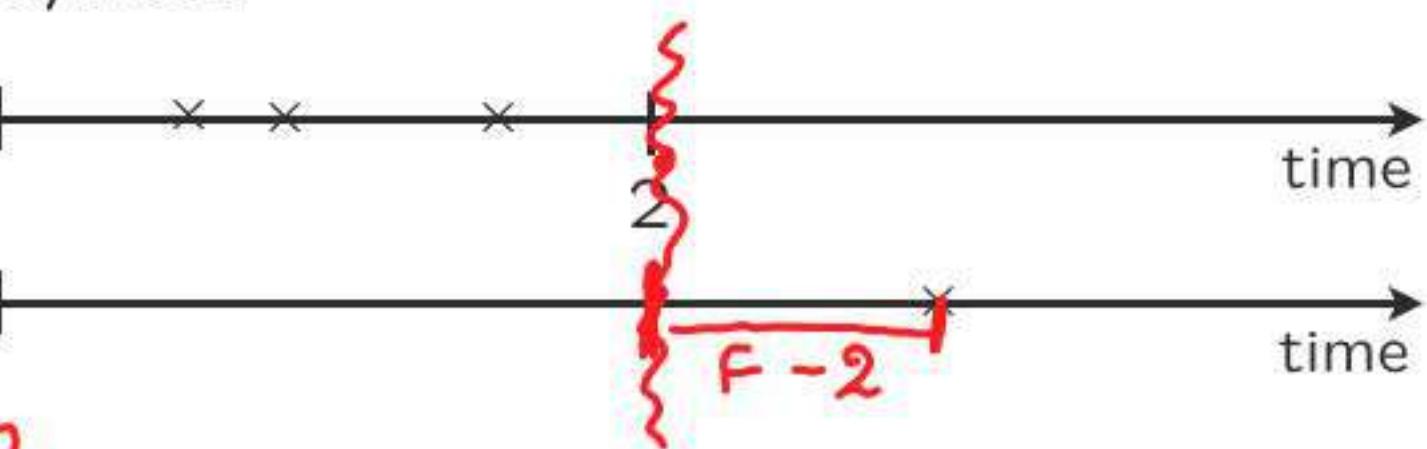
$$P(k, \tau) = \frac{(\lambda \tau)^k e^{-\lambda \tau}}{k!}$$

$$E[N_\tau] = \lambda \tau$$

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}$$

Example: Poisson fishing

- Fish are caught as a Poisson process, $\lambda = 0.6/\text{hour}$
 - fish for two hours;
 - if you caught at least one fish, stop
 - else continue until first fish is caught



$$E[\text{total fishing time}] = E[F] = 2 + E[F-2]$$

$$\begin{aligned} &= 2 + P(F=2) \cdot 0 + P(F>2) E[F-2 | F>2] \\ &= 2 + P(0,2) \cdot 1/\lambda \end{aligned}$$

$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}$$

$$E[N_\tau] = \lambda\tau$$

$$\begin{aligned} E[\text{number of fish}] &= \lambda\tau + P(0,2) \cdot 1 \\ &= 0.6 \cdot 2 \end{aligned}$$

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}$$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

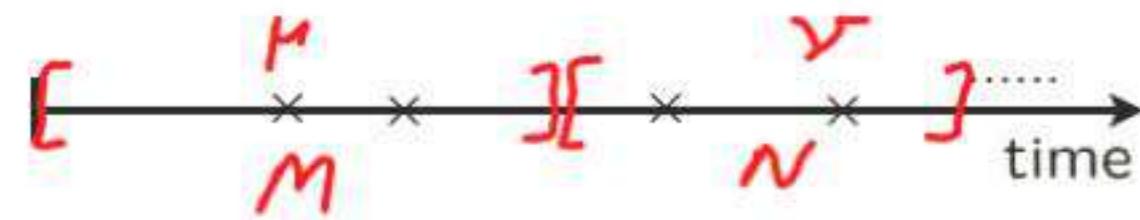
For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

LECTURE 23: More on the Poisson process

- The sum of independent Poisson r.v.s
- Merging and splitting
- Random incidence

The sum of independent Poisson random variables

- Poisson process of rate $\lambda = 1$



- Consecutive intervals of length μ and ν

$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}$$

- Numbers of arrivals during these intervals: M and N

Poisson($\lambda\tau$)

- M : Poisson(μ)

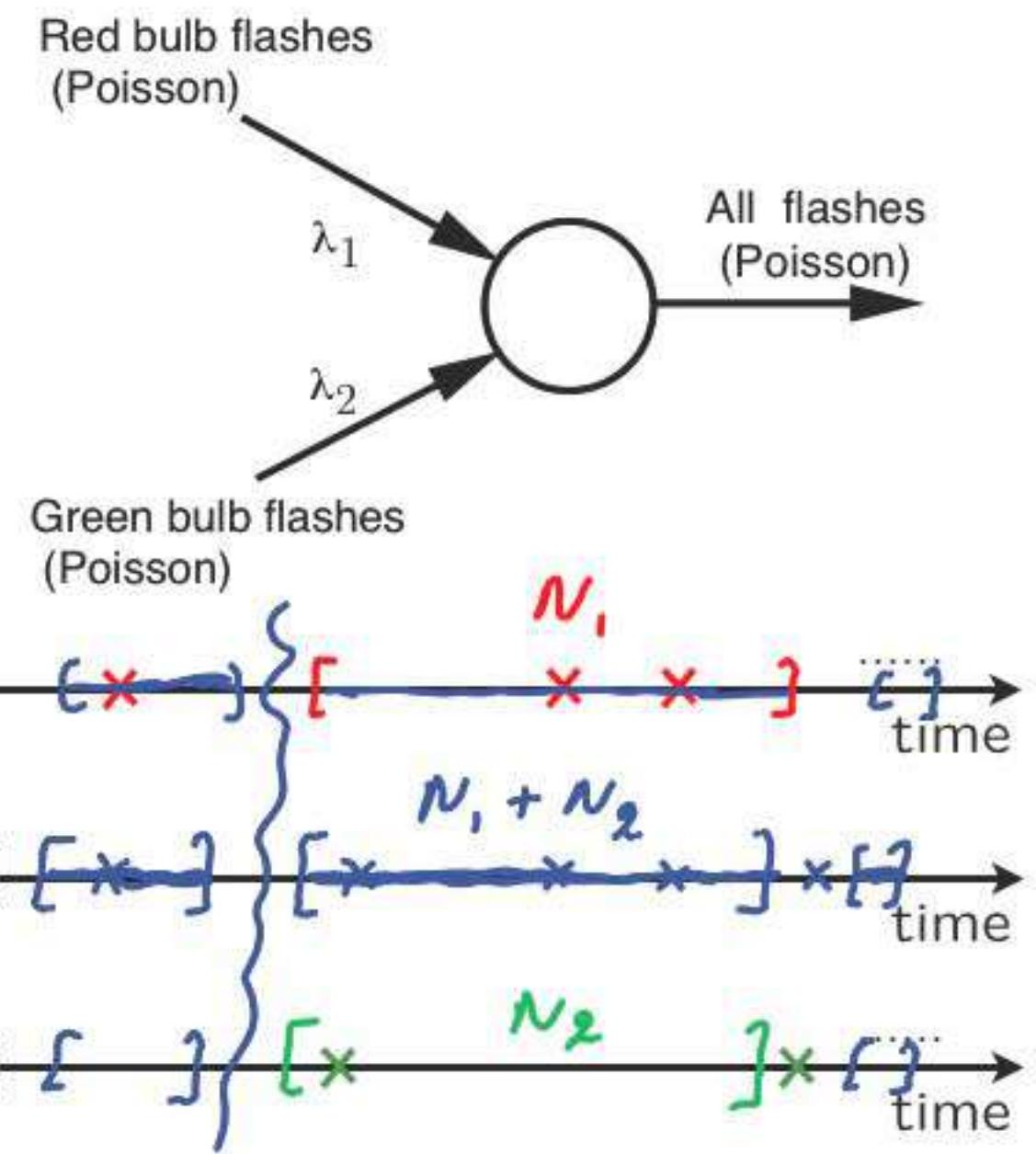
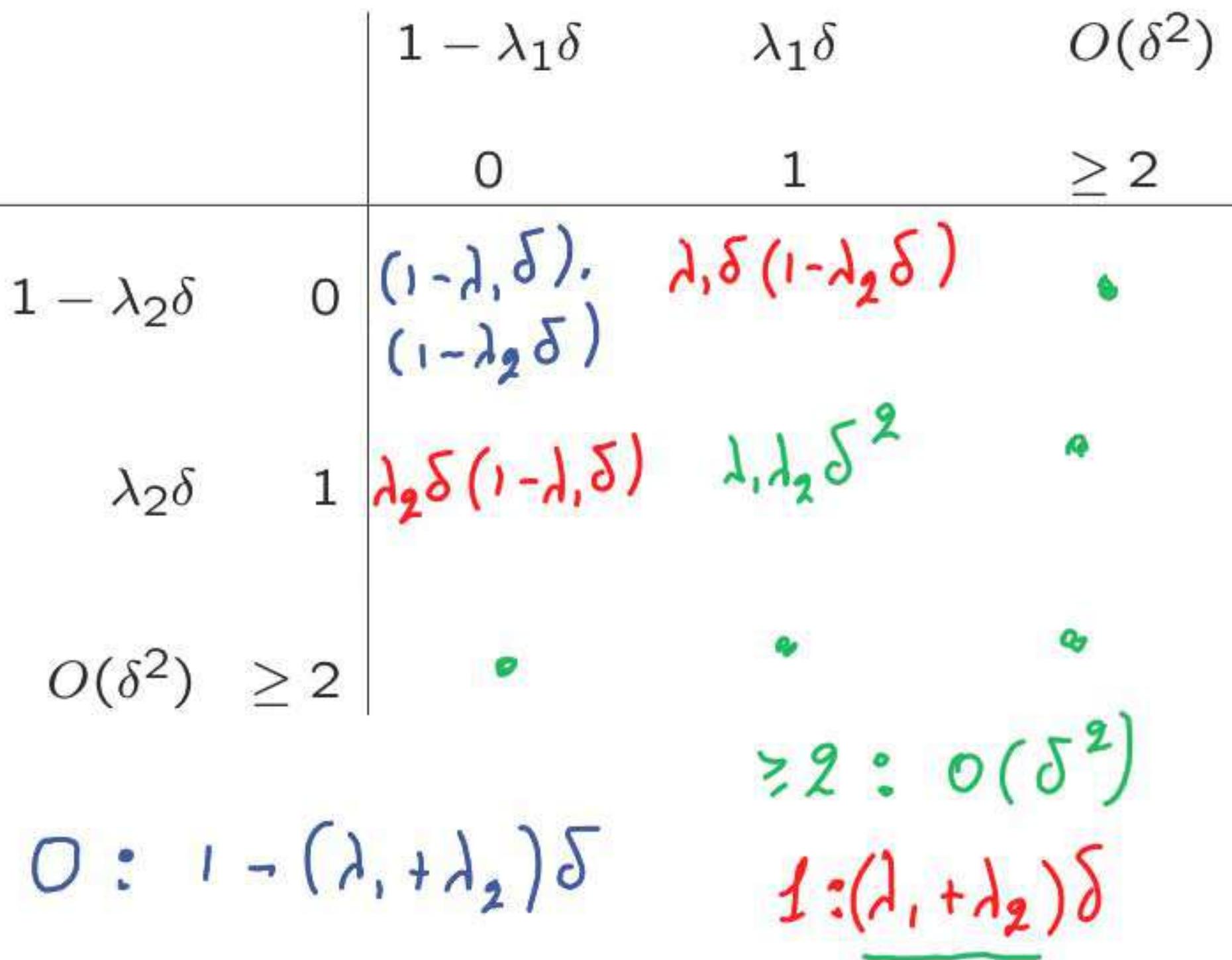
- Independent? Yes

- N : Poisson(ν)

- $M + N$: Poisson($\mu + \nu$)

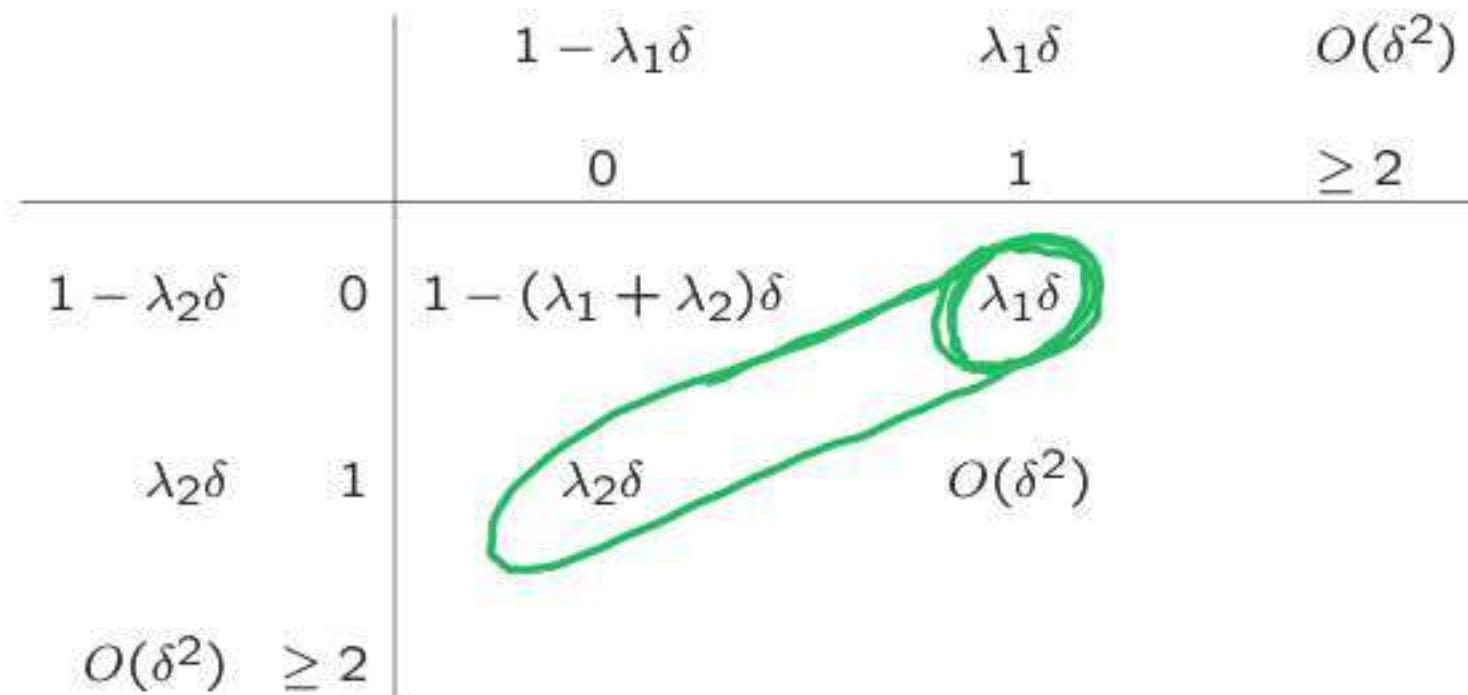
The sum of independent Poisson random variables, with means/parameters μ and ν , is Poisson with mean/parameter $\mu + \nu$

Merging of independent Poisson processes



Where is an arrival of the merged process coming from?

$$P(\text{Red} \mid \text{arrival at time } t) = \lambda_1 / (\lambda_1 + \lambda_2)$$

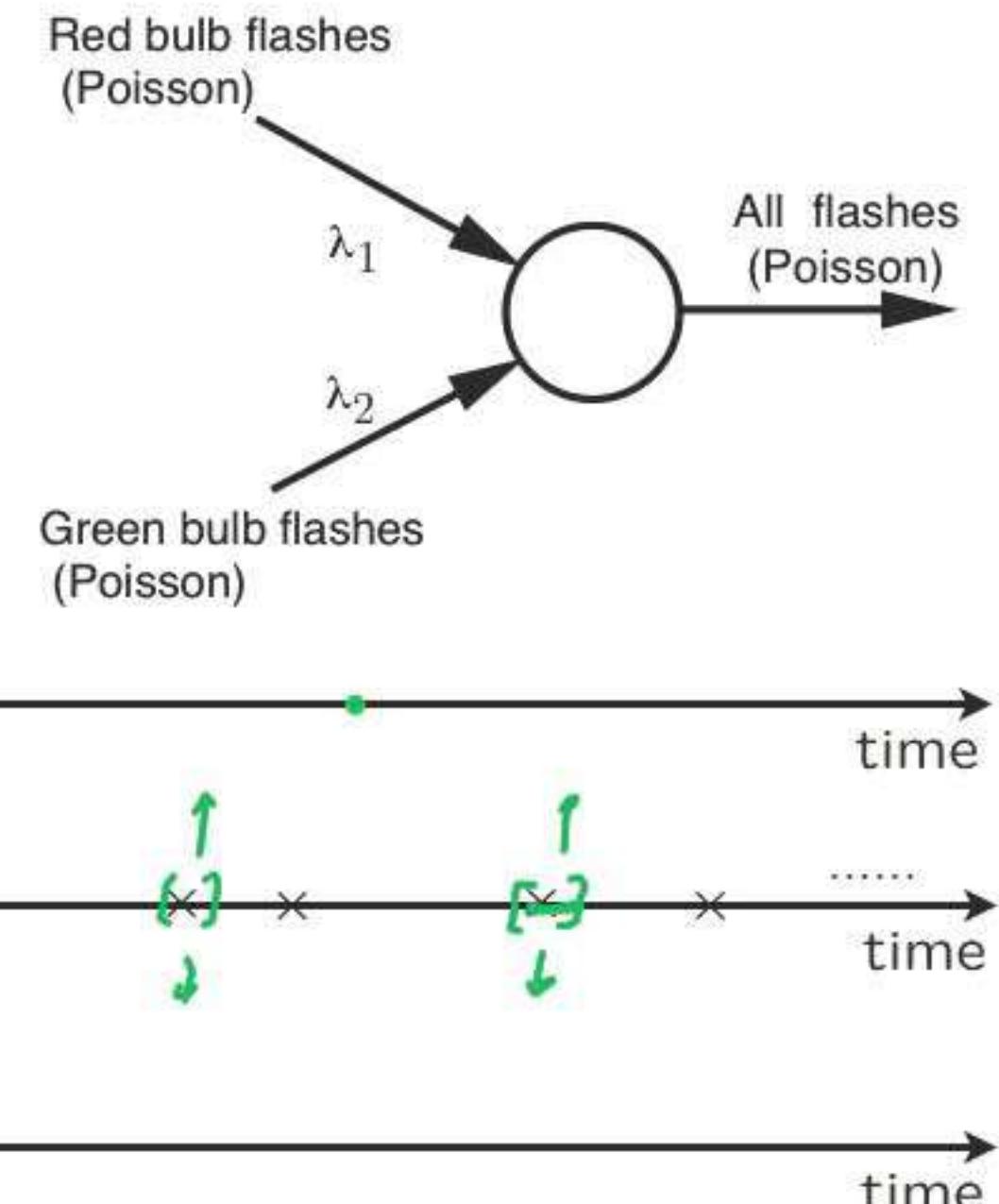


$$P(k\text{th arrival is Red}) = \lambda_1 / (\lambda_1 + \lambda_2)$$

- Independence for different arrivals

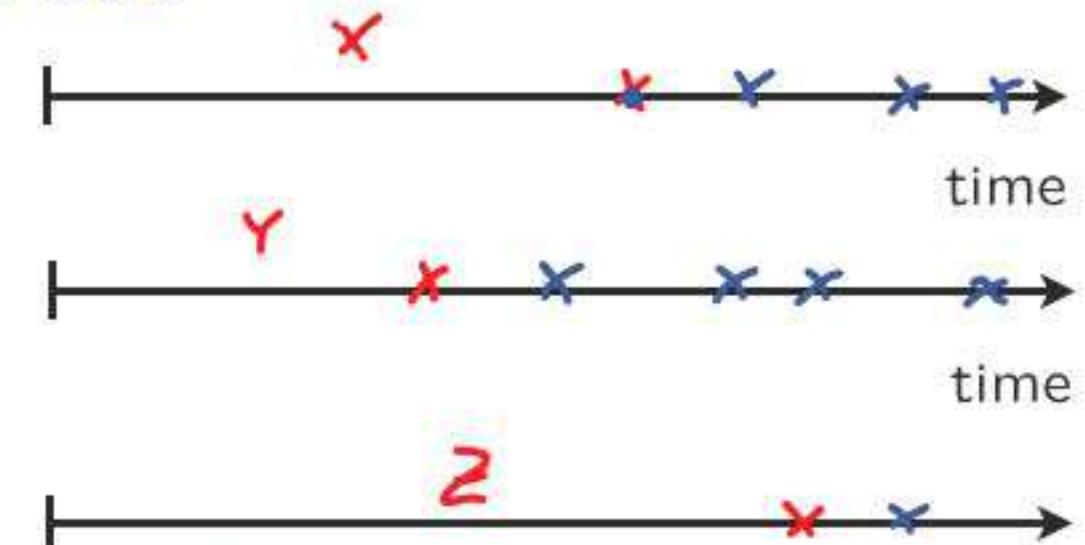
$P(4 \text{ out of first 10 arrivals are Red}) =$

$$\binom{10}{4} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^4 \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^6$$



The time the first (or the last) light bulb burns out

- Three lightbulbs
 - independent lifetimes X, Y, Z ; exponential(λ)
- Find expected time until first burnout = $1/3\lambda$



$$E[\min\{X, Y, Z\}] = \iiint_{000}^{000000} \min\{x, y, z\} \lambda e^{-\lambda x} \lambda e^{-\lambda y} \lambda e^{-\lambda z} dx dy dz$$

$$\Pr(\min\{X, Y, Z\} \geq t) = \Pr(X \geq t, Y \geq t, Z \geq t) = e^{-\lambda t} e^{-\lambda t} e^{-\lambda t} = e^{-3\lambda t}$$

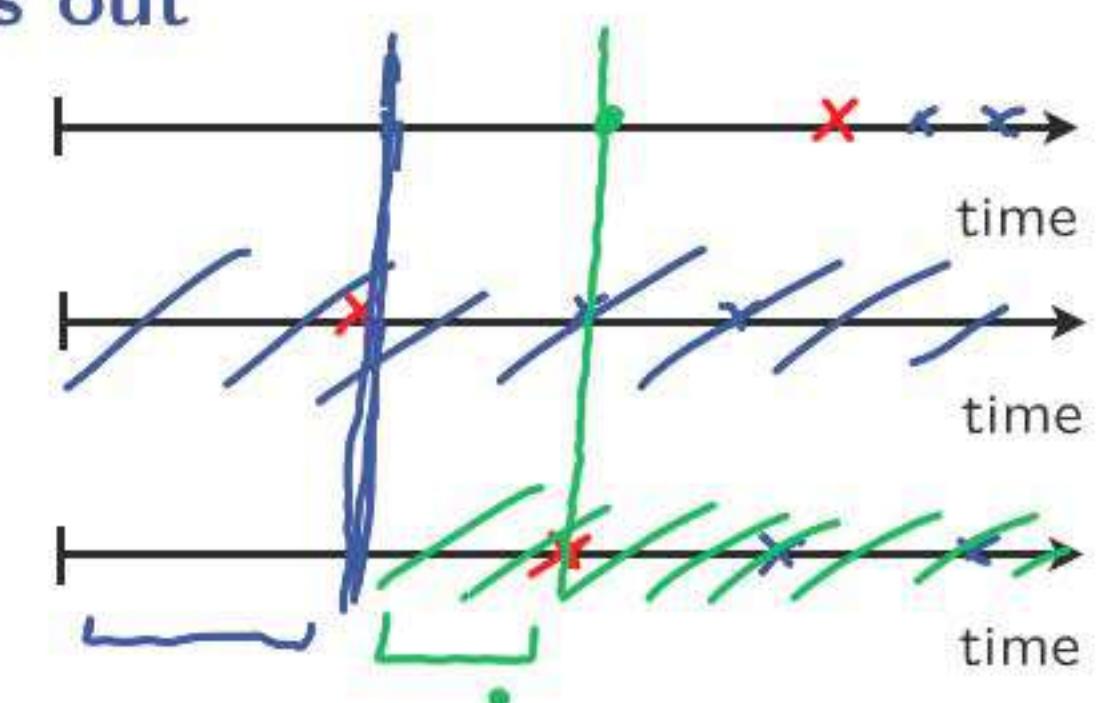
$\text{Exp}(3\lambda)$

- X, Y, Z : first arrivals in independent Poisson processes
- Merged process: $\text{Poisson}(3\lambda)$
- $\min\{X, Y, Z\}$: 1st arrival in merged process $\leftarrow \text{Exp}(3\lambda)$

The time the first (or the last) light bulb burns out

- Three lightbulbs
 - independent lifetimes X, Y, Z ; exponential(λ)
- Find expected time until all burn out

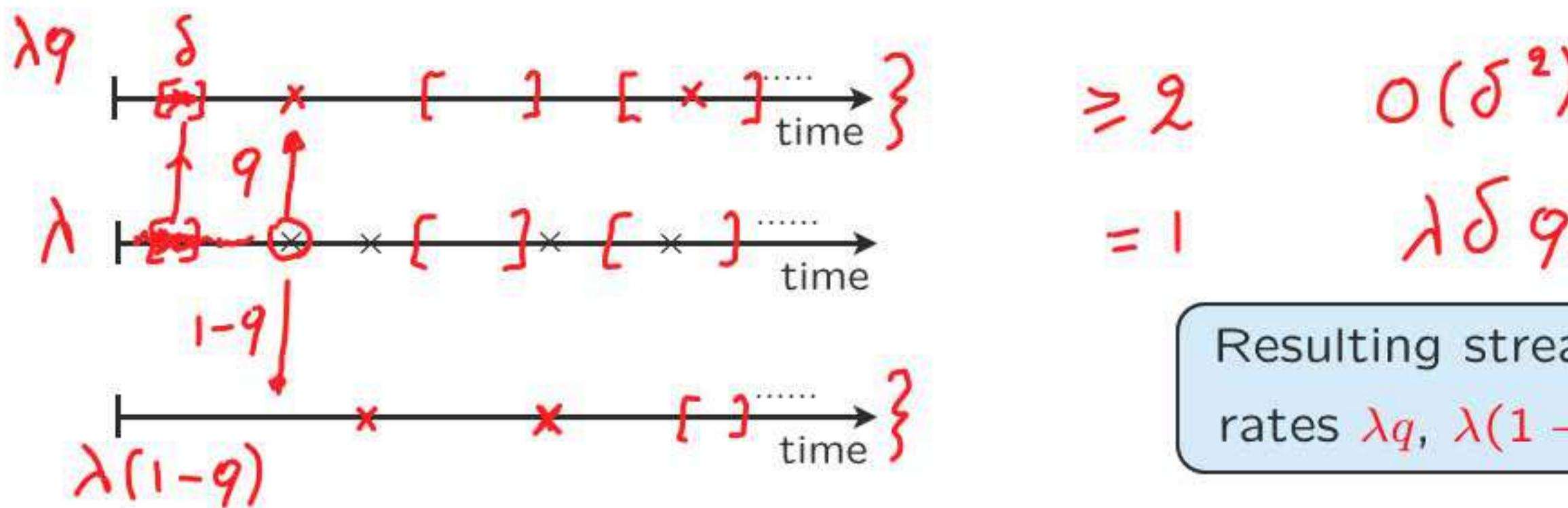
$$\max\{X, Y, Z\}$$



$$\frac{1}{3\lambda} + \frac{1}{2\lambda} + \frac{1}{\lambda}$$

Splitting of a Poisson process

- Split arrivals into two streams, using independent coin flips of a coin with bias q
 - assume that coin flips are independent from the original Poisson process



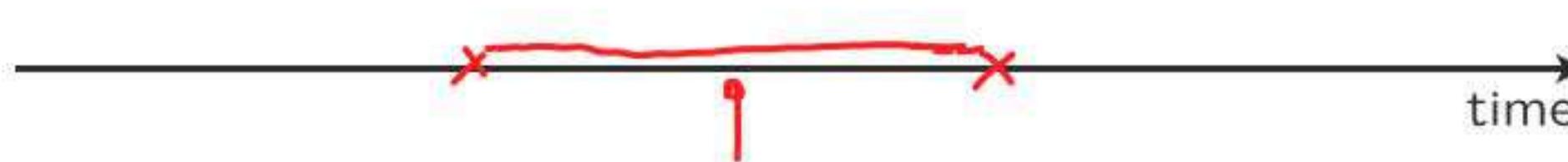
Resulting streams are Poisson,
rates $\lambda q, \lambda(1 - q)$

- Are the two resulting streams independent?

Surprisingly, yes!

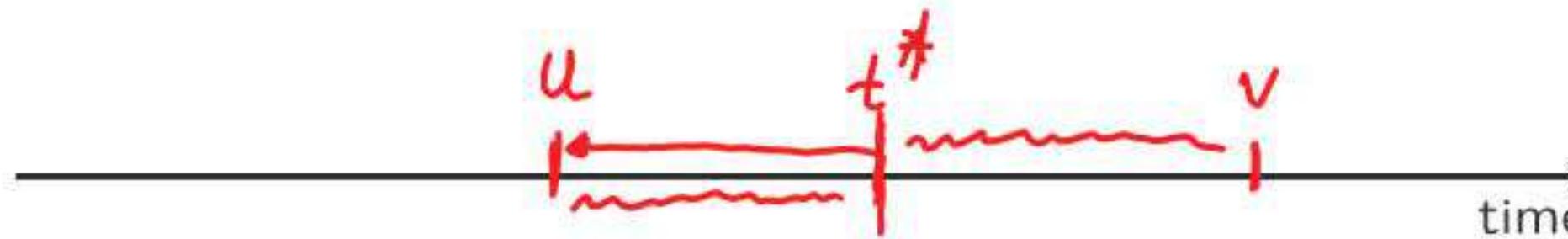
“Random incidence” in the Poisson process

- Poisson process that has been running forever



- Believe that $\lambda = 4/\text{hour}$, so that $E[T_k] = \frac{1}{\lambda} \text{ hrs} = 15 \text{ mins}$
- Show up at some time and measure interarrival time
 - do it many times, average results, see something around 30 mins! Why?

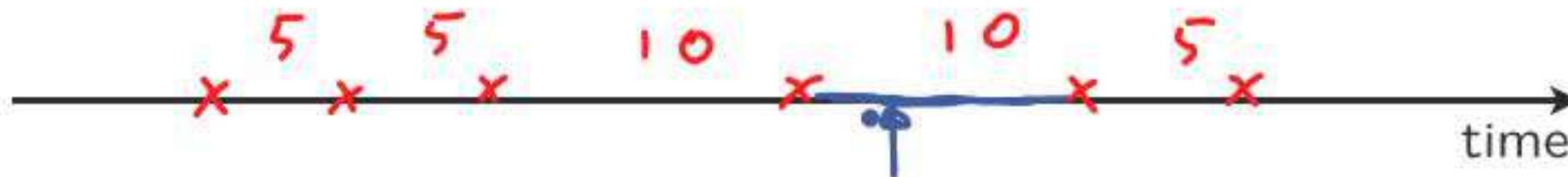
“Random incidence” in the Poisson process — analysis



- Arrive at time t^*
- U : last arrival time • V : next arrival time
- $V - U = \frac{(V - t^*)}{\text{Exp}(\lambda)} + \frac{(t^* - U)}{\text{Exp}(\lambda)}$
- $E[V - U] = \frac{1}{\lambda} + \frac{1}{\lambda} = \frac{2}{\lambda}$
- $V - U$: interarrival time you see, versus k th interarrival time



Random incidence “paradox” is not special to the Poisson process



- **Example:** interarrival times, i.i.d., equally likely to be 5 or 10 minutes

expected value of k th interarrival time: $\frac{1}{2} \cdot 5 + \frac{1}{2} \cdot 10 = 7.5$

- you show up at a “random time”

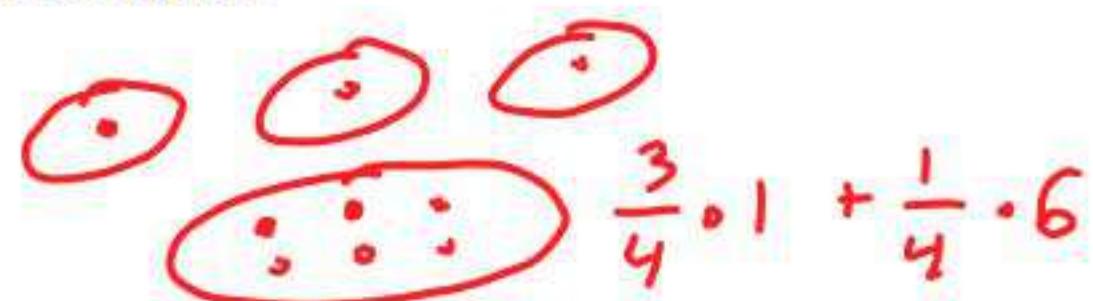
$$P(\text{arrive during a 5-minute interarrival interval}) = \frac{1}{3}$$

$$\begin{aligned} \text{expected length of interarrival interval during which you arrive} &= \frac{1}{3} \cdot 5 + \frac{2}{3} \cdot 10 \\ &\approx 8.3 \end{aligned}$$

- Calculation generalizes to “renewal processes:”
i.i.d. interarrival times, from some general distribution
- “Sampling method” matters

Different sampling methods can give different results

- Average family size?



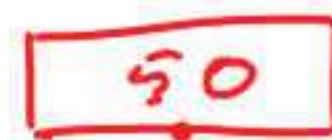
- look at a “random” family (uniformly chosen)

- look at a “random” person’s (uniformly chosen) family

$$\frac{3}{9} \cdot 1 + \frac{6}{9} \cdot 6$$

- Average bus occupancy?

- look at a “random” bus (uniformly chosen)



- look at a “random” passenger’s bus

- Average class size?

MIT OpenCourseWare
<https://ocw.mit.edu>

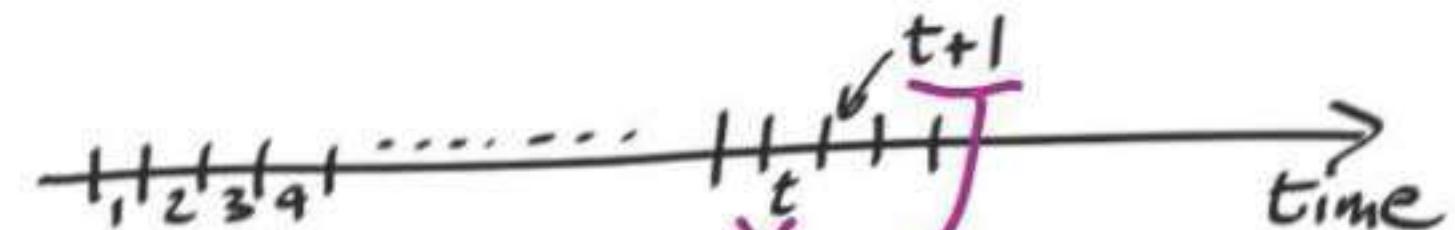
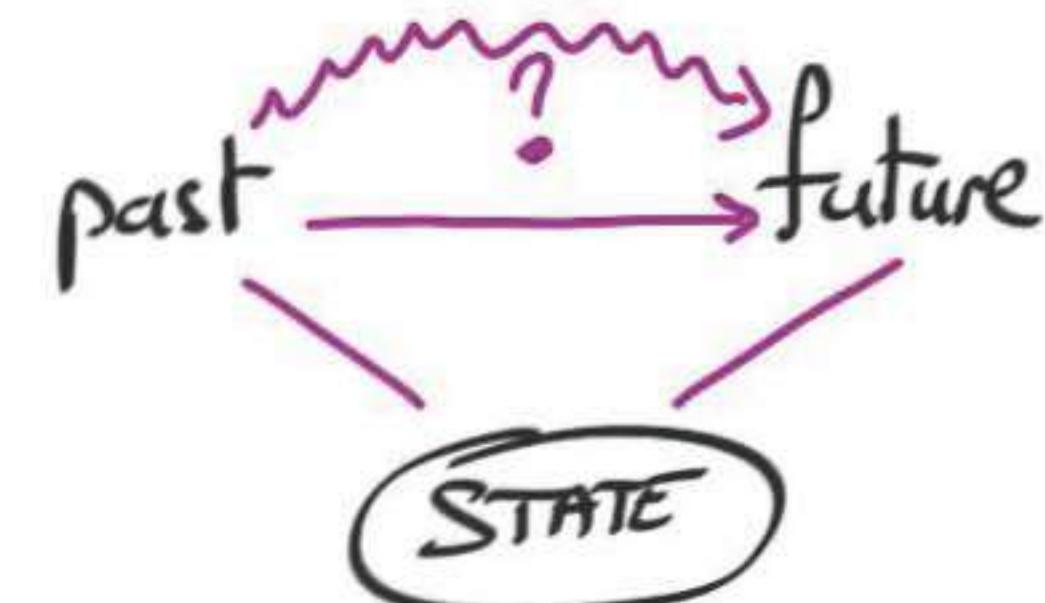
Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

Markov processes – I

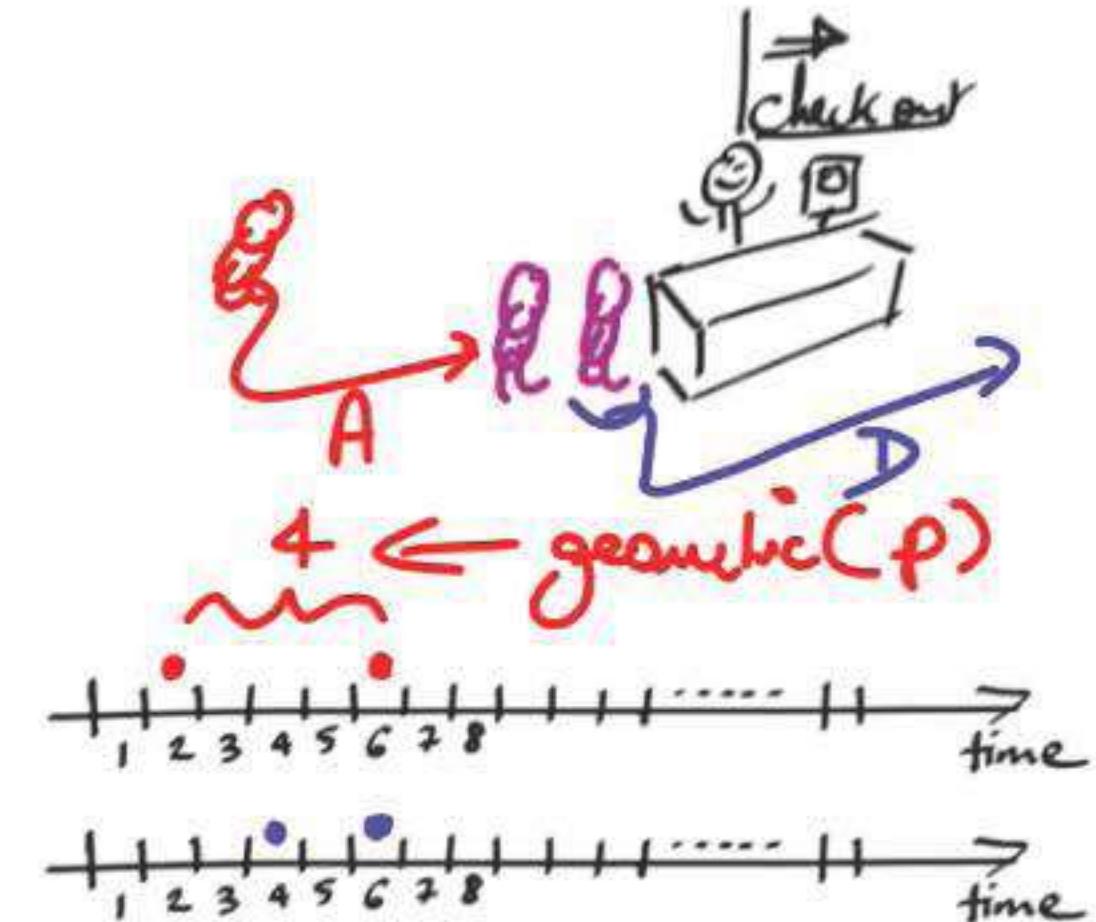
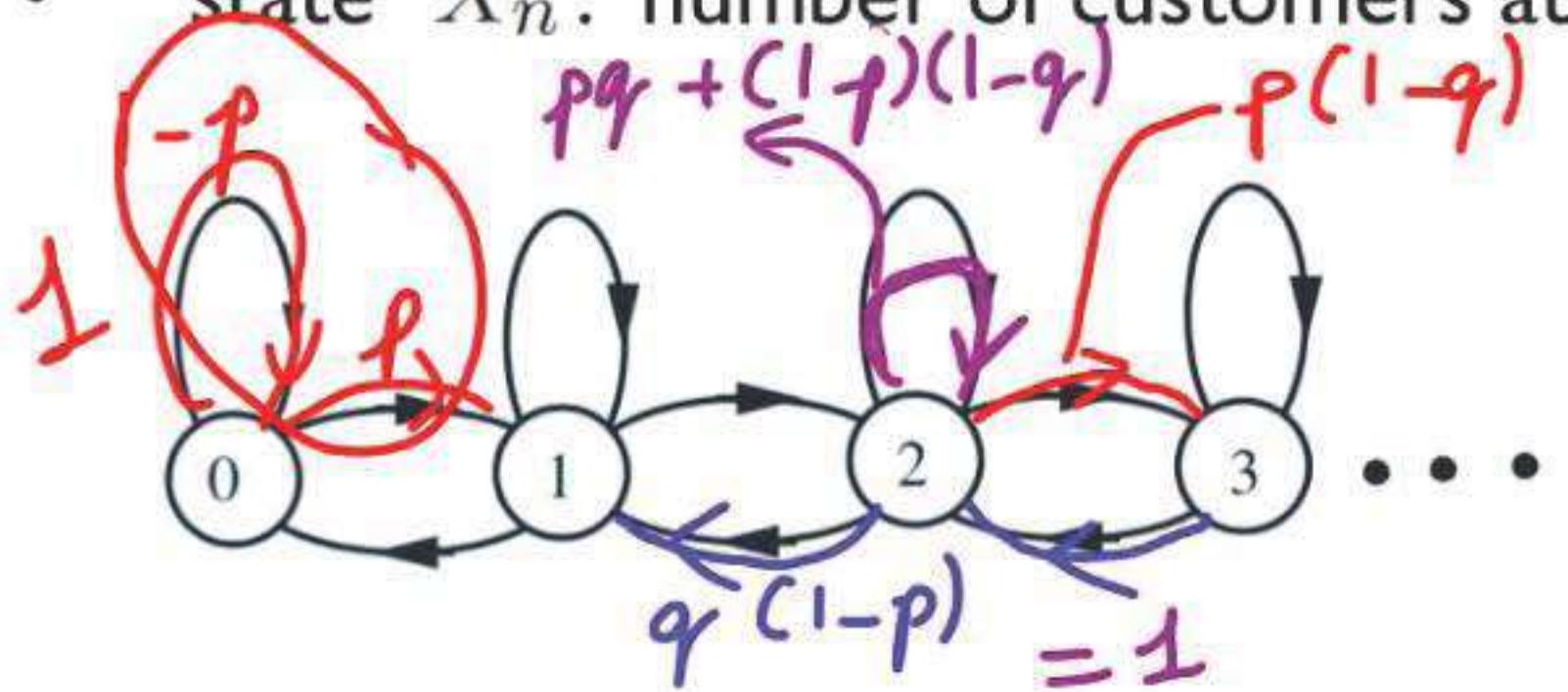
- checkout counter example ✓
- Markov process definition ✓
- n-step transition probabilities ✓
- classification of states ✓



$$\underline{\text{state}(t+1)} = f(\text{state}(t), \text{noise})$$

checkout counter example

- discrete time $n = 0, 1, \dots$
- customer arrivals: Bernoulli(p)
- customer service times: geometric(q)
- "state" X_n : number of customers at time n



$$\begin{aligned}
 X_0 &= 2 & X_1 &= 2 & X_2 &= 2+1 \\
 &&&&&= 3 \\
 X_3 &= 3 & X_4 &= 3-1 & X_5 &= 2 \\
 &&&&& X_6 = 2+1-1 \\
 &&&&&= 2, \dots \\
 &&&&&= 1 \} \text{ transition probability graph}
 \end{aligned}$$

discrete-time finite state Markov chains

- X_n : state after n transitions
 - belongs to a finite set
 - initial state X_0 either given or random
 - transition probabilities:

$$\begin{aligned} p_{ij} &= P(X_1 = j | X_0 = i) \\ &= P(X_{n+1} = j | X_n = i) \end{aligned} \quad \left. \begin{array}{l} \text{the} \\ \text{homogeneous} \\ \sum p_{ij} = 1 \end{array} \right\}$$

"at time n "



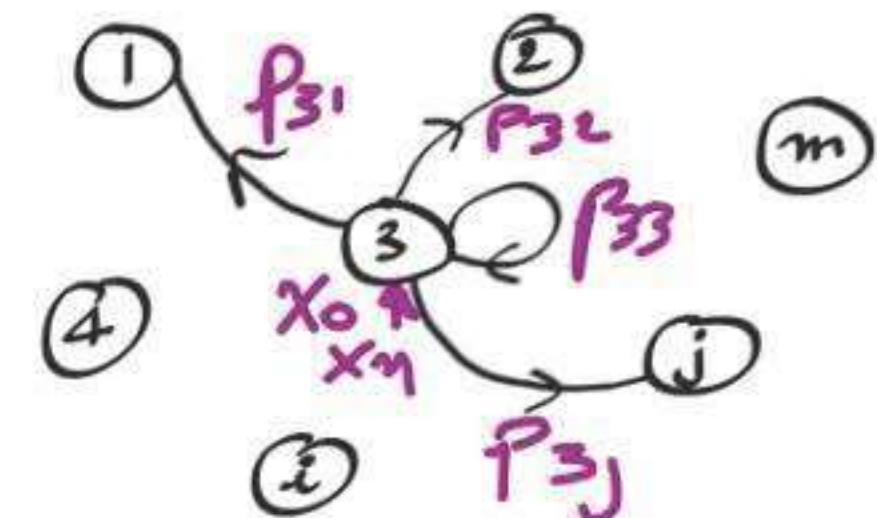
$$p_{31} + p_{32} + p_{33} + p_{3j} = 1$$

- Markov property/assumption:

"given current state, the past doesn't matter"

$$\begin{aligned} p_{ij} &\stackrel{?}{=} P(X_{n+1} = j | X_n = i) \\ &= P(X_{n+1} = j | \underline{X_n = i}, \underline{X_{n-1}, \dots, X_0}) \end{aligned}$$

- model specification: identify states, transitions, and transition probabilities

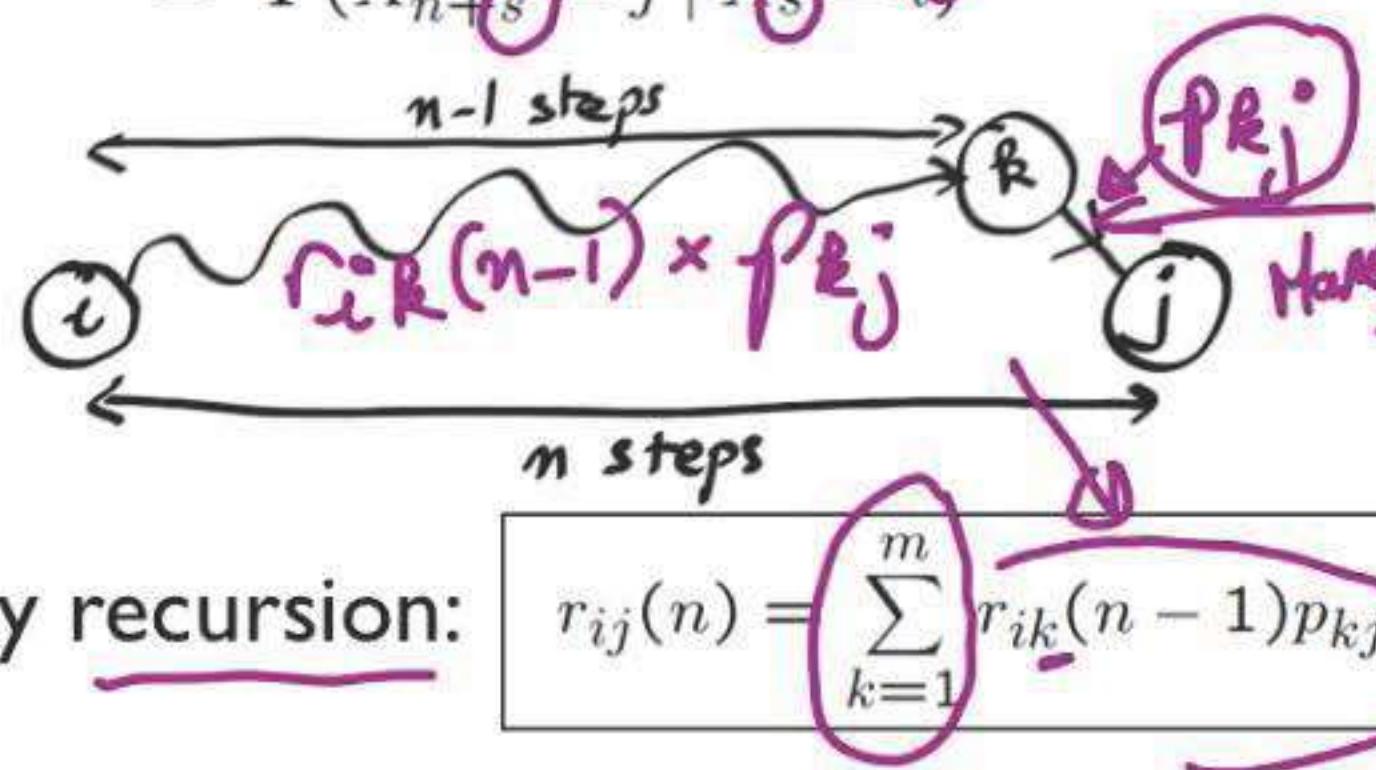


n-step transition probabilities

$$r_{ij}(0) = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases} \quad r_{ij}(1) = p_{ij} + \epsilon_i, \quad b_j$$

- state probabilities, given initial state i :

$$r_{ij}(n) = P(X_n = j \mid X_0 = i) = \sum_{k=1}^m r_{ik}(n) = 1$$

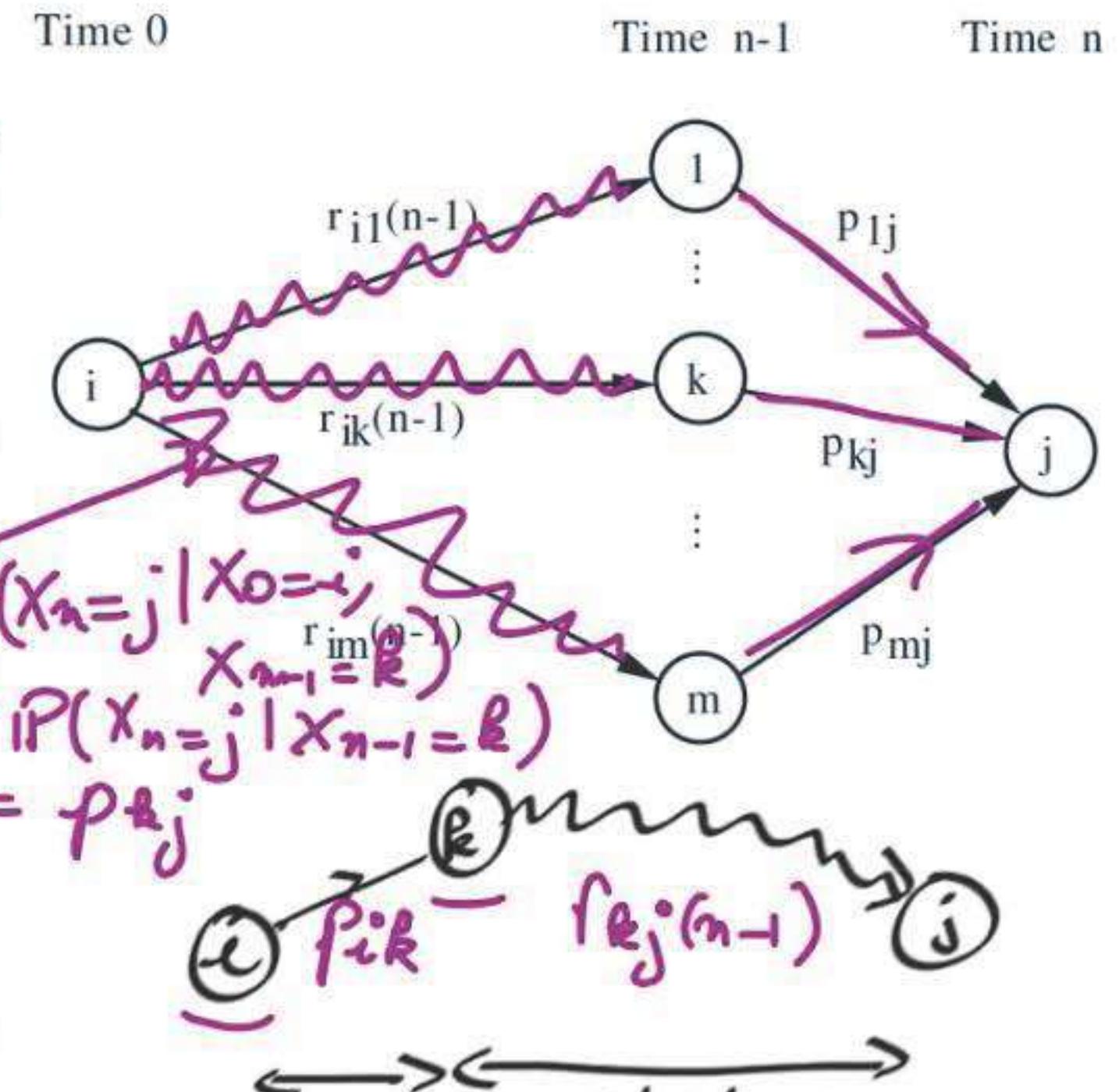


- key recursion:

$$r_{ij}(n) = \sum_{k=1}^m r_{ik}(n-1) p_{kj}$$

- random initial state:

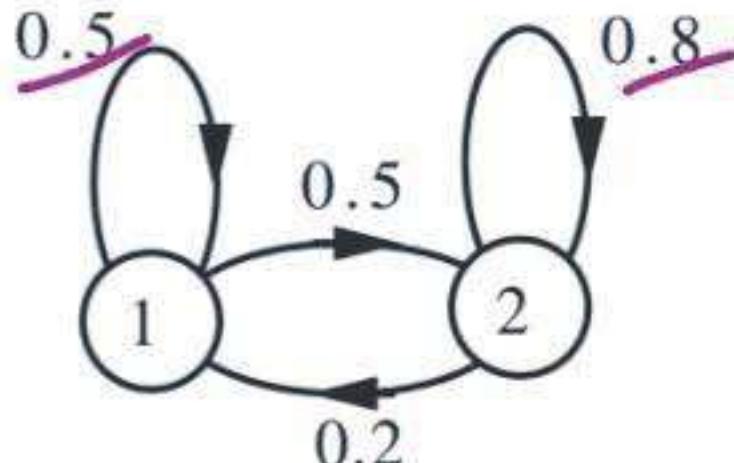
$$P(X_n = j) = \sum_{i=1}^m P(X_0 = i) \cdot p_{ij}(n)$$



$$r_{ij}^1(n) = \sum_{k=1}^{n-1} p_{ik} r_{kj}(n-1)$$

example

$$\begin{cases} r_{11}(n) = r_{11}(n-1) \times 0.5 + r_{12}(n-1) \times 0.2 \\ r_{12}(n) = 1 - r_{11}(n) \end{cases}$$



$$r_{ij}(n) = P(X_n = j \mid X_0 = i)$$

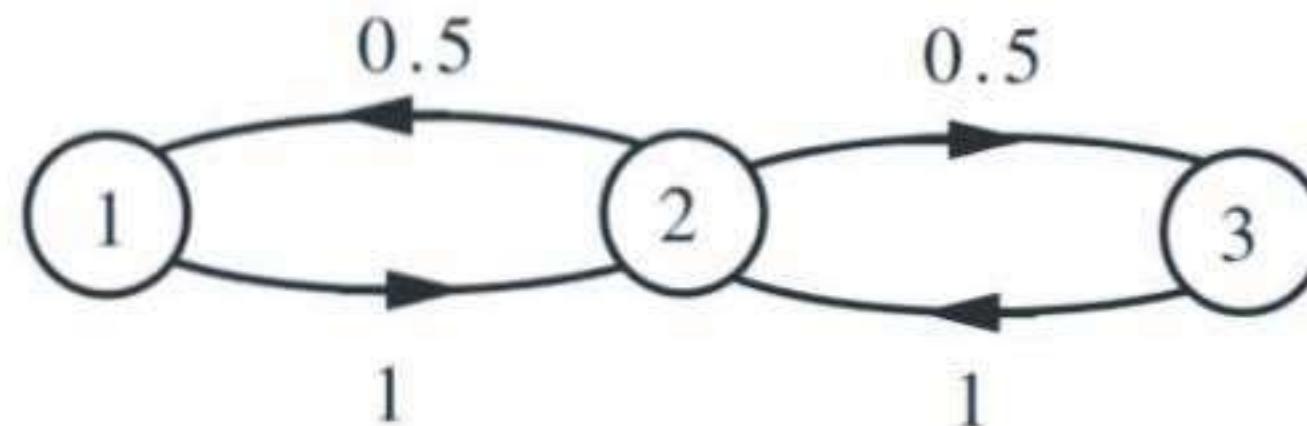
	$n = 0$	$n = 1$	$n = 2$	$n = 100$	$n = 101$
$r_{11}(n)$	1	0.5	$\xrightarrow{0.5} 0.25$ $\xrightarrow{0.2} 0.35$	$\approx 2/7$	$\approx 2/7$
$r_{12}(n)$	0	0.5	0.65	$\approx 5/7$	$\approx 5/7$
$r_{21}(n)$	0	0.2		$\approx 2/7$	
$r_{22}(n)$	1	0.8		$\approx 5/7$	

Red annotations: $r_{11}(101) = \frac{2}{7} \times 0.5 + \frac{5}{7} \times 0.2 = \frac{1}{7} + \frac{1}{7} = \frac{2}{7}$. A red bracket groups the last three columns. Red arrows show transitions from $r_{11}(1)$ to $r_{11}(2)$, and from $r_{12}(1)$ to $r_{12}(2)$.

generic convergence questions

$$r_{ij}(n) \xrightarrow{n \rightarrow \infty} \pi_j ?$$

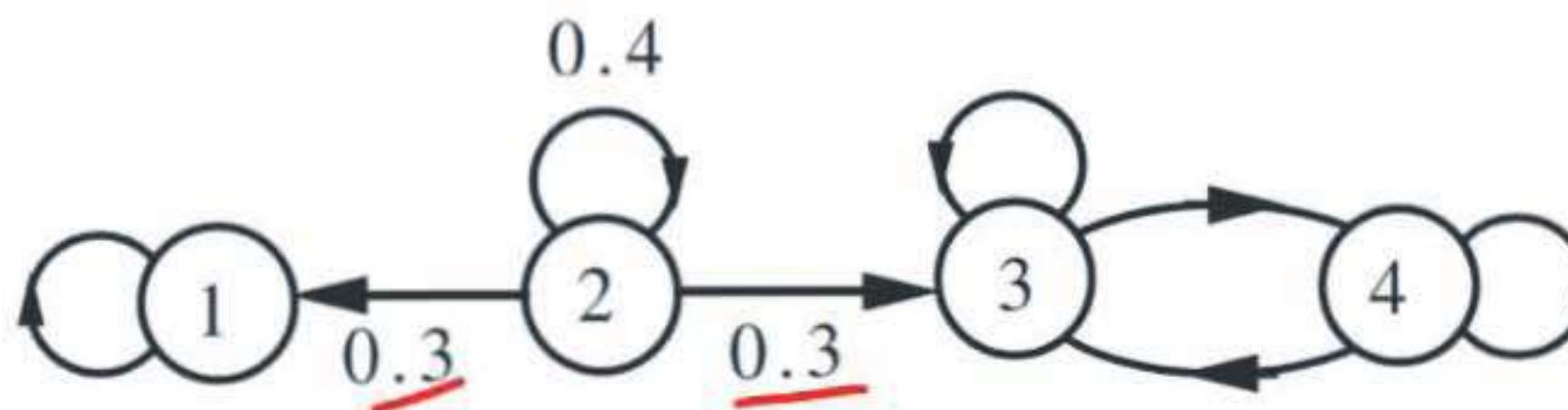
- does $r_{ij}(n)$ converge to something?



$$n \text{ odd: } r_{22}(n) = 0$$

$$n \text{ even: } r_{22}(n) = 1$$

- does the limit depend on initial state?



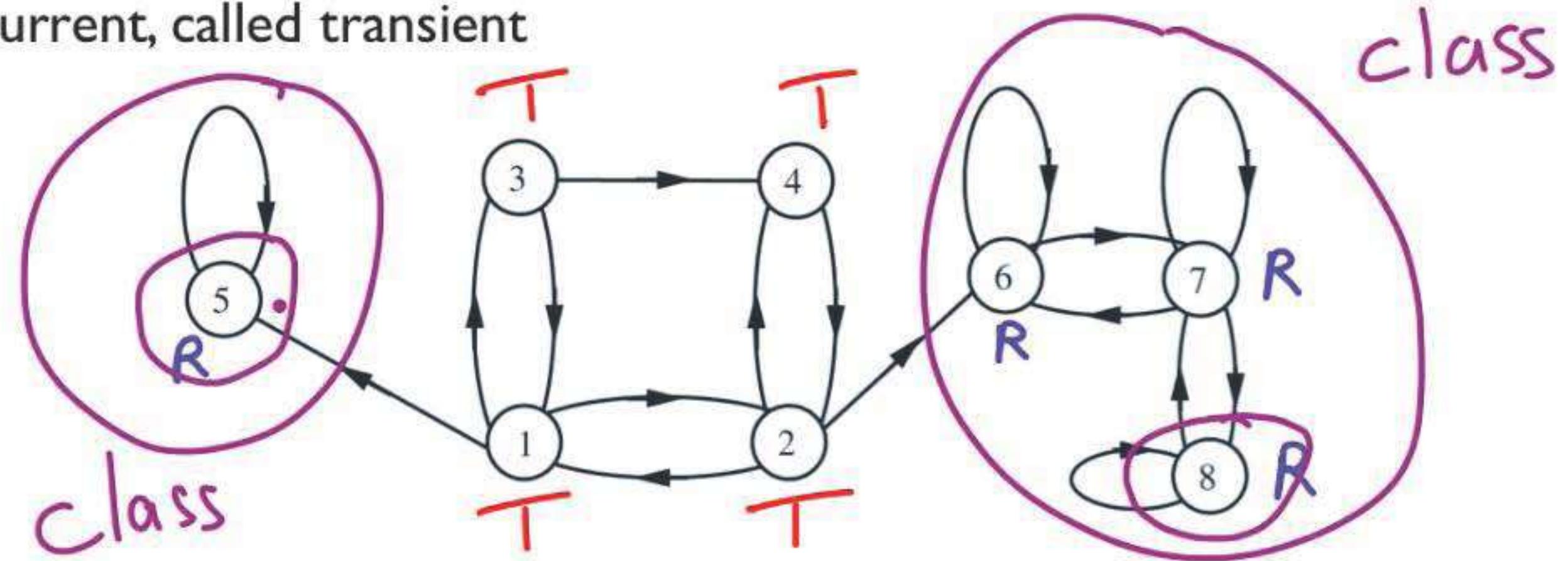
$$r_{11}(n) = 1$$

$$r_{31}(n) = 0$$

$$r_{21}(n) = \frac{1}{2}$$

recurrent and transient states

- state i is recurrent if “starting from i, and from wherever you can go, there is a way of returning to i”
- if not recurrent, called transient



MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

Markov processes – II

- review and some warm-up
 - definitions, Markov property
 - calculating the probabilities of trajectories
- steady-state behavior
 - recurrent states, transient states, recurrent classes
 - periodic states
 - convergence theorem
 - balance equations
- birth-death processes

review

- discrete time, discrete state space, **time-homogeneous**

– transition probabilities $p_{ij} \triangleq P(X_{s+1}=j | X_s=i) \leftarrow \text{for } s=0, 1, 2, \dots$

– Markov property $P(X_{s+1}=j | X_s=i, X_0=i_0, \dots, X_{s-1}=i_{s-1})$

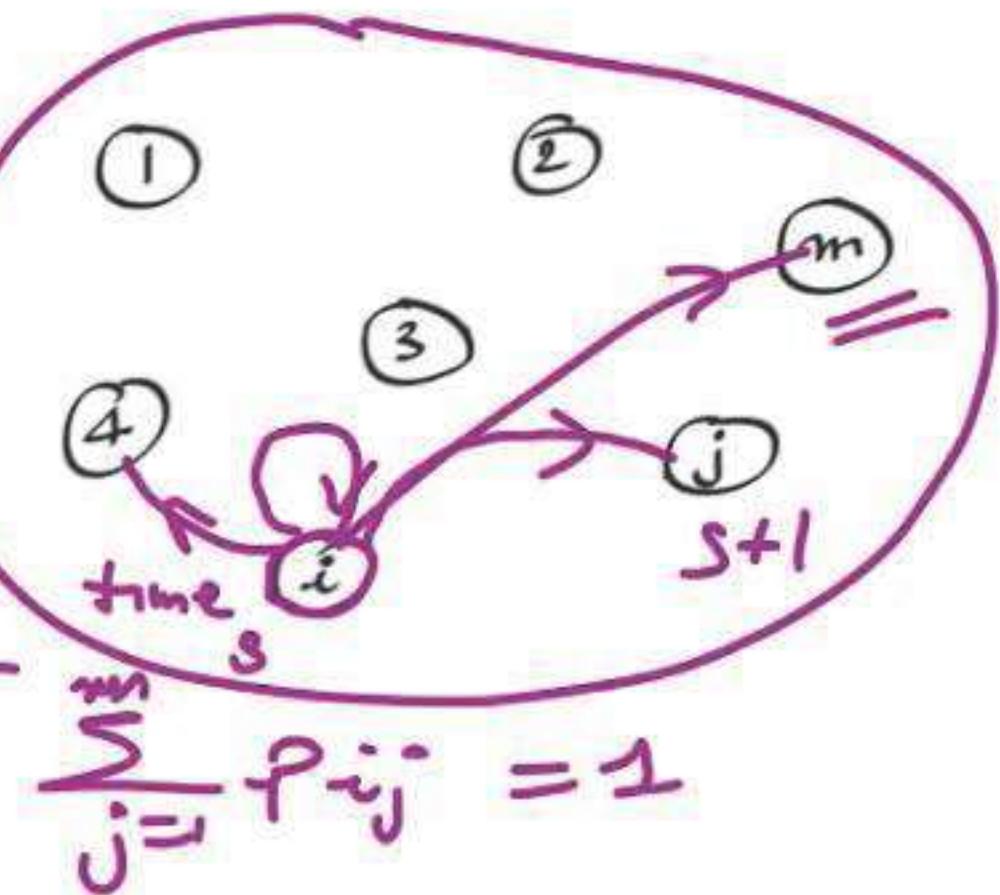
$$= p_{ij}$$

$$\begin{aligned} r_{ij}(n) &= P(X_n=j | X_0=i) \\ &= P(X_{n+s}=j | X_s=i) \end{aligned}$$

$$n=1 \quad r_{ij}(1) = p_{ij}, \quad n \geq 2$$

- key recursion:

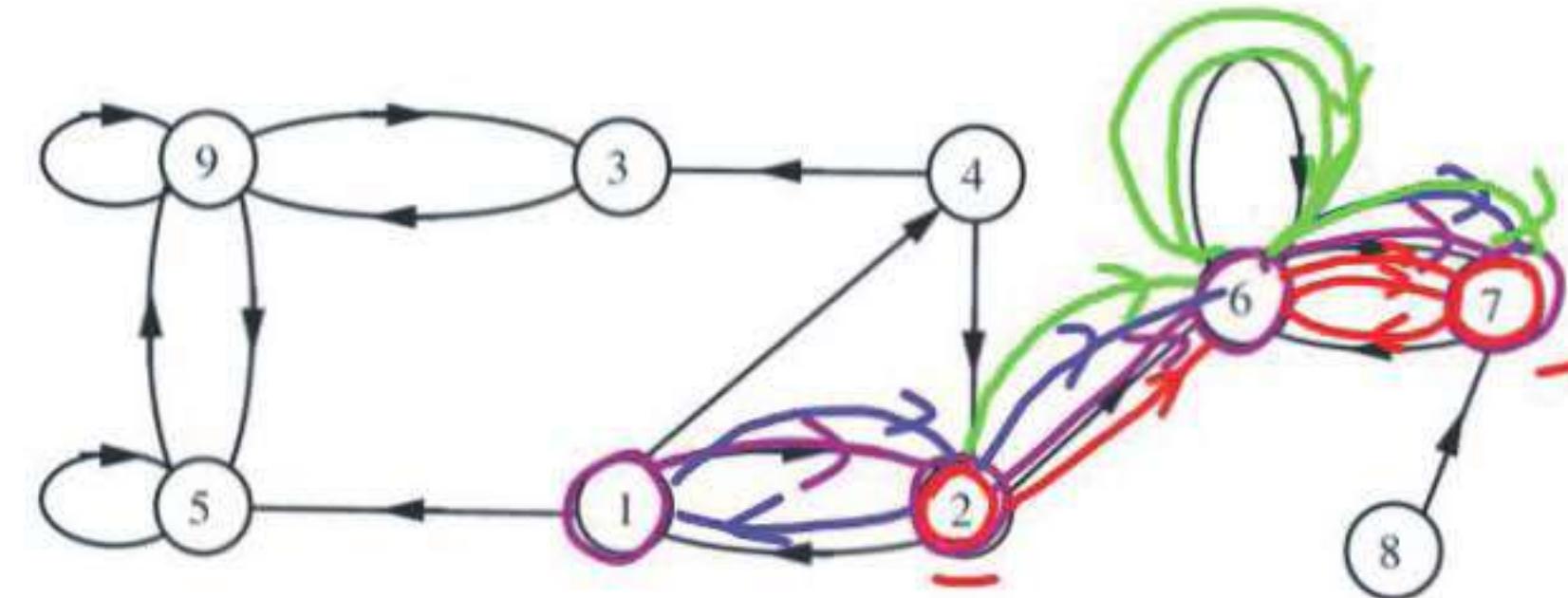
$$r_{ij}(n) = \sum_{k=1}^m r_{ik}(n-1) p_{kj}$$



warmup

"multiplicative rule"

$$\begin{aligned} \text{IP}(B \wedge C \wedge D | A) &= \\ \text{IP}(B | A) \times \text{IP}(C | A \wedge B) & \\ \times \text{IP}(D | A \wedge B \wedge C) & \end{aligned}$$



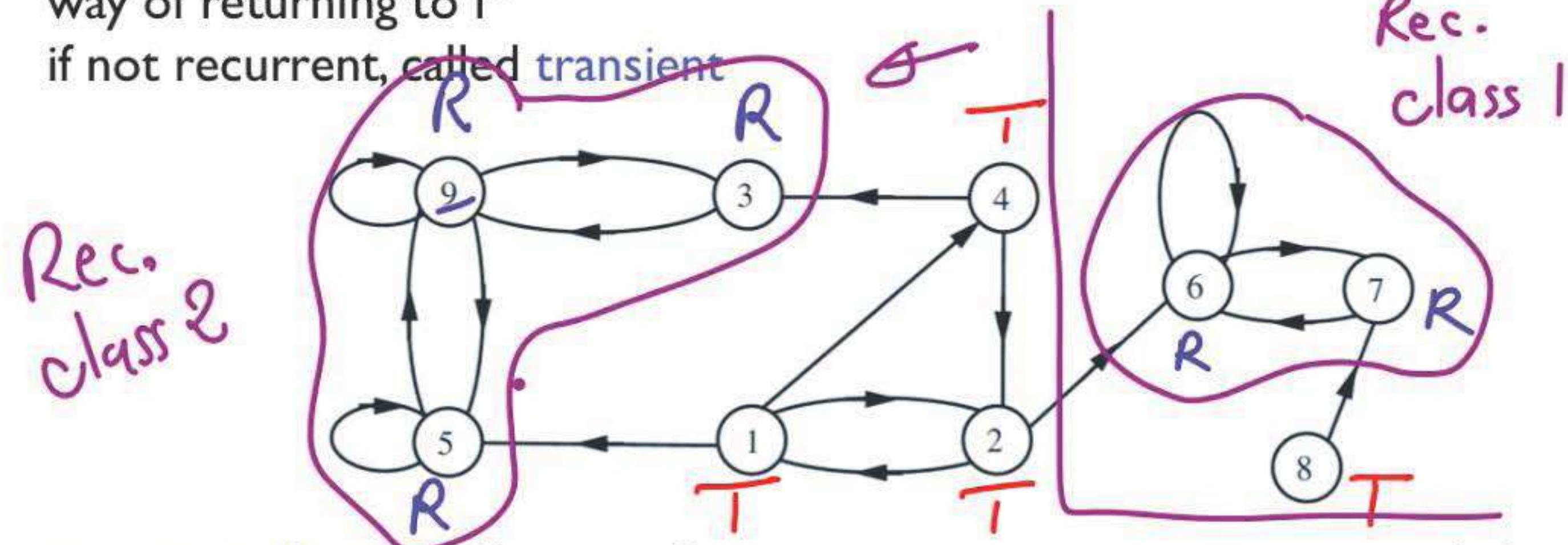
$$P(X_1 = 2, X_2 = 6, X_3 = 7 | X_0 = 1) =$$

$$\underbrace{\text{IP}(X_1=2 | X_0=1)}_{P_{12}} \times \underbrace{\text{IP}(X_2=6 | X_0=1, X_1=2)}_{P_{26}} \times \underbrace{\text{IP}(X_3=7 | X_0=1, X_1=2, X_2=6)}_{P_{67}}$$

$$\begin{aligned} P(X_4=7 | X_0=2) &= \underbrace{P_{26} \cdot P_{67} \cdot P_{76} \cdot P_{67}}_{\text{m}} + \underbrace{P_{21} \cdot P_{12} \cdot P_{26} \cdot P_{67}}_{\text{m}^2} \\ &\quad + \underbrace{P_{26} \cdot P_{67} \cdot P_{76} \cdot P_{67}}_{\text{n} \times \text{m}} \end{aligned}$$

review: recurrent and transient states

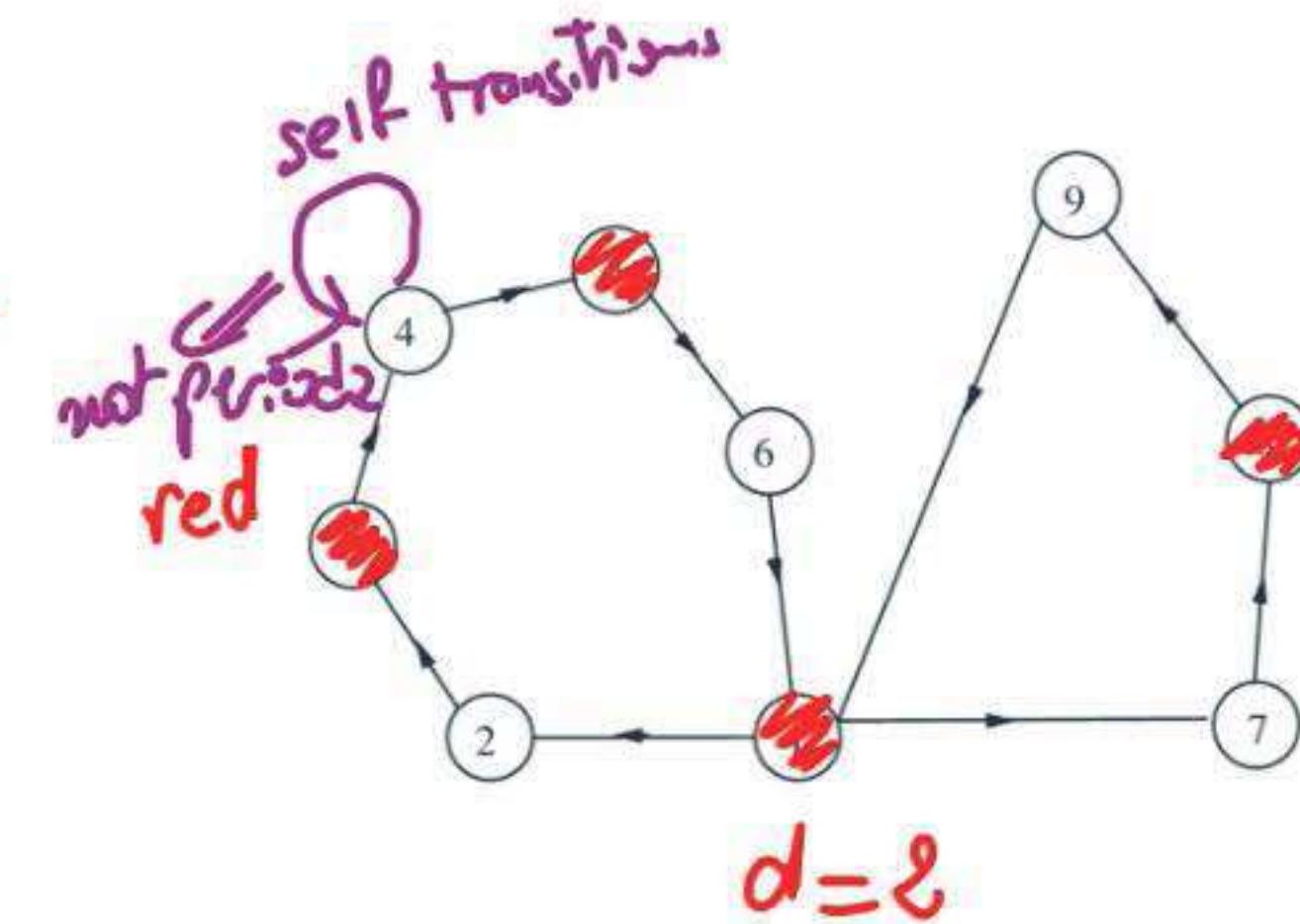
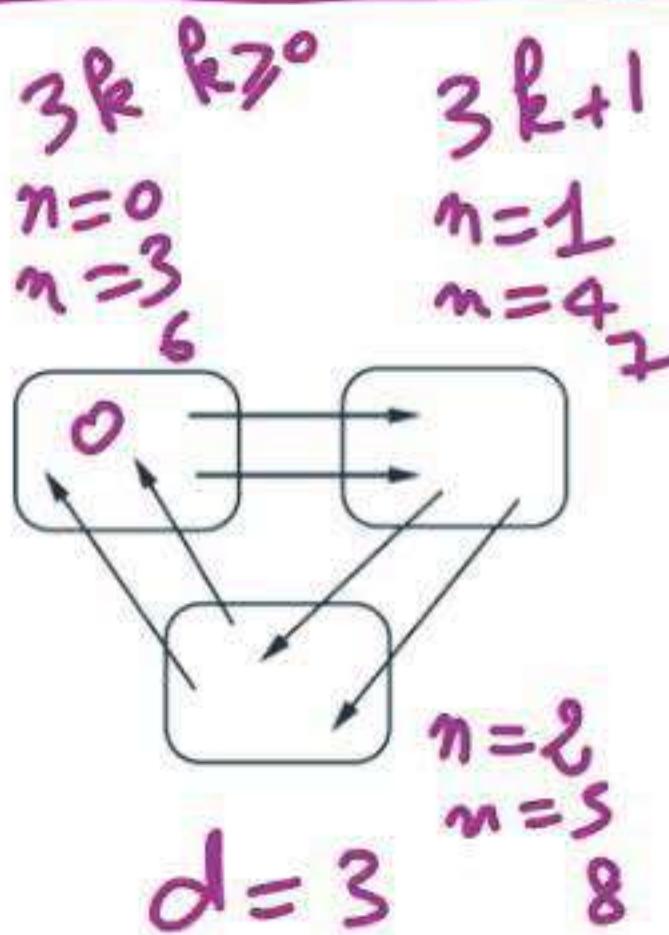
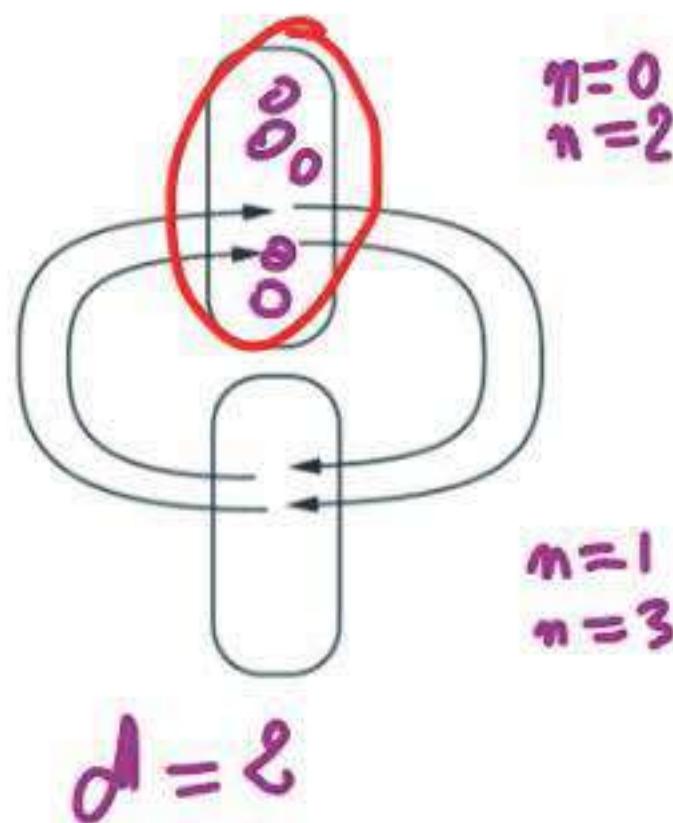
- state i is **recurrent** if “starting from i , and from wherever you can go, there is a way of returning to i ”
- if not recurrent, called **transient**



- **recurrent class:** a collection of recurrent states communicating only between each other

periodic states in a recurrent class

The states in a recurrent class are periodic if they can be grouped into $d > 1$ groups so that all transitions from one group lead to the next group

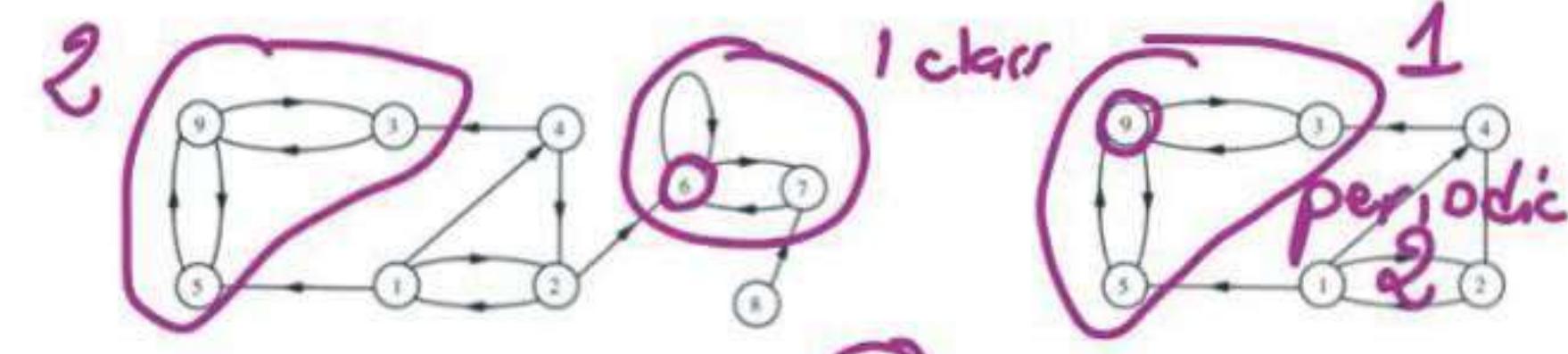


steady-state probabilities

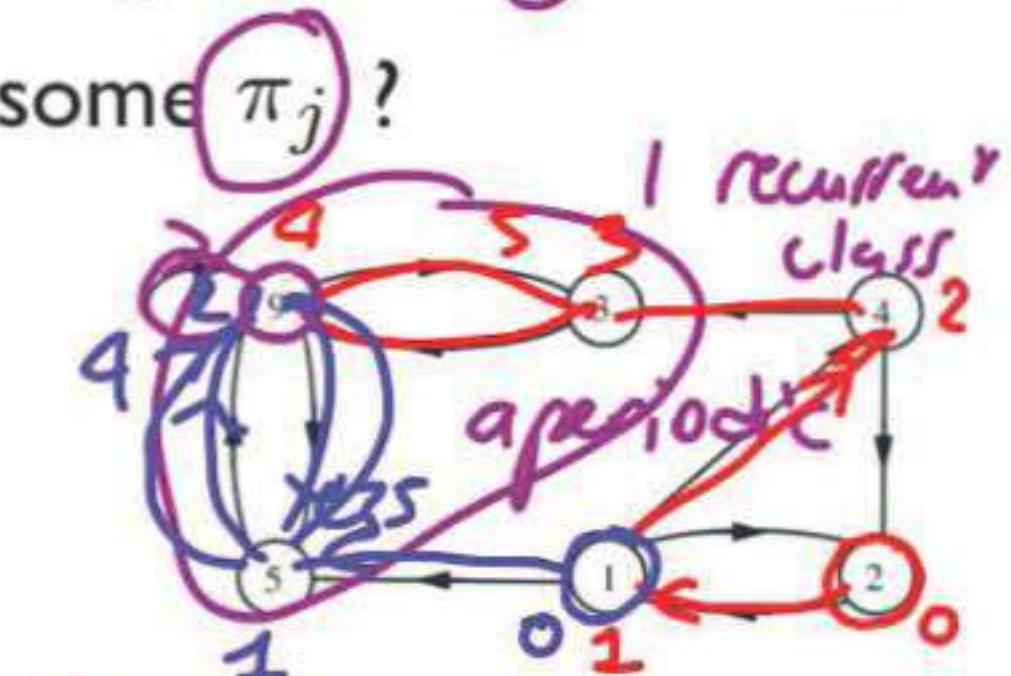
steady-state pi
- convergence

- converges
- is independent of i

- does $r_{ij}(n) = P(\underbrace{X_n=j}_{\text{independent}} \mid \underbrace{X_0=i}_{\text{initial}})$ converge to some π_j ?
 - theorem: yes, if:



- recurrent states are all in a single class, and
 - single recurrent class is not periodic



- assuming “yes”, start from key recursion $r_{ij}(n) = \sum r_{ik}(n-1)p_{kj}$

- take the limit as $n \rightarrow \infty$

– need also:

$$\sum_{j=1} \pi_j = 1$$

and
unique sol?

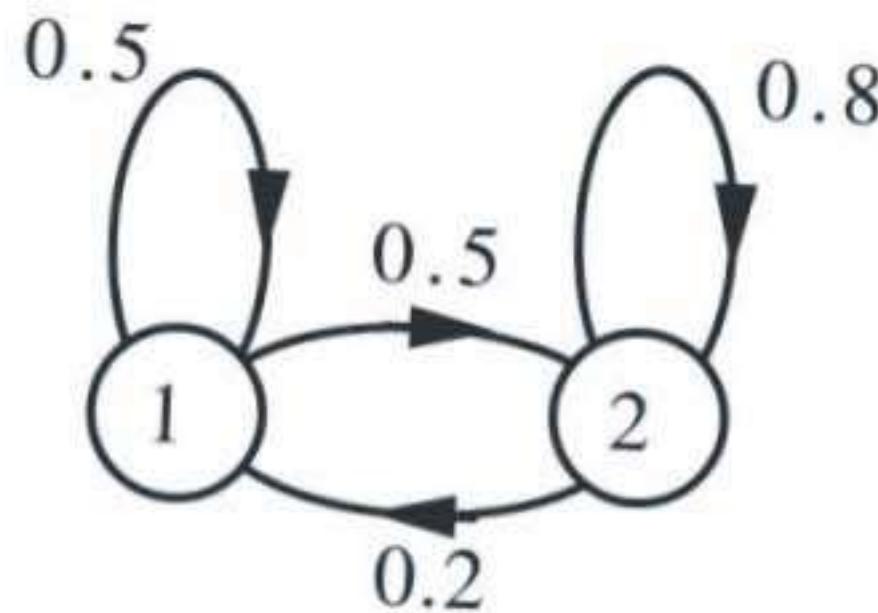
$$r_{ij}(n) = \sum_{k=1}^m r_{ik}(n-1)p_{kj}$$

for $j=1, \dots, m$

| m equations
| m unknowns

~~$\pi_i = \pi_j$~~

example



- single recurrent class
- not periodic

$$\left\{ \begin{array}{l} \pi_j = \sum_{k=1}^m \pi_k p_{kj}, \quad j=1, \dots, m \\ m=2 \end{array} \right.$$

$$\left\{ \begin{array}{l} \overline{\pi}_1 = \overline{\pi}_1 \times 0.5 + \overline{\pi}_2 \times 0.2 \\ \overline{\pi}_2 = \overline{\pi}_1 \times 0.5 + \overline{\pi}_2 \times 0.8 \end{array} \right. \quad \left\{ \begin{array}{l} \overline{\pi}_1 \times 0.5 = \overline{\pi}_2 \times 0.2 \\ \overline{\pi}_2 \times 0.2 = \overline{\pi}_1 \times 0.5 \end{array} \right. \xrightarrow{\text{same}}$$

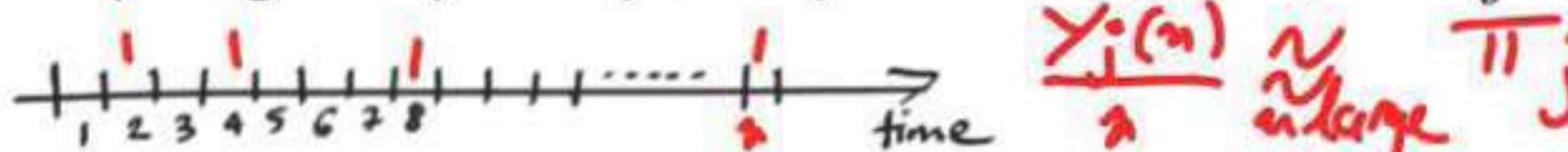
$$\left\{ \begin{array}{l} \overline{\pi}_1 \times \frac{1}{2} = \overline{\pi}_2 \times \frac{1}{5} \\ \overline{\pi}_1 + \overline{\pi}_2 = 1 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \overline{\pi}_1 = \frac{2}{5} \overline{\pi}_2 \\ (\overline{\pi}_2 \left(\frac{2}{5} + 1 \right)) = 1 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \overline{\pi}_1 = \frac{2}{5} \times \frac{5}{7} = \frac{2}{7} \\ \overline{\pi}_2 = \frac{5}{7} \end{array} \right.$$

visit frequency interpretation

- balance equations

$$\pi_j = \sum_k \pi_k p_{kj}$$

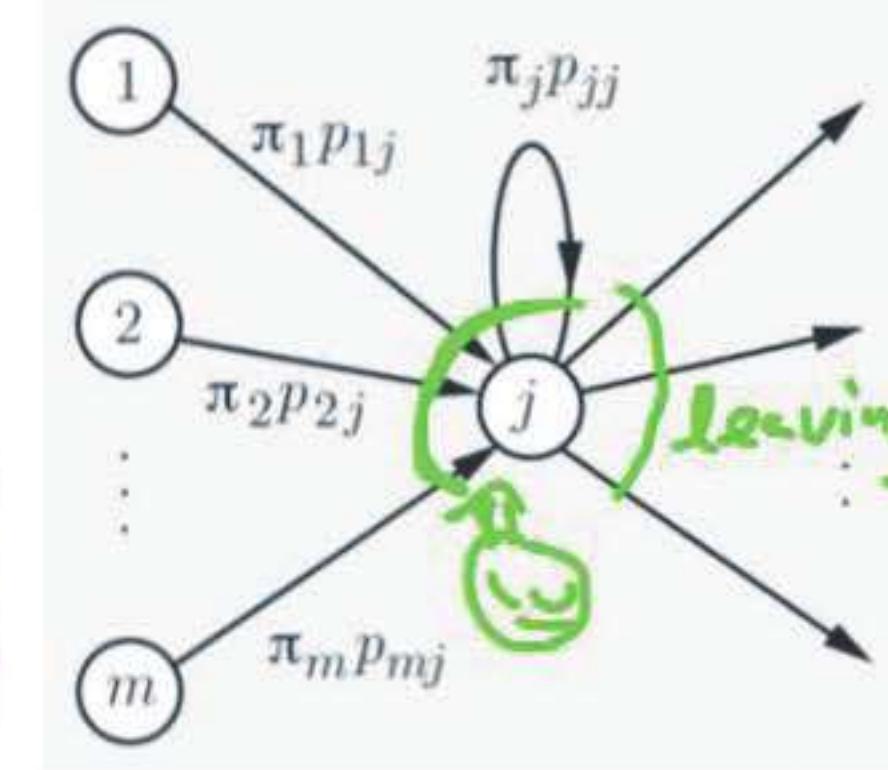
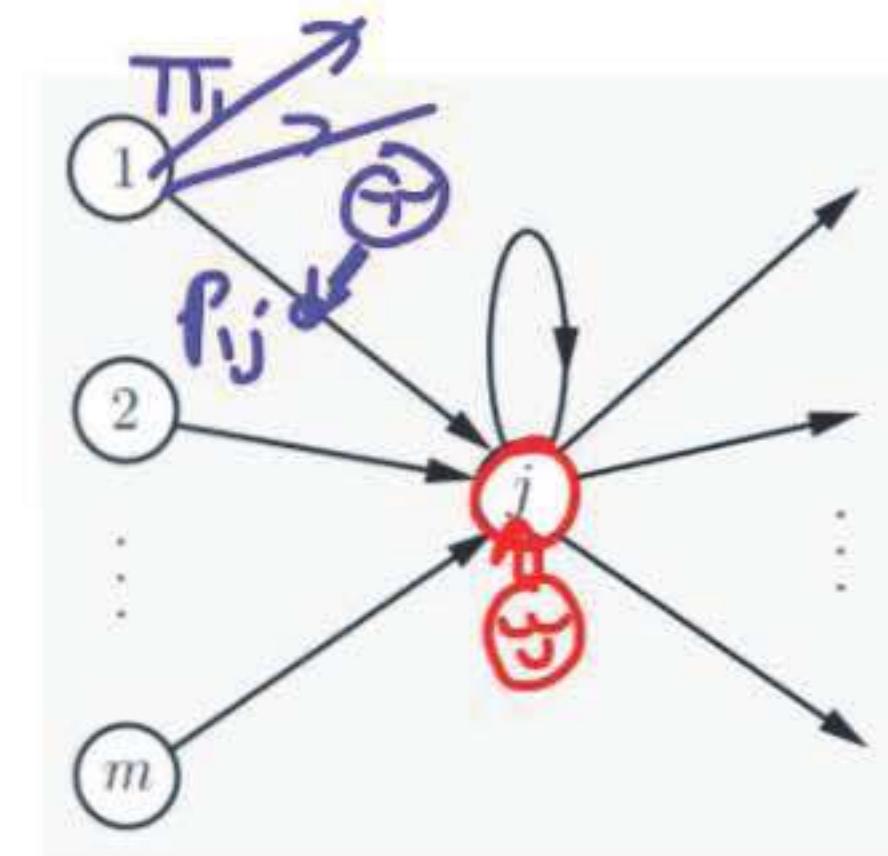
- (long run) frequency of being in j : π_j



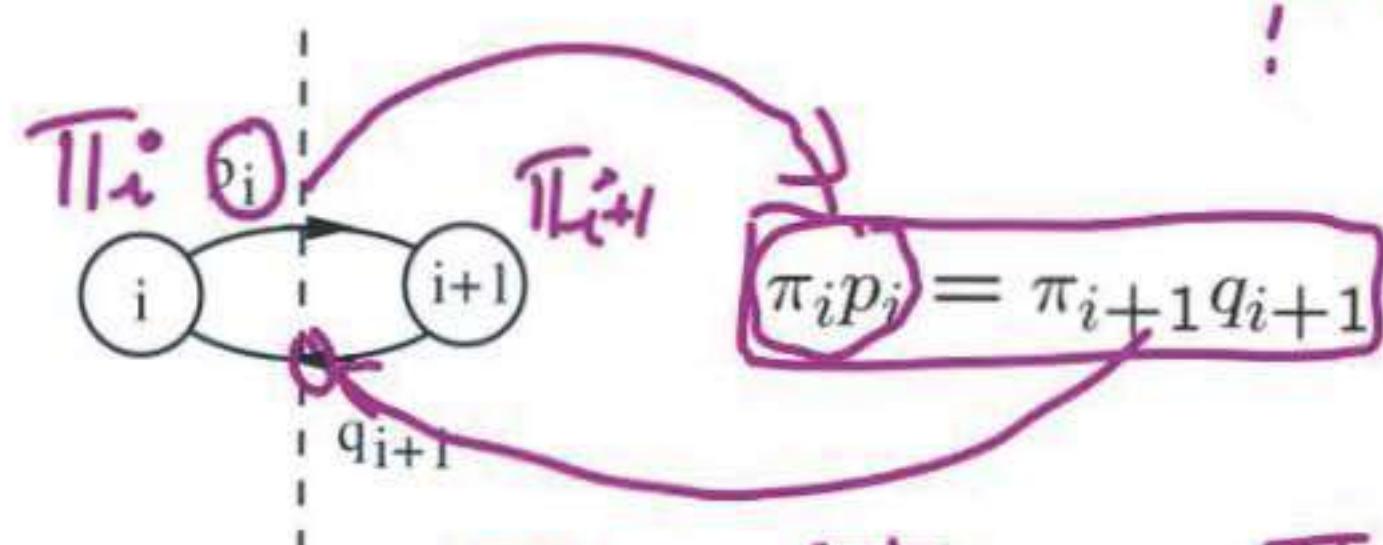
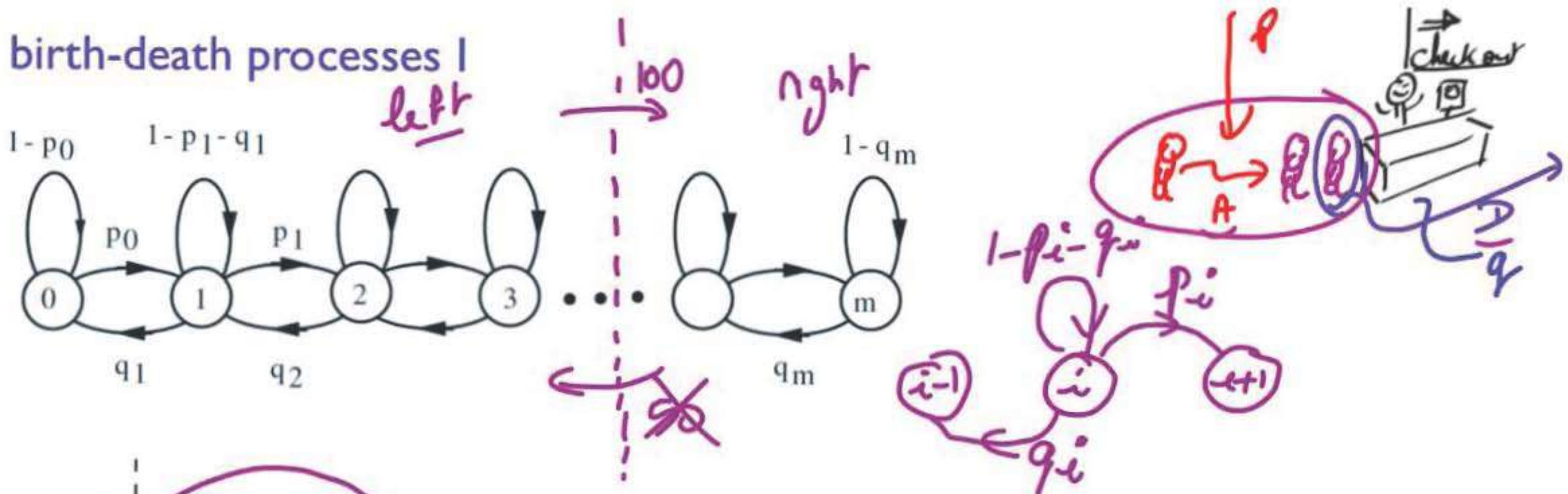
- frequency of transitions $1 \rightarrow j$: $\pi_1 p_{1j}$



- frequency of transitions into j : $\sum_k \pi_k p_{kj}$



birth-death processes I

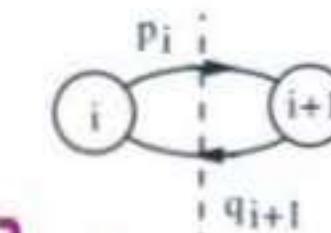
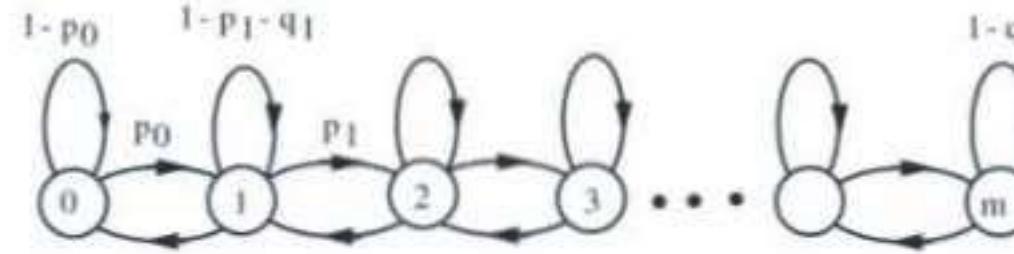


$$\pi_{i+1} = \pi_i \times \frac{p_i}{q_{i+1}} \quad i = 0, 1, \dots$$

$\pi_0 \rightarrow \pi_1, \rightarrow \pi_2, \rightarrow \pi_3 / \pi_0 ?$

$$\sum_j \pi_j = 1 \Rightarrow \pi_0 + \pi_0 \times \frac{p_0}{q_1} + \pi_0 \times \frac{p_0 \times p_1}{q_1 q_2} + \dots = 1$$

birth-death processes II



$$\pi_i p_i = \pi_{i+1} q_{i+1}$$

$$\sum_j \pi_j = 1$$

$$\pi_{i+1} = \pi_i \times \frac{p_i}{q_{i+1}}, \quad i=0, \dots, m \quad \pi_0 \left[1 + \frac{p_0}{q_1} + \dots \right] = 1$$

special case: $p_i = p$ and $q_i = q$ for all i

$$\rho = p/q \quad \pi_{i+1} = \pi_i \frac{p}{q} = \pi_i \rho \quad \pi_1 = \pi_0 \rho, \quad \pi_2 = \pi_1 \rho = \pi_0 \rho^2, \quad \dots$$

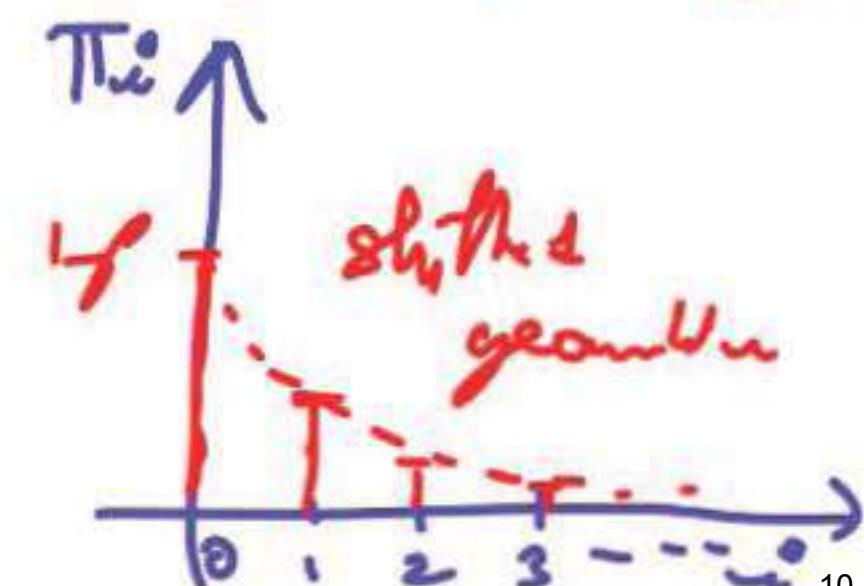
$$\pi_i = \pi_0 \rho^i \quad i = 0, 1, \dots, m \quad \sum_j \pi_j = 1 \Rightarrow \pi_0 \left[1 + \rho + \rho^2 + \dots + \rho^m \right] = 1$$

- assume $p = q \Rightarrow \pi_i = \pi_0 \quad i = 0, \dots, m \quad \pi_0 [1+m] = 1, \quad \pi_0 = \frac{1}{1+m}$

- assume $p < q$ and $m \approx \infty \Rightarrow \sum_{i=1}^{\infty} \rho^i = \frac{1}{1-\rho}$

$\boxed{\pi_0 = 1 - \rho}$ $\boxed{E[X_n] = \frac{\rho}{1-\rho}}$ (in steady-state)

$\Rightarrow \pi_i = \pi_0 \rho^i = (1-\rho) \rho^i, \quad i=0, \dots$



MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

Markov processes – III

- review of steady-state behavior
- probability of blocked phone calls
- calculating absorption probabilities
- calculating expected time to absorption



review of steady state behavior

- Markov chain with a single class of recurrent states, aperiodic; and some transient states; then,

$$\lim_{n \rightarrow \infty} r_{ij}(n) = \lim_{n \rightarrow \infty} P(X_n = j | X_0 = i) = \pi_j, \quad \forall i$$

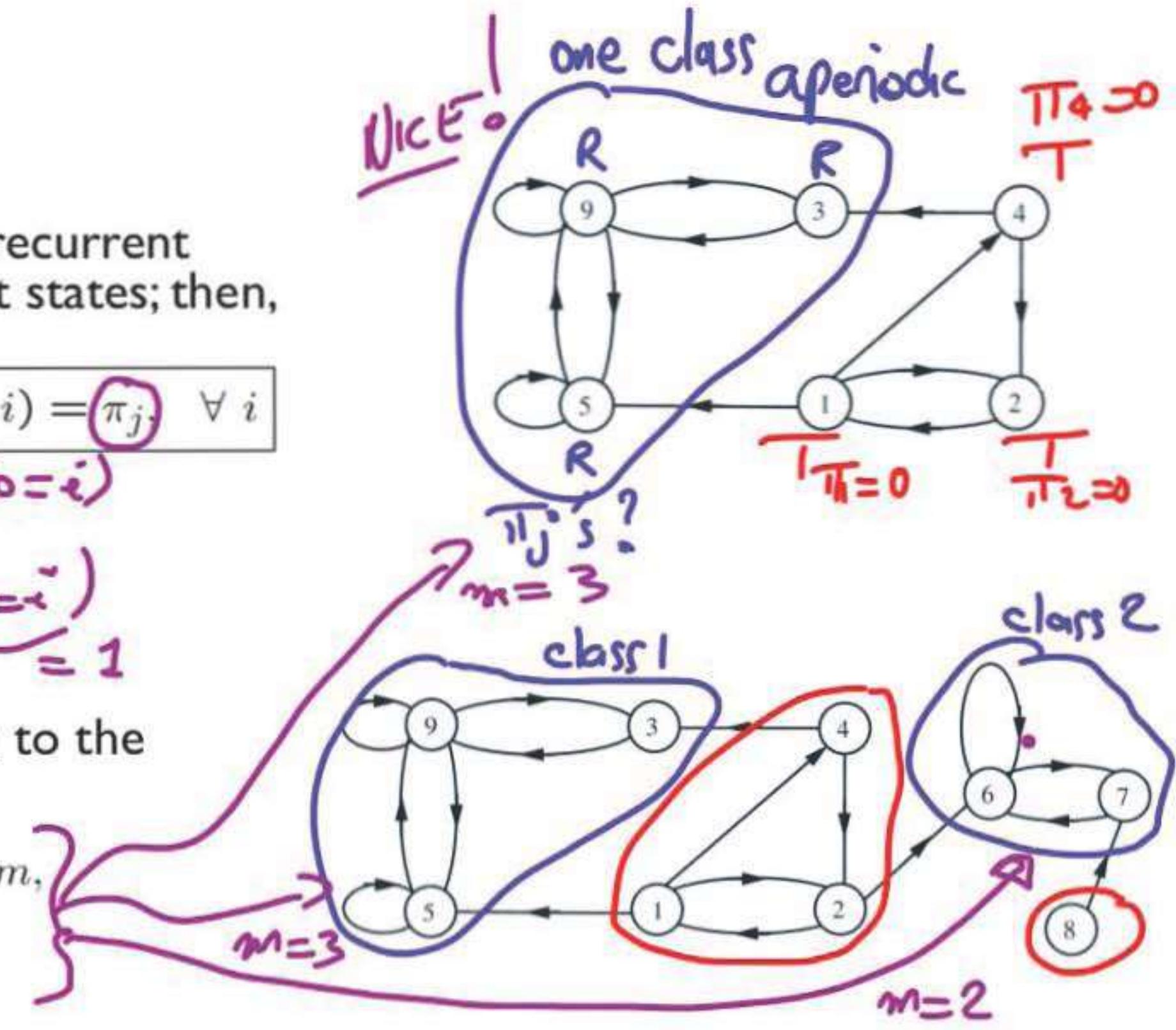
$$P(X_n = j) = \sum_i r_{ij}(n) \times P(X_0 = i)$$

$\pi_j = \sum_i P(X_0 = i) = 1$

- can be found as the unique solution to the balance equations

$$\pi_j = \sum_k \pi_k p_{kj}, \quad j = 1, \dots, m,$$

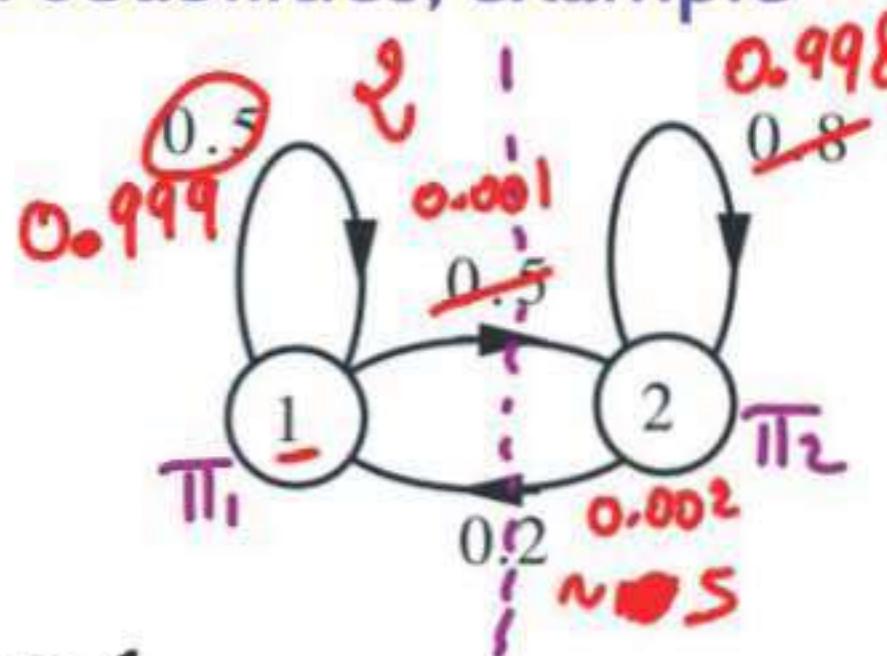
- together with $\sum_j \pi_j = 1$



on the use of steady state probabilities, example

$$\begin{cases} \pi_1 \times 0.5 = \pi_2 \times 0.2 \\ \pi_1 + \pi_2 = 1 \end{cases}$$

$$\pi_1 = 2/7, \pi_2 = 5/7$$



assume process starts in state 1

$$P(X_1 = 1 \text{ and } X_{100} = 1 | X_0 = 1) =$$

$$P(X_1 = 1 | X_0 = 1) \times P(X_{100} = 1 | X_1 = 1, X_2 \neq 1) \\ p_{11} \times \pi_{11}(99) \approx p_{11} \times \pi_1 = 0.5 \times \frac{2}{7}$$

$$P(X_{100} = 1 \text{ and } X_{101} = 2 | X_0 = 1) =$$

$$P(X_{100} = 1 | X_0 = 1) \times P(X_{101} = 2 | X_{100} = 1, X_0 \neq 1) \\ \pi_{11}(100) \times p_{12} \approx \pi_1 \times p_{12} = \frac{2}{7} \times 0.5$$

$$P(X_{100} = 1 \text{ and } X_{200} = 1 | X_0 = 1) =$$

$$P(X_{100} = 1 | X_0 = 1) \times P(X_{200} = 1 | X_{100} = 1, X_0 \neq 1) \\ = \pi_{11}(100) \times \pi_{11}(100) \approx \pi_1 \times \pi_1 = \pi_1^2 = \left(\frac{2}{7}\right)^2$$

is $n=99, 100$ large enough?

Simulation

$$f_{11}(n)$$

$$1$$

$$0.5$$

$$2/7$$

$$0$$

$$1$$

$$5$$

$$n$$

exponential
decrease

$n=5$ 2 correct decimal

$n=10$ correct up to 5 decimal

2) order of magnitude

3) by theory

$$(c)^n \quad 0 < c < 1$$

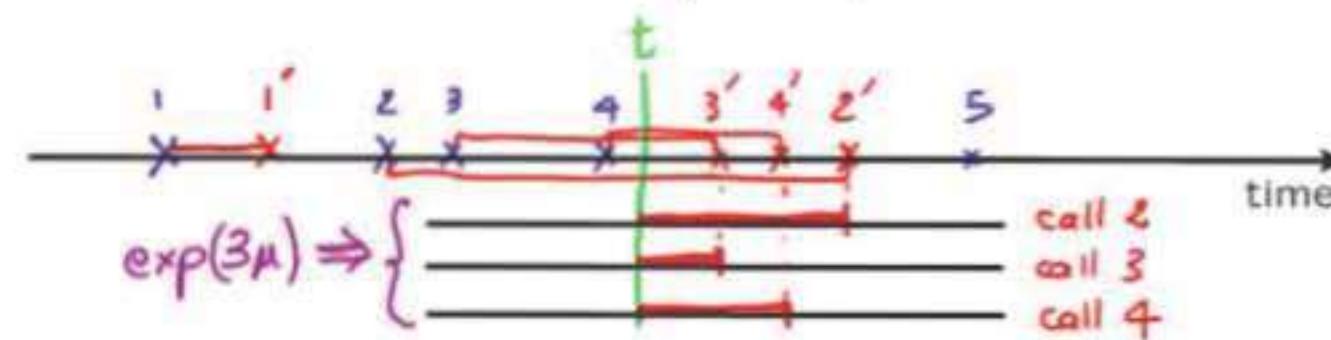
$$-c = 0.3$$

$$-c = 0.997$$

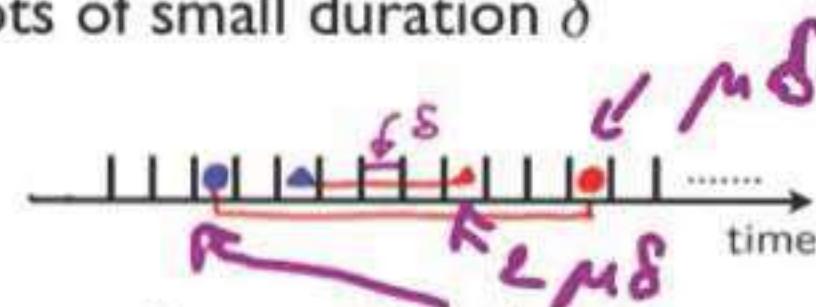
design of a phone system

(Erlang)

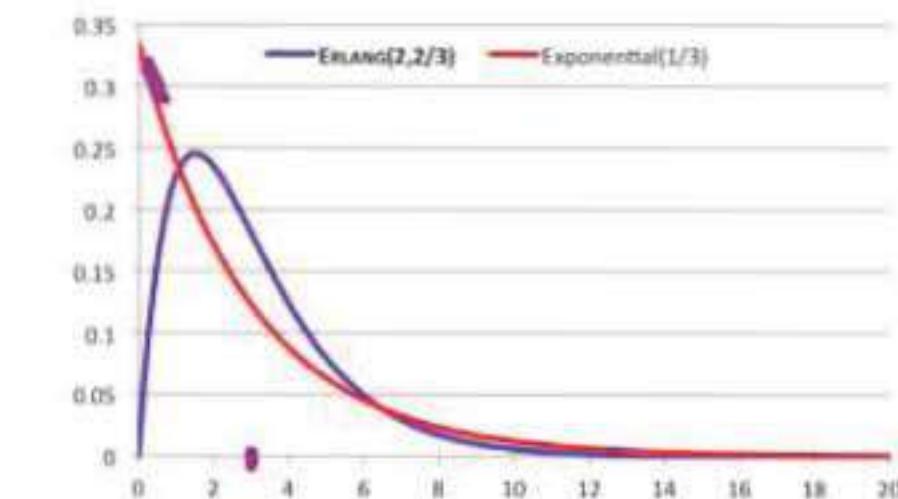
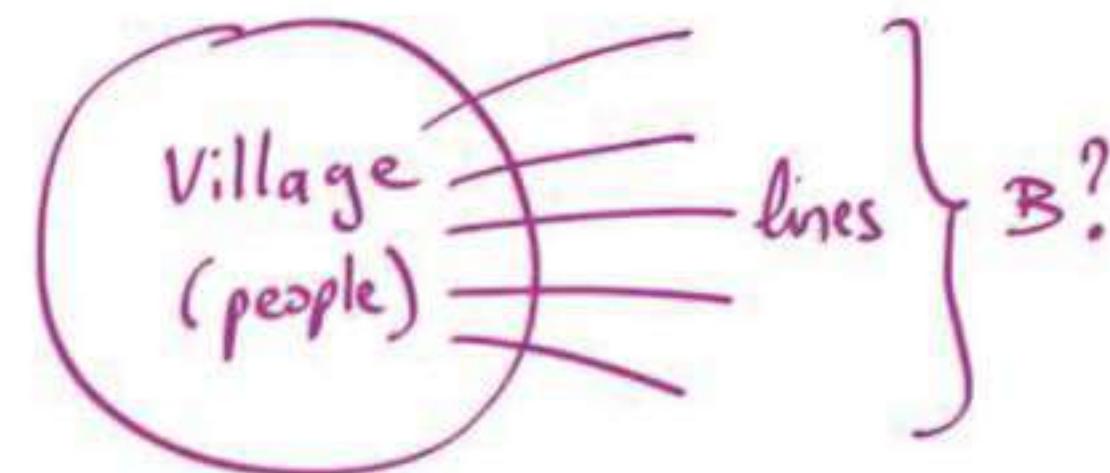
- calls originate as a Poisson process, rate λ
- each call duration is exponential (parameter μ)
- need to decide on how many lines, B ?



- for time slots of small duration δ



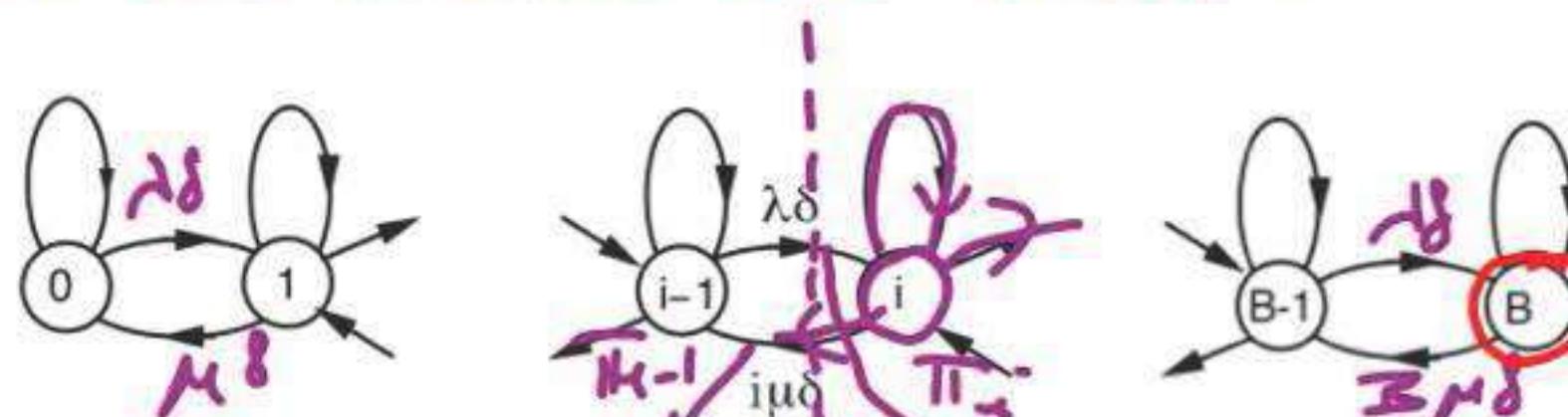
- $P(\text{a new call arrives}) \approx \lambda\delta$
- if you have i active calls, then $P(\text{a departure}) \approx i\mu\delta$



design of a phone system, a discrete time approximation

- approximation: discrete time slots of (small) duration δ

$$P(1 \text{ new call}) \approx \lambda \delta ; P(1 \text{ call ends} | i \text{ busy}) \approx i \mu \delta$$



$$\pi_i (\rightarrow \cancel{\pi_B}) \Rightarrow \pi_{i-1} \times \lambda \delta$$

$$\pi_0 \rightarrow \pi_1 \rightarrow \pi_2 \dots$$

$$\sum_i \pi_i = 1 \rightarrow \pi_0 = 1 / \sum_{i=0}^B \frac{\lambda^i}{\mu^i i!} \Rightarrow \pi_i = f(B, \lambda, \mu)$$

- balance equations

$$\lambda \pi_{i-1} = i \mu \pi_i$$

$$\pi_i = \pi_0 \frac{\lambda^i}{\mu^i i!}$$

- P(arriving customer finds busy system) is $\boxed{\pi_B}$

$$\pi_B \leq p_0 \Rightarrow B \geq \underline{106}$$



$$\lambda = 30 \text{ calls / minute}$$

$$\mu = \gamma_3$$

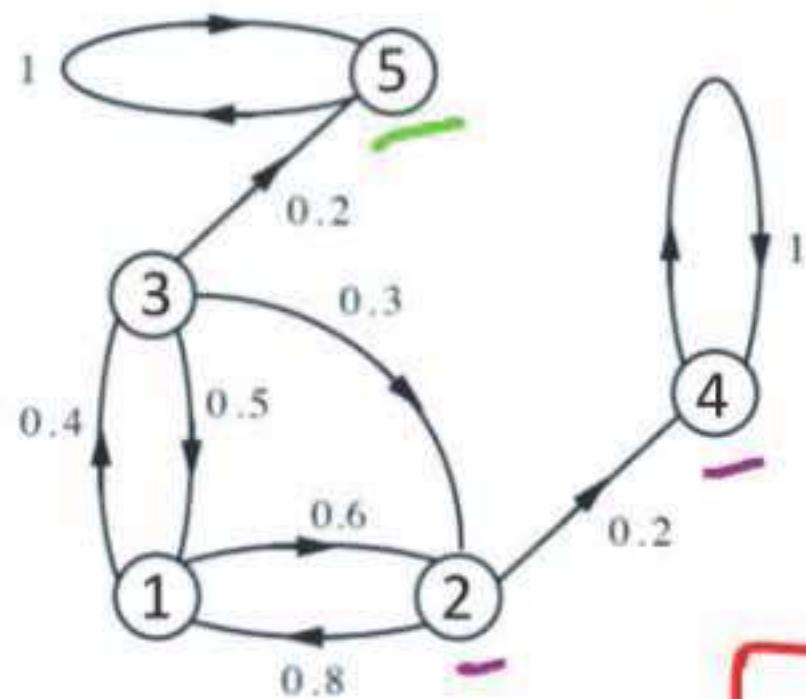
3 minutes

90 calls

$$B = 90 ?$$

calculating absorption probabilities

- absorbing state: recurrent state k with $p_{kk} = 1$
- what is the probability a_i that the chain eventually settles in 4 given it started in i ?



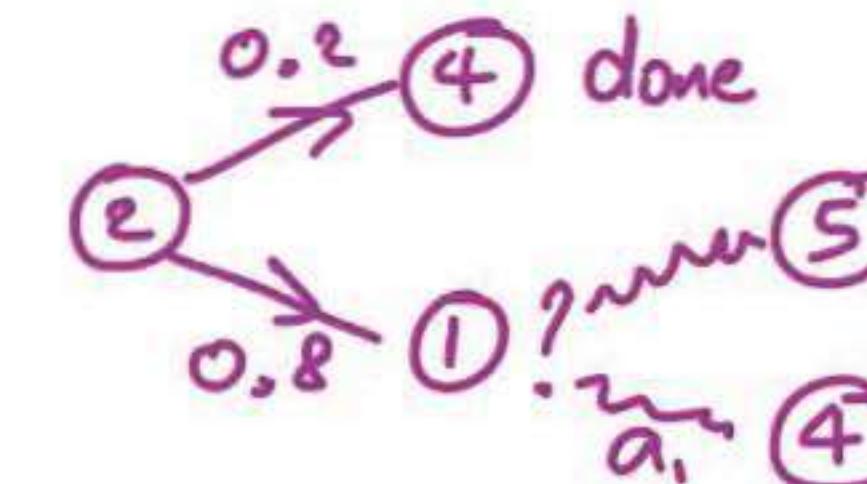
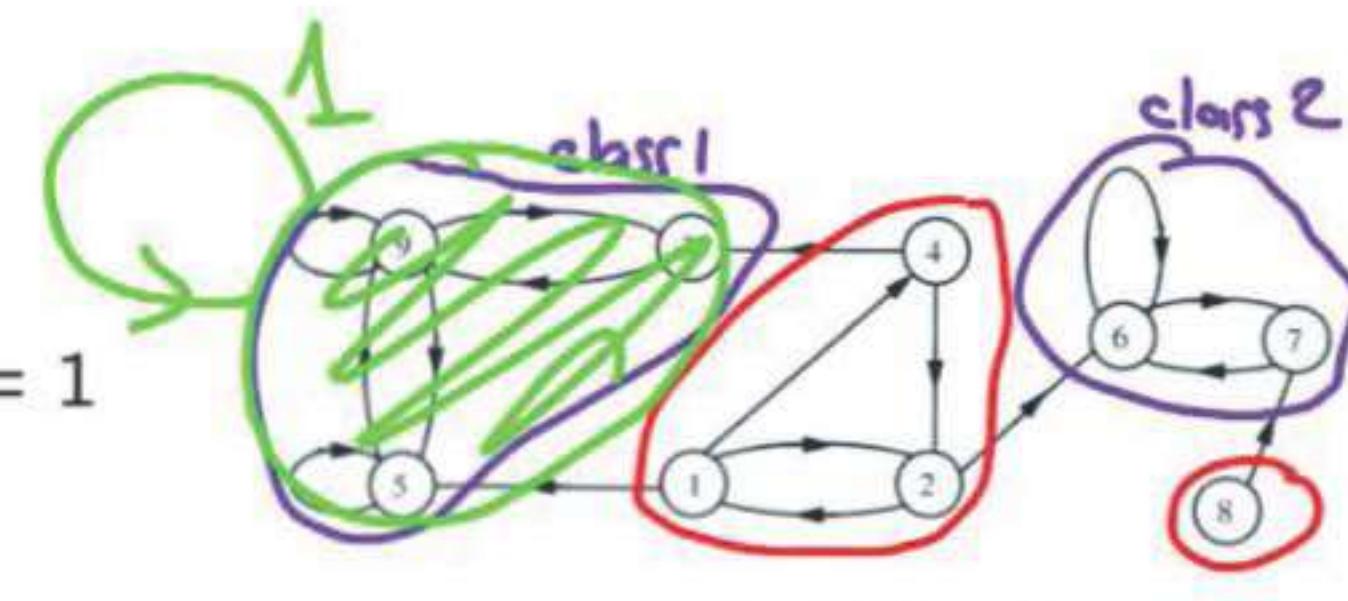
$$i = 4, a_i = 1$$

$$i = 5, a_i = 0$$

otherwise, $a_i = ?$

$$\begin{cases} a_1 = 18/28 \\ a_2 = 20/28 \\ a_3 = 15/28 \end{cases}$$

- unique solution from $a_i = \sum_{j=1}^m p_{ij} a_j$ the i
state s $a_s = 1$, and $a_{s'} = 0$ for the other absorbing state



$$\left\{ \begin{array}{l} a_2 = 0.2 a_4 + 0.8 a_1 \\ a_1 = 0.6 a_2 + 0.4 a_3 \\ a_3 = 0.3 a_2 + 0.5 a_1 + 0.2 a_5 \end{array} \right.$$

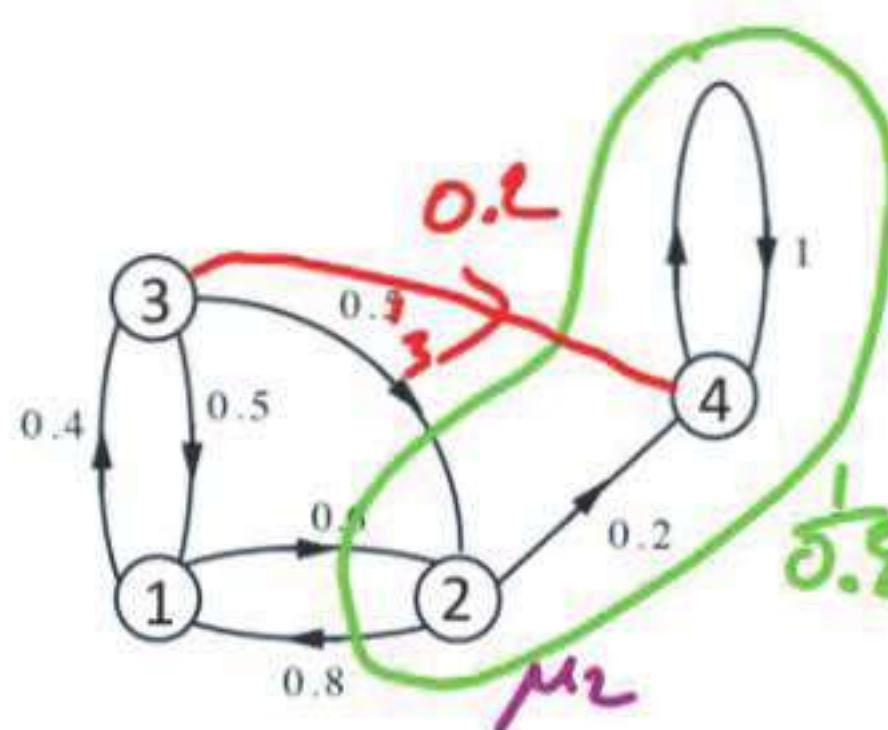
$a_1 + b_1 = 1$

$a_2 + b_2 = 1$

$a_3 + b_3 = 1$

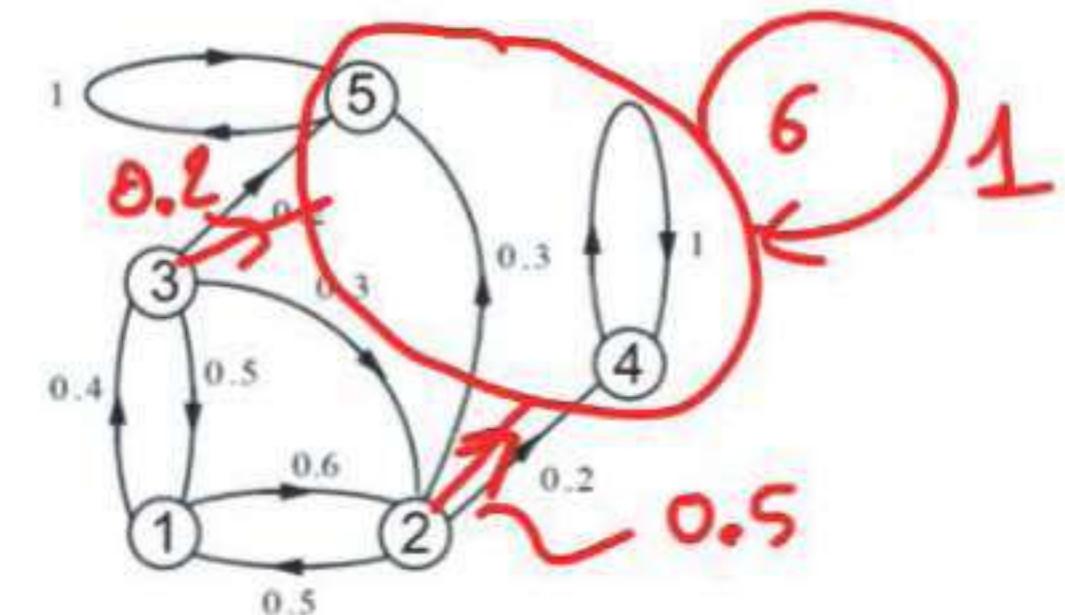
expected time to absorption

- find expected number of transitions μ_i until reaching 4, given that the initial state is i



$$\begin{aligned} \mu_i = 0 & \text{ for } i = 4 \\ \text{for all others, } \mu_i = ? & \\ \mu_2 = 5 & \quad \left| \begin{array}{l} \mu_1 = 110/8 \\ \mu_2 = 96/8 = 12 \\ \mu_3 = 111/8 \end{array} \right. \end{aligned}$$

- unique solution from $\mu_i = 1 + \sum_j p_{ij}\mu_j$



+1

$$\begin{aligned} \mu_2 &= 0.2\mu_4 + 0.8\mu_1 \\ \mu_2 &= 1 + 0.8\mu_1 \\ \mu_1 &= 1 + 0.6\mu_2 + 0.4\mu_3 \\ \mu_3 &= 1 + 0.5\mu_1 + 0.5\mu_2 \end{aligned}$$

mean first passage and recurrence times

- chain with one recurrent class; fix a recurrent state s

- mean first passage time from i to s :

$$t_i = E[\min\{n \geq 0 \text{ such that } X_n = s \mid X_0 = i\}]$$

– unique solution to:

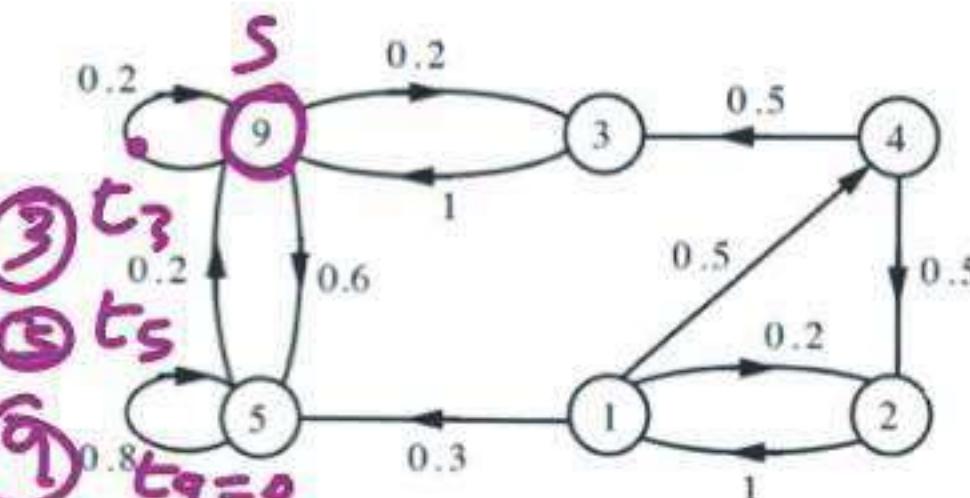
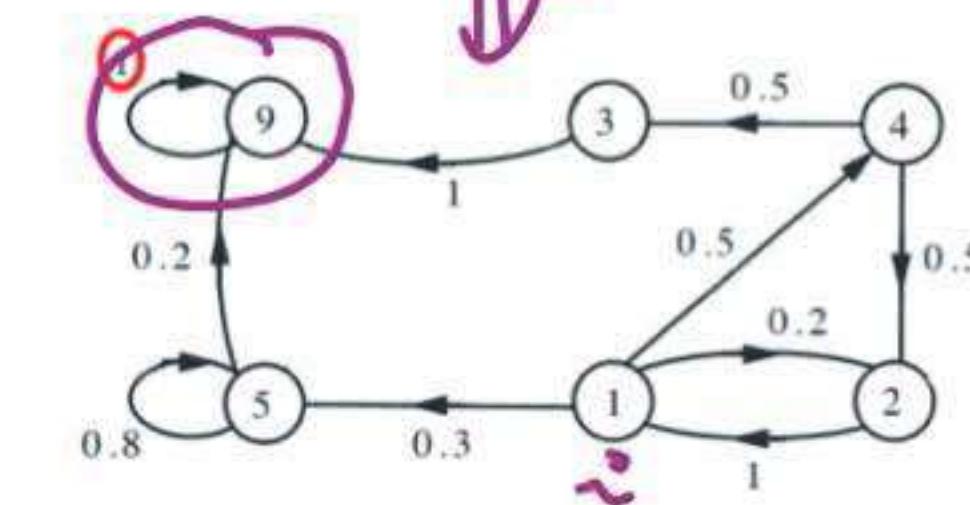
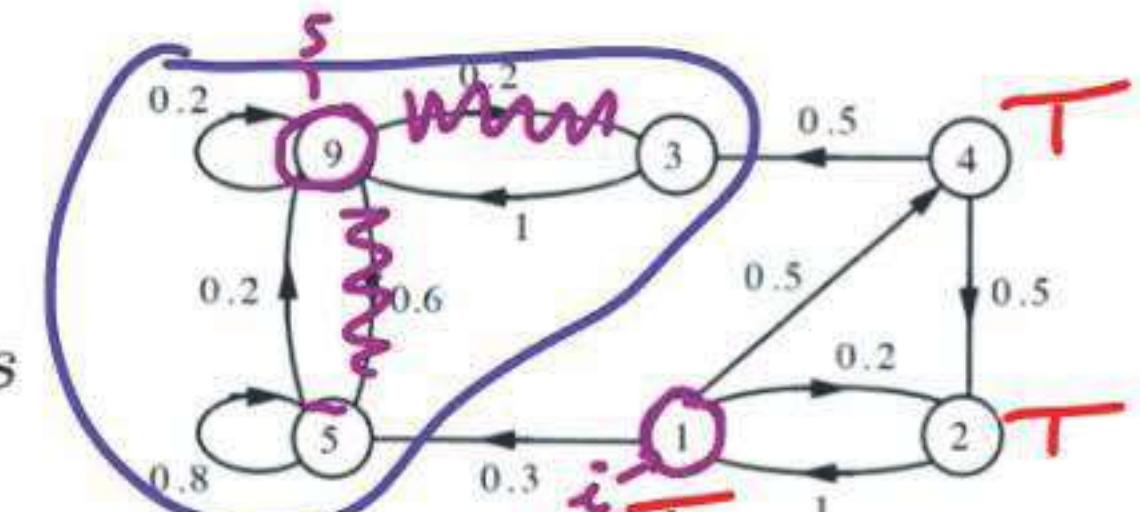
$$\begin{aligned} t_s &= 0, \\ t_i &= 1 + \sum_j p_{ij} t_j, \end{aligned} \quad \text{for all } i \neq s$$

- mean recurrence time of s

$$t_s^* = E[\min\{n \geq 1 \text{ such that } X_n = s \mid X_0 = s\}]$$

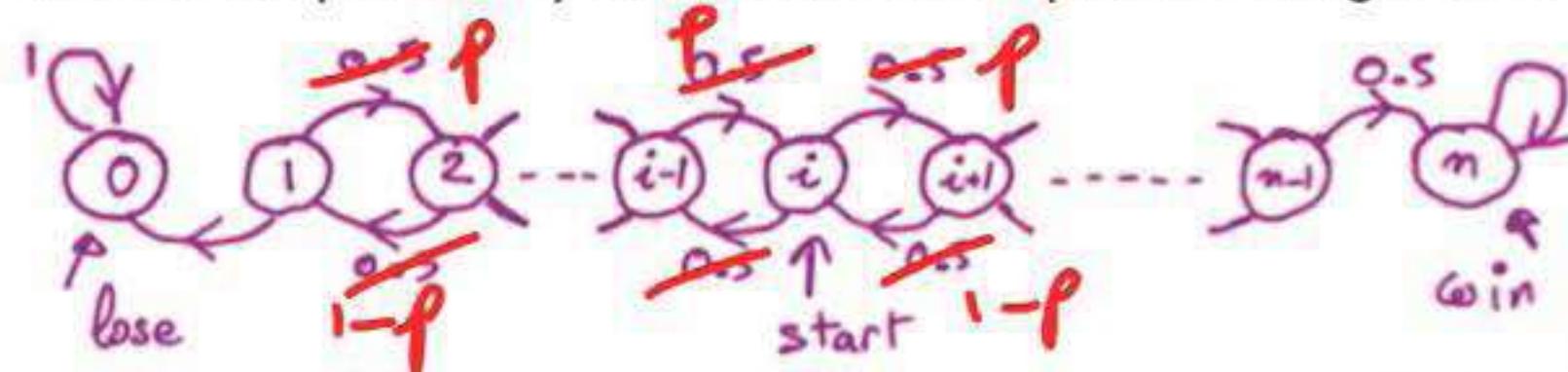
– solution to:

$$t_s^* = 1 + \sum_j p_{sj} t_j$$



gambler's example

- a gambler starts with i dollars; each time, she bets \$1 in a fair game, until she either has 0 or n dollars.
- what is the probability a_i that she ends up with having n dollars?



- expected wealth at the end? $0 \cdot (1 - a_i) + n \cdot a_i = n \times \frac{i}{n} = i$!!

- how long does the gambler expect to stay in the game?

- μ_i = expected number of plays, starting from i

- for $i = 0, n$: $\mu_i = 0$

- in general

$$\mu_i = 1 + \sum_j p_{ij} \mu_j$$

- in case of unfavorable odds?

$$r = \frac{1-p}{p}$$

$$p \neq 0.5 \quad a_i = \frac{1-r^i}{1-r^n}$$

$$\begin{aligned} i = 0, a_i &= 0 \checkmark & i = n, a_i &= 1 \checkmark \\ 0 < i < n, a_i &= ? & a_i &= 0.5 a_{i+1} + 0.5 a_{i-1} \\ a_i &= \frac{i}{n} & \boxed{a_i = \frac{i}{n}} & n \rightarrow \infty \end{aligned}$$

$$\begin{aligned} 1 < i < n \quad \mu_i &= 1 + \cancel{0.5 \mu_{i+1}} + \cancel{0.5 \mu_{i-1}} \\ \boxed{\mu_i = i(n-i)} & \end{aligned}$$

$$\begin{aligned} \mu_i &= \left(\frac{n!}{(i-1)!}\right) \left(i-n \cdot \frac{1-r^i}{1-r^n}\right) \end{aligned}$$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

Mathematical background

- Sets and De Morgan's laws
- Sequences and their limits
- Infinite series
 - The geometric series
- Sums with multiple indices
- Countable and uncountable sets

Sets

- A collection of distinct elements

$\{a, b, c, d\}$

finite

\mathbb{R} : real numbers infinite

$\{x \in \mathbb{R} : \cos(x) > 1/2\}$

Ω : universal set

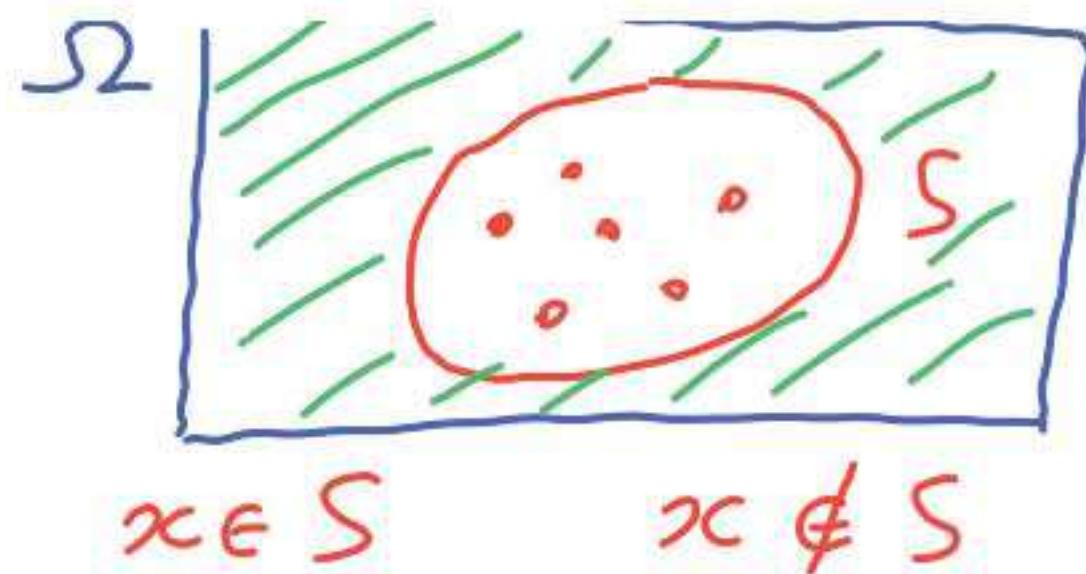
\emptyset : empty set

$\Omega^c = \emptyset$



$S \subset T : x \in S \Rightarrow x \in T$

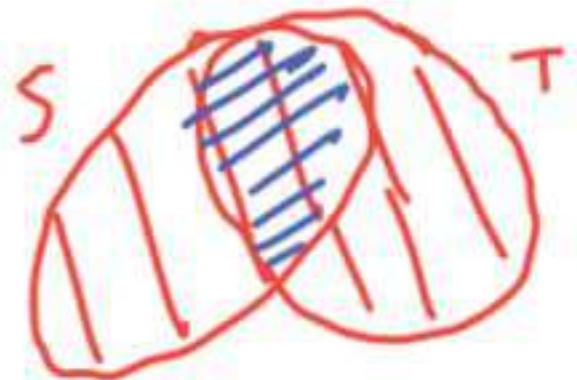
\subseteq



S^c
 $x \in S^c$ if $x \in \Omega$,
 $x \notin S$

$(S^c)^c = S$

Unions and intersections



$S \cup T$

$x \in S \cup T \Leftrightarrow x \in S \text{ or } x \in T$

$S \cap T$

$x \in S \cap T \Leftrightarrow x \in S \text{ and } x \in T$

S_n $n = 1, 2, \dots$



$x \in \bigcup_n S_n$ iff $x \in S_n$, for some n

$x \in \bigcap_n S_n$ iff $x \in S_n$, for all n

Set properties

- $S \cup T = T \cup S,$
- $S \cap (T \cup U) = (S \cap T) \cup (S \cap U),$
- $(S^c)^c = S,$
- $S \cup \Omega = \Omega,$



$S \cup T \cup U$

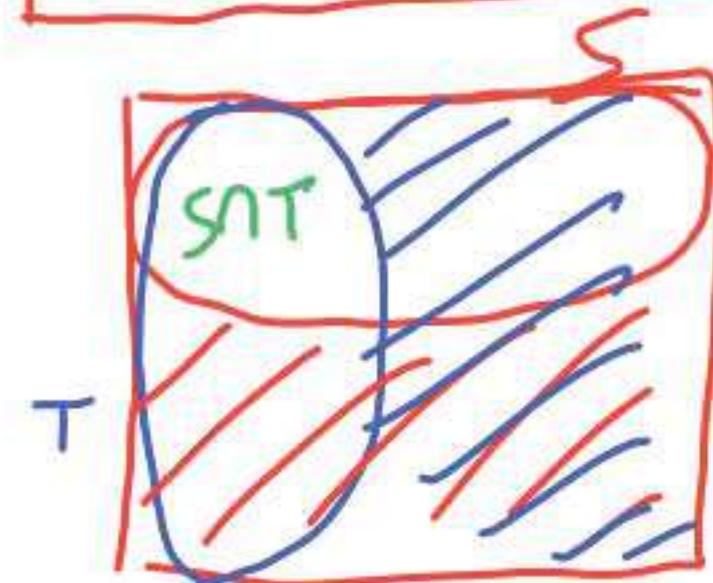
- $S \cup (T \cup U) = (S \cup T) \cup U,$
- $S \cup (T \cap U) = (S \cup T) \cap (S \cup U),$
- $S \cap S^c = \emptyset,$
- $S \cap \Omega = S.$

$S \cap (T \cap U) = (S \cap T) \cap U$

$$\left. \begin{array}{l} S \subset T \\ T \subset S \end{array} \right\} \Rightarrow S = T$$

De Morgan's laws

$$(S \cap T)^c = S^c \cup T^c$$



$$\begin{array}{ll} S \rightarrow S^c & T \rightarrow T^c \\ S^c \rightarrow S & T^c \rightarrow T \end{array}$$

$$(S^c \cap T^c)^c = S \cup T$$

$$S^c \cap T^c = (S \cup T)^c$$

$$\begin{aligned} (\bigcap_n S_n)^c &= \bigcup_n S_n^c \\ \bullet (\bigcup_n S_n)^c &= \bigcap_n S_n^c \end{aligned}$$

$$x \in (S \cap T)^c \Leftrightarrow x \notin S \cap T \Leftrightarrow \left\{ \begin{array}{l} x \notin S \\ \text{or} \\ x \notin T \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{l} x \in S^c \\ \text{or} \\ x \in T^c \end{array} \right\} \Leftrightarrow x \in S^c \cup T^c$$

Mathematical background: Sequences and their limits

a_1, a_2, a_3, \dots

$i \in \mathbb{N} = \{1, 2, 3, \dots\}$

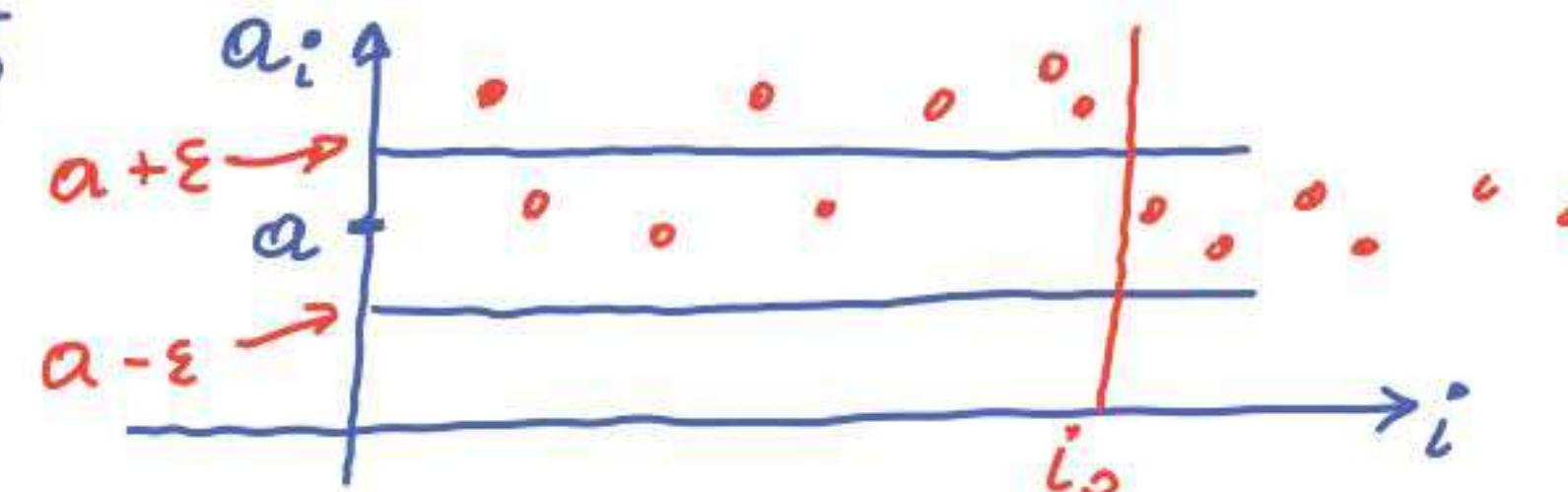
sequence $a_i, \{a_i\}$

$a_i \in S$ $S = \mathbb{R}$ \mathbb{R}^n

function $f: \mathbb{N} \rightarrow S$

$$f(i) = a_i$$

$$\left. \begin{array}{l} a_i \rightarrow a \\ i \rightarrow \infty \\ \lim_{i \rightarrow \infty} a_i = a \end{array} \right\}$$



For any $\varepsilon > 0$, there exists i_0 , such that
if $i \geq i_0$, then $|a_i - a| < \varepsilon$

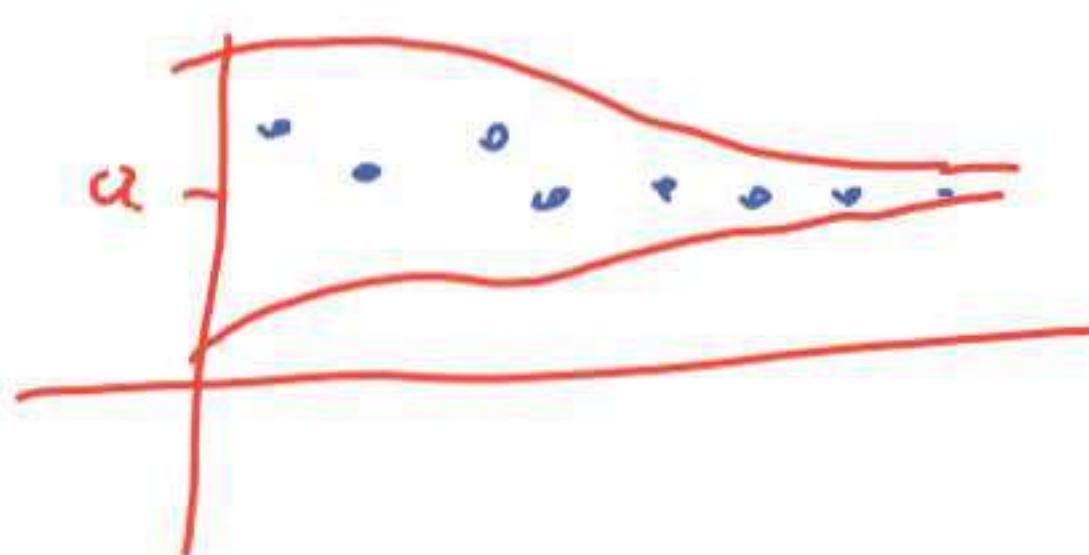
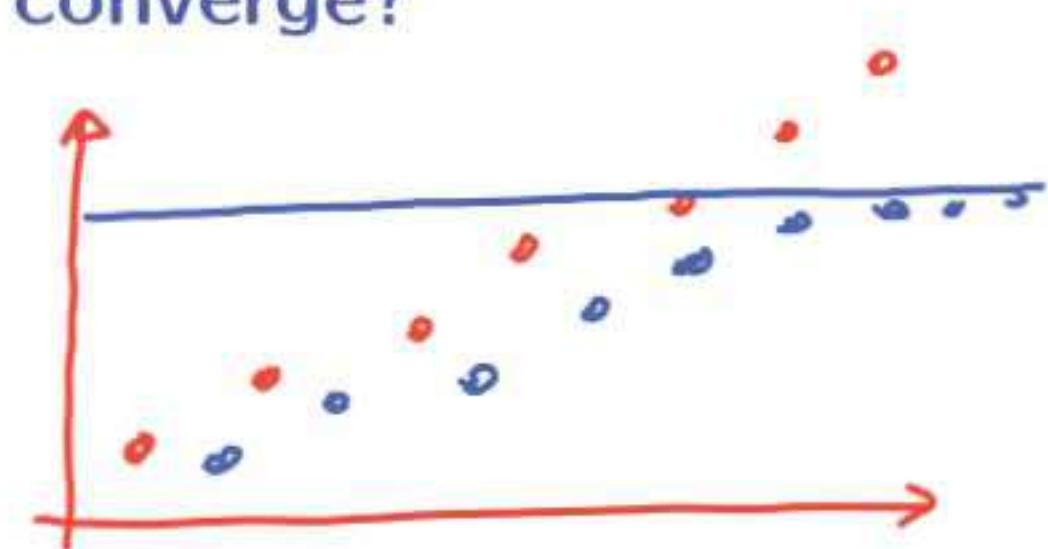
$$\left. \begin{array}{l} a_i \rightarrow a \\ b_i \rightarrow b \end{array} \right\} \Rightarrow a_i + b_i \rightarrow a + b$$

$$\begin{aligned} g: &\text{continuous} \\ &\Rightarrow g(a_i) \rightarrow g(a) \end{aligned}$$

$$a_i^2 \rightarrow a^2$$

Mathematical background: When does a sequence converge?

- If $a_i \leq a_{i+1}$, for all i , then either:
 - the sequence “converges to ∞ ”
 - the sequence converges to some real number a
- If $|a_i - a| \leq b_i$, for all i , and $b_i \rightarrow 0$, then $a_i \rightarrow a$



Mathematical background: Infinite series

$$\sum_{i=1}^{\infty} a_i = \lim_{n \rightarrow \infty} \sum_{i=1}^n a_i.$$

provided limit exists

- If $a_i \geq 0$: limit exists ←
- if terms a_i do not all have the same sign:
 - limit need not exist
 - limit may exist but be different if we sum in a different order
 - **Fact:** limit exists and independent of order of summation if $\sum_{i=1}^{\infty} |a_i| < \infty$

Mathematical background: Geometric series

$$S = \sum_{i=0}^{\infty} \alpha^i = 1 + \alpha + \alpha^2 + \dots = \frac{1}{1 - \alpha} \quad |\alpha| < 1$$

$$(1 - \alpha)(1 + \alpha + \dots + \alpha^n) = 1 - \alpha^{n+1}$$

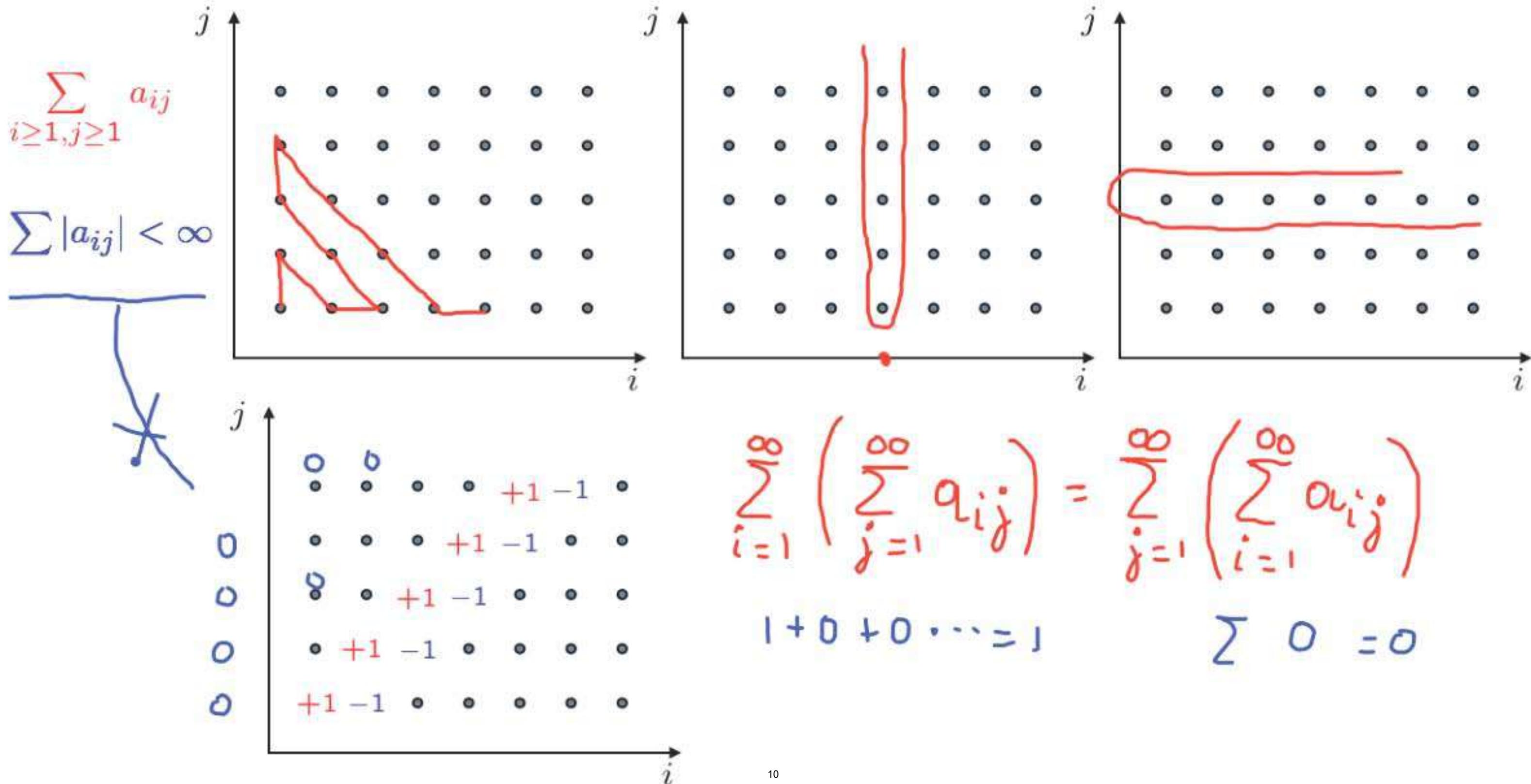
$$n \rightarrow \infty$$

$$(1 - \alpha) S = 1$$

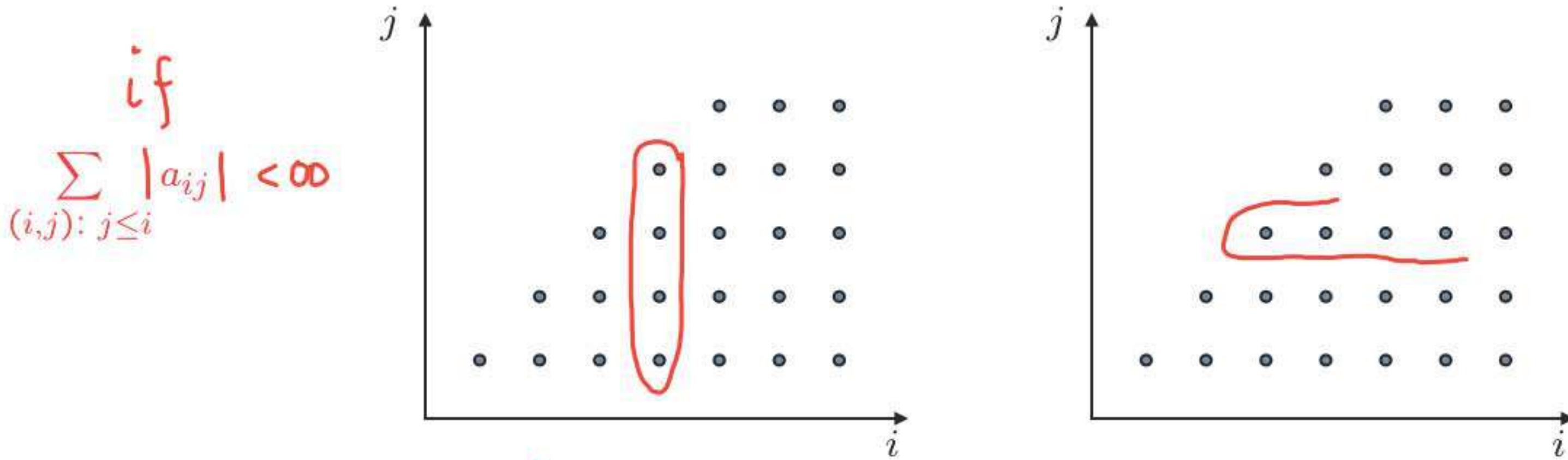
$$S = 1 + \sum_{i=1}^{\infty} \alpha^i = 1 + \alpha \sum_{i=0}^{\infty} \alpha^i = 1 + \alpha S \Rightarrow S(1 - \alpha) = 1$$

$S < \infty$ taken for granted

About the order of summation in series with multiple indices



About the order of summation in series with multiple indices



$$\sum_{i=1}^{\infty} \sum_{j=1}^i a_{ij} =$$

$$\sum_{j=1}^{\infty} \sum_{i=j}^{\infty} a_{ij}$$

Countable versus uncountable infinite sets

- Countable: can be put in 1-1 correspondence with positive integers

- positive integers $1, 2, 3, \dots$

- integers $0, 1, -1, 2, -2, 3, -3, \dots$

- pairs of positive integers

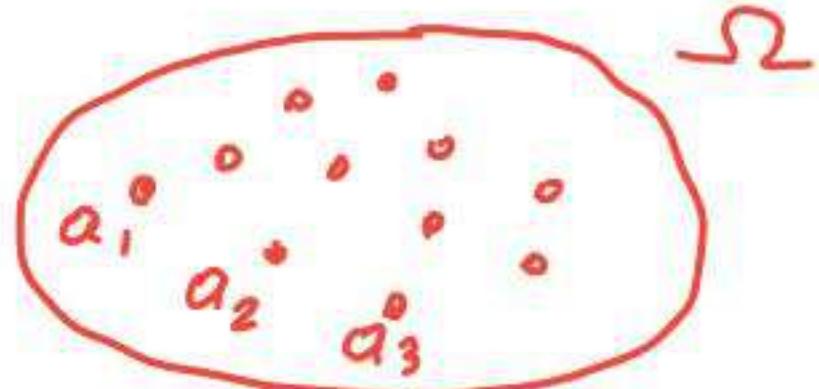
- rational numbers q , with $0 < q < 1$

$$\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \cancel{\frac{2}{4}}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \dots$$

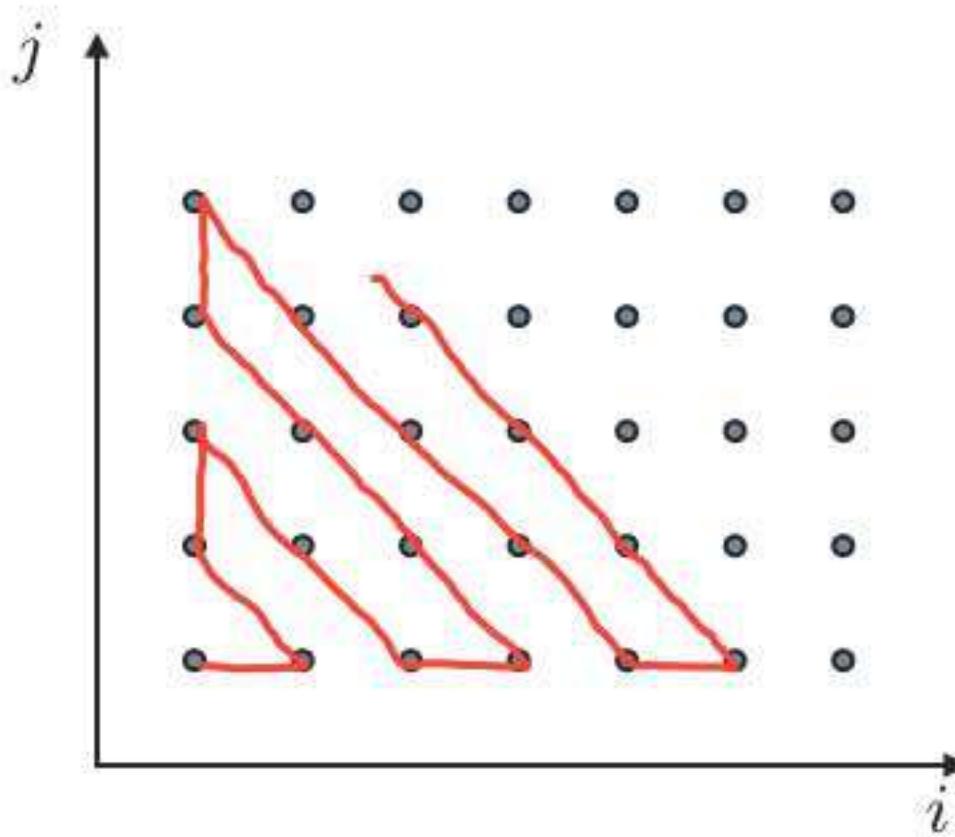
- Uncountable: not countable

- the interval $[0, 1]$

- the reals, the plane, ...



$$\{a_1, a_2, a_3, \dots\} = \Omega$$



The reals are uncountable

- Cantor's diagonalization argument

$\rightarrow \{x \in (0,1) : \text{decimal expansion only has } 3,4\}$

If countable

" $\{x_1, x_2, x_3, \dots\}$

$$x_1: \quad 0.\underline{\underline{3}}43443\dots$$

$$\underline{\underline{.433\dots}} = x$$

$$x_2: \quad 0.4\underline{\underline{3}}43443$$

$$\neq x_i$$

$$x_3: \quad 0.33\underline{\underline{0}}3444$$

for all i

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.

Introduction to Probability

SECOND EDITION

Dimitri P. Bertsekas and John N. Tsitsiklis

Massachusetts Institute of Technology

Selected Summary Material – All Rights Reserved

WWW site for book information and orders

<http://www.athenasc.com>



Athena Scientific, Belmont, Massachusetts

1

Sample Space and Probability

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

1.1. Sets	p. 3
1.2. Probabilistic Models	p. 6
1.3. Conditional Probability	p. 18
1.4. Total Probability Theorem and Bayes' Rule	p. 28
1.5. Independence	p. 34
1.6. Counting	p. 44
1.7. Summary and Discussion	p. 51
Problems	p. 53

1.1 SETS

1.2 PROBABILISTIC MODELS

Elements of a Probabilistic Model

- The **sample space** Ω , which is the set of all possible **outcomes** of an experiment.
- The **probability law**, which assigns to a set A of possible outcomes (also called an **event**) a nonnegative number $\mathbf{P}(A)$ (called the **probability** of A) that encodes our knowledge or belief about the collective “likelihood” of the elements of A . The probability law must satisfy certain properties to be introduced shortly.

Probability Axioms

1. **(Nonnegativity)** $\mathbf{P}(A) \geq 0$, for every event A .
2. **(Additivity)** If A and B are two disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

More generally, if the sample space has an infinite number of elements and A_1, A_2, \dots is a sequence of disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A_1 \cup A_2 \cup \dots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \dots$$

3. **(Normalization)** The probability of the entire sample space Ω is equal to 1, that is, $\mathbf{P}(\Omega) = 1$.

Discrete Probability Law

If the sample space consists of a finite number of possible outcomes, then the probability law is specified by the probabilities of the events that consist of a single element. In particular, the probability of any event $\{s_1, s_2, \dots, s_n\}$ is the sum of the probabilities of its elements:

$$\mathbf{P}(\{s_1, s_2, \dots, s_n\}) = \mathbf{P}(s_1) + \mathbf{P}(s_2) + \dots + \mathbf{P}(s_n).$$

Discrete Uniform Probability Law

If the sample space consists of n possible outcomes which are equally likely (i.e., all single-element events have the same probability), then the probability of any event A is given by

$$\mathbf{P}(A) = \frac{\text{number of elements of } A}{n}.$$

Some Properties of Probability Laws

Consider a probability law, and let A , B , and C be events.

- (a) If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.
- (b) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.
- (c) $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.
- (d) $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A^c \cap B^c \cap C)$.

1.3 CONDITIONAL PROBABILITY

Properties of Conditional Probability

- The conditional probability of an event A , given an event B with $\mathbf{P}(B) > 0$, is defined by

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

and specifies a new (conditional) probability law on the same sample space Ω . In particular, all properties of probability laws remain valid for conditional probability laws.

- Conditional probabilities can also be viewed as a probability law on a new universe B , because all of the conditional probability is concentrated on B .
- If the possible outcomes are finitely many and equally likely, then

$$\mathbf{P}(A | B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.$$

1.4 TOTAL PROBABILITY THEOREM AND BAYES' RULE

Total Probability Theorem

Let A_1, \dots, A_n be disjoint events that form a partition of the sample space (each possible outcome is included in exactly one of the events A_1, \dots, A_n) and assume that $\mathbf{P}(A_i) > 0$, for all i . Then, for any event B , we have

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1 \cap B) + \cdots + \mathbf{P}(A_n \cap B) \\ &= \mathbf{P}(A_1)\mathbf{P}(B | A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B | A_n).\end{aligned}$$

1.5 INDEPENDENCE

Independence

- Two events A and B are said to be **independent** if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

If in addition, $\mathbf{P}(B) > 0$, independence is equivalent to the condition

$$\mathbf{P}(A | B) = \mathbf{P}(A).$$

- If A and B are independent, so are A and B^c .
- Two events A and B are said to be **conditionally independent**, given another event C with $\mathbf{P}(C) > 0$, if

$$\mathbf{P}(A \cap B | C) = \mathbf{P}(A | C)\mathbf{P}(B | C).$$

If in addition, $\mathbf{P}(B \cap C) > 0$, conditional independence is equivalent to the condition

$$\mathbf{P}(A | B \cap C) = \mathbf{P}(A | C).$$

- Independence does not imply conditional independence, and vice versa.

Definition of Independence of Several Events

We say that the events A_1, A_2, \dots, A_n are **independent** if

$$\mathbf{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbf{P}(A_i), \quad \text{for every subset } S \text{ of } \{1, 2, \dots, n\}.$$

1.6 COUNTING

The Counting Principle

Consider a process that consists of r stages. Suppose that:

- (a) There are n_1 possible results at the first stage.
- (b) For every possible result at the first stage, there are n_2 possible results at the second stage.
- (c) More generally, for any sequence of possible results at the first $i - 1$ stages, there are n_i possible results at the i th stage.

Then, the total number of possible results of the r -stage process is

$$n_1 n_2 \cdots n_r.$$

Summary of Counting Results

- **Permutations** of n objects: $n!$.
- **k -permutations** of n objects: $n!/(n - k)!$.
- **Combinations** of k out of n objects: $\binom{n}{k} = \frac{n!}{k!(n - k)!}$.
- **Partitions** of n objects into r groups, with the i th group having n_i objects:

$$\binom{n}{n_1, n_2, \dots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}.$$

1.7 SUMMARY AND DISCUSSION

2

Discrete Random Variables

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

2.1. Basic Concepts	p. 72
2.2. Probability Mass Functions	p. 74
2.3. Functions of Random Variables	p. 80
2.4. Expectation, Mean, and Variance	p. 81
2.5. Joint PMFs of Multiple Random Variables	p. 92
2.6. Conditioning	p. 97
2.7. Independence	p. 109
2.8. Summary and Discussion	p. 115
Problems	p. 119

2.1 BASIC CONCEPTS

Main Concepts Related to Random Variables

Starting with a probabilistic model of an experiment:

- A **random variable** is a real-valued function of the outcome of the experiment.
- A **function of a random variable** defines another random variable.
- We can associate with each random variable certain “averages” of interest, such as the **mean** and the **variance**.
- A random variable can be **conditioned** on an event or on another random variable.
- There is a notion of **independence** of a random variable from an event or from another random variable.

Concepts Related to Discrete Random Variables

Starting with a probabilistic model of an experiment:

- A **discrete random variable** is a real-valued function of the outcome of the experiment that can take a finite or countably infinite number of values.
- A discrete random variable has an associated **probability mass function (PMF)**, which gives the probability of each numerical value that the random variable can take.
- A **function of a discrete random variable** defines another discrete random variable, whose PMF can be obtained from the PMF of the original random variable.

2.2 PROBABILITY MASS FUNCTIONS

Calculation of the PMF of a Random Variable X

For each possible value x of X :

1. Collect all the possible outcomes that give rise to the event $\{X = x\}$.
2. Add their probabilities to obtain $p_X(x)$.

2.3 FUNCTIONS OF RANDOM VARIABLES

2.4 EXPECTATION, MEAN, AND VARIANCE

Expectation

We define the **expected value** (also called the **expectation** or the **mean**) of a random variable X , with PMF p_X , by

$$\mathbf{E}[X] = \sum_x x p_X(x).$$

Expected Value Rule for Functions of Random Variables

Let X be a random variable with PMF p_X , and let $g(X)$ be a function of X . Then, the expected value of the random variable $g(X)$ is given by

$$\mathbf{E}[g(X)] = \sum_x g(x) p_X(x).$$

Variance

The variance $\text{var}(X)$ of a random variable X is defined by

$$\text{var}(X) = \mathbf{E} \left[(X - \mathbf{E}[X])^2 \right],$$

and can be calculated as

$$\text{var}(X) = \sum_x (x - \mathbf{E}[X])^2 p_X(x).$$

It is always nonnegative. Its square root is denoted by σ_X and is called the **standard deviation**.

Mean and Variance of a Linear Function of a Random Variable

Let X be a random variable and let

$$Y = aX + b,$$

where a and b are given scalars. Then,

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b, \quad \text{var}(Y) = a^2 \text{var}(X).$$

Variance in Terms of Moments Expression

$$\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

2.5 JOINT PMFS OF MULTIPLE RANDOM VARIABLES

Summary of Facts About Joint PMFs

Let X and Y be random variables associated with the same experiment.

- The **joint PMF** $p_{X,Y}$ of X and Y is defined by

$$p_{X,Y}(x,y) = \mathbf{P}(X = x, Y = y).$$

- The **marginal PMFs** of X and Y can be obtained from the joint PMF, using the formulas

$$p_X(x) = \sum_y p_{X,Y}(x,y), \quad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

- A function $g(X, Y)$ of X and Y defines another random variable, and

$$\mathbf{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y).$$

If g is linear, of the form $aX + bY + c$, we have

$$\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c.$$

- The above have natural extensions to the case where more than two random variables are involved.

2.6 CONDITIONING

Summary of Facts About Conditional PMFs

Let X and Y be random variables associated with the same experiment.

- Conditional PMFs are similar to ordinary PMFs, but pertain to a universe where the conditioning event is known to have occurred.
- The conditional PMF of X given an event A with $\mathbf{P}(A) > 0$, is defined by

$$p_{X|A}(x) = \mathbf{P}(X = x | A)$$

and satisfies

$$\sum_x p_{X|A}(x) = 1.$$

- If A_1, \dots, A_n are disjoint events that form a partition of the sample space, with $\mathbf{P}(A_i) > 0$ for all i , then

$$p_X(x) = \sum_{i=1}^n \mathbf{P}(A_i) p_{X|A_i}(x).$$

(This is a special case of the total probability theorem.) Furthermore, for any event B , with $\mathbf{P}(A_i \cap B) > 0$ for all i , we have

$$p_{X|B}(x) = \sum_{i=1}^n \mathbf{P}(A_i | B) p_{X|A_i \cap B}(x).$$

- The conditional PMF of X given $Y = y$ is related to the joint PMF by

$$p_{X,Y}(x,y) = p_Y(y) p_{X|Y}(x | y).$$

- The conditional PMF of X given Y can be used to calculate the marginal PMF of X through the formula

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x | y).$$

- There are natural extensions of the above involving more than two random variables.

Summary of Facts About Conditional Expectations

Let X and Y be random variables associated with the same experiment.

- The conditional expectation of X given an event A with $\mathbf{P}(A) > 0$, is defined by

$$\mathbf{E}[X | A] = \sum_x x p_{X|A}(x).$$

For a function $g(X)$, we have

$$\mathbf{E}[g(X) | A] = \sum_x g(x) p_{X|A}(x).$$

- The conditional expectation of X given a value y of Y is defined by

$$\mathbf{E}[X | Y = y] = \sum_x x p_{X|Y}(x | y).$$

- If A_1, \dots, A_n be disjoint events that form a partition of the sample space, with $\mathbf{P}(A_i) > 0$ for all i , then

$$\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i) \mathbf{E}[X | A_i].$$

Furthermore, for any event B with $\mathbf{P}(A_i \cap B) > 0$ for all i , we have

$$\mathbf{E}[X | B] = \sum_{i=1}^n \mathbf{P}(A_i | B) \mathbf{E}[X | A_i \cap B].$$

- We have

$$\mathbf{E}[X] = \sum_y p_Y(y) \mathbf{E}[X | Y = y].$$

2.7 INDEPENDENCE

Summary of Facts About Independent Random Variables

Let A be an event, with $\mathbf{P}(A) > 0$, and let X and Y be random variables associated with the same experiment.

- X is independent of the event A if

$$p_{X|A}(x) = p_X(x), \quad \text{for all } x,$$

that is, if for all x , the events $\{X = x\}$ and A are independent.

- X and Y are independent if for all pairs (x, y) , the events $\{X = x\}$ and $\{Y = y\}$ are independent, or equivalently

$$p_{X,Y}(x, y) = p_X(x)p_Y(y), \quad \text{for all } x, y.$$

- If X and Y are independent random variables, then

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y].$$

Furthermore, for any functions g and h , the random variables $g(X)$ and $h(Y)$ are independent, and we have

$$\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)]\mathbf{E}[h(Y)].$$

- If X and Y are independent, then

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

2.8 SUMMARY AND DISCUSSION

Summary of Results for Special Random Variables

Discrete Uniform over $[a, b]$:

$$p_X(k) = \begin{cases} \frac{1}{b-a+1}, & \text{if } k = a, a+1, \dots, b, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)(b-a+2)}{12}.$$

Bernoulli with Parameter p : (Describes the success or failure in a single trial.)

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

$$\mathbf{E}[X] = p, \quad \text{var}(X) = p(1-p).$$

Binomial with Parameters p and n : (Describes the number of successes in n independent Bernoulli trials.)

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n,$$

$$\mathbf{E}[X] = np, \quad \text{var}(X) = np(1-p).$$

Geometric with Parameter p : (Describes the number of trials until the first success, in a sequence of independent Bernoulli trials.)

$$p_X(k) = (1-p)^{k-1} p, \quad k = 1, 2, \dots,$$

$$\mathbf{E}[X] = \frac{1}{p}, \quad \text{var}(X) = \frac{1-p}{p^2}.$$

Poisson with Parameter λ : (Approximates the binomial PMF when n is large, p is small, and $\lambda = np$.)

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \dots,$$

$$\mathbf{E}[X] = \lambda, \quad \text{var}(X) = \lambda.$$

3

General Random Variables

Excerpts from **Introduction to Probability: Second Edition**

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

3.1. Continuous Random Variables and PDFs	p. 140
3.2. Cumulative Distribution Functions	p. 148
3.3. Normal Random Variables	p. 153
3.4. Joint PDFs of Multiple Random Variables	p. 158
3.5. Conditioning	p. 164
3.6. The Continuous Bayes' Rule	p. 178
3.7. Summary and Discussion	p. 182
Problems	p. 184

3.1 CONTINUOUS RANDOM VARIABLES AND PDFS

Summary of PDF Properties

Let X be a continuous random variable with PDF f_X .

- $f_X(x) \geq 0$ for all x .
- $\int_{-\infty}^{\infty} f_X(x) dx = 1$.
- If δ is very small, then $\mathbf{P}([x, x + \delta]) \approx f_X(x) \cdot \delta$.
- For any subset B of the real line,

$$\mathbf{P}(X \in B) = \int_B f_X(x) dx.$$

Expectation of a Continuous Random Variable and its Properties

Let X be a continuous random variable with PDF f_X .

- The expectation of X is defined by

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

- The expected value rule for a function $g(X)$ has the form

$$\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

- The variance of X is defined by

$$\text{var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2] = \int_{-\infty}^{\infty} (x - \mathbf{E}[X])^2 f_X(x) dx.$$

- We have

$$0 \leq \text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

- If $Y = aX + b$, where a and b are given scalars, then

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b, \quad \text{var}(Y) = a^2\text{var}(X).$$

3.2 CUMULATIVE DISTRIBUTION FUNCTIONS

Properties of a CDF

The CDF F_X of a random variable X is defined by

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

and has the following properties.

- F_X is monotonically nondecreasing:

$$\text{if } x \leq y, \text{ then } F_X(x) \leq F_X(y).$$

- $F_X(x)$ tends to 0 as $x \rightarrow -\infty$, and to 1 as $x \rightarrow \infty$.
- If X is discrete, then $F_X(x)$ is a piecewise constant function of x .
- If X is continuous, then $F_X(x)$ is a continuous function of x .
- If X is discrete and takes integer values, the PMF and the CDF can be obtained from each other by summing or differencing:

$$F_X(k) = \sum_{i=-\infty}^k p_X(i),$$

$$p_X(k) = \mathbf{P}(X \leq k) - \mathbf{P}(X \leq k-1) = F_X(k) - F_X(k-1),$$

for all integers k .

- If X is continuous, the PDF and the CDF can be obtained from each other by integration or differentiation:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt, \quad f_X(x) = \frac{dF_X}{dx}(x).$$

(The second equality is valid for those x at which the PDF is continuous.)

3.3 NORMAL RANDOM VARIABLES

Normality is Preserved by Linear Transformations

If X is a normal random variable with mean μ and variance σ^2 , and if $a \neq 0$, b are scalars, then the random variable

$$Y = aX + b$$

is also normal, with mean and variance

$$\mathbf{E}[Y] = a\mu + b, \quad \text{var}(Y) = a^2\sigma^2.$$

CDF Calculation for a Normal Random Variable

For a normal random variable X with mean μ and variance σ^2 , we use a two-step procedure.

- (a) “Standardize” X , i.e., subtract μ and divide by σ to obtain a standard normal random variable Y .
- (b) Read the CDF value from the standard normal table:

$$\mathbf{P}(X \leq x) = \mathbf{P}\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = \mathbf{P}\left(Y \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

The standard normal table. The entries in this table provide the numerical values of $\Phi(y) = \mathbf{P}(Y \leq y)$, where Y is a standard normal random variable, for y between 0 and 3.49. For example, to find $\Phi(1.71)$, we look at the row corresponding to 1.7 and the column corresponding to 0.01, so that $\Phi(1.71) = .9564$. When y is negative, the value of $\Phi(y)$ can be found using the formula $\Phi(y) = 1 - \Phi(-y)$.

3.4 JOINT PDFS OF MULTIPLE RANDOM VARIABLES

Summary of Facts about Joint PDFs

Let X and Y be jointly continuous random variables with joint PDF $f_{X,Y}$.

- The **joint PDF** is used to calculate probabilities:

$$\mathbf{P}((X, Y) \in B) = \int \int_{(x,y) \in B} f_{X,Y}(x, y) dx dy.$$

- The **marginal PDFs** of X and Y can be obtained from the joint PDF, using the formulas

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx.$$

- The **joint CDF** is defined by $F_{X,Y}(x, y) = \mathbf{P}(X \leq x, Y \leq y)$, and determines the joint PDF through the formula

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x, y),$$

for every (x, y) at which the joint PDF is continuous.

- A function $g(X, Y)$ of X and Y defines a new random variable, and

$$\mathbf{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy.$$

If g is linear, of the form $aX + bY + c$, we have

$$\mathbf{E}[aX + bY + c] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c.$$

- The above have natural extensions to the case where more than two random variables are involved.

3.5 CONDITIONING

Conditional PDF Given an Event

- The conditional PDF $f_{X|A}$ of a continuous random variable X , given an event A with $\mathbf{P}(A) > 0$, satisfies

$$\mathbf{P}(X \in B | A) = \int_B f_{X|A}(x) dx.$$

- If A is a subset of the real line with $\mathbf{P}(X \in A) > 0$, then

$$f_{X|\{X \in A\}}(x) = \begin{cases} \frac{f_X(x)}{\mathbf{P}(X \in A)}, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

- Let A_1, A_2, \dots, A_n be disjoint events that form a partition of the sample space, and assume that $\mathbf{P}(A_i) > 0$ for all i . Then,

$$f_X(x) = \sum_{i=1}^n \mathbf{P}(A_i) f_{X|A_i}(x)$$

(a version of the total probability theorem).

Conditional PDF Given a Random Variable

Let X and Y be jointly continuous random variables with joint PDF $f_{X,Y}$.

- The joint, marginal, and conditional PDFs are related to each other by the formulas

$$f_{X,Y}(x,y) = f_Y(y) f_{X|Y}(x|y),$$

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x|y) dy.$$

The conditional PDF $f_{X|Y}(x|y)$ is defined only for those y for which $f_Y(y) > 0$.

- We have

$$\mathbf{P}(X \in A | Y = y) = \int_A f_{X|Y}(x|y) dx.$$

Summary of Facts About Conditional Expectations

Let X and Y be jointly continuous random variables, and let A be an event with $\mathbf{P}(A) > 0$.

- **Definitions:** The conditional expectation of X given the event A is defined by

$$\mathbf{E}[X | A] = \int_{-\infty}^{\infty} x f_{X|A}(x) dx.$$

The conditional expectation of X given that $Y = y$ is defined by

$$\mathbf{E}[X | Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x | y) dx.$$

- **The expected value rule:** For a function $g(X)$, we have

$$\mathbf{E}[g(X) | A] = \int_{-\infty}^{\infty} g(x) f_{X|A}(x) dx,$$

and

$$\mathbf{E}[g(X) | Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x | y) dx.$$

- **Total expectation theorem:** Let A_1, A_2, \dots, A_n be disjoint events that form a partition of the sample space, and assume that $\mathbf{P}(A_i) > 0$ for all i . Then,

$$\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i) \mathbf{E}[X | A_i].$$

Similarly,

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} \mathbf{E}[X | Y = y] f_Y(y) dy.$$

- There are natural analogs for the case of functions of several random variables. For example,

$$\mathbf{E}[g(X, Y) | Y = y] = \int g(x, y) f_{X|Y}(x | y) dx,$$

and

$$\mathbf{E}[g(X, Y)] = \int \mathbf{E}[g(X, Y) | Y = y] f_Y(y) dy.$$

Independence of Continuous Random Variables

Let X and Y be jointly continuous random variables.

- X and Y are **independent** if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \text{for all } x,y.$$

- If X and Y are independent, then

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y].$$

Furthermore, for any functions g and h , the random variables $g(X)$ and $h(Y)$ are independent, and we have

$$\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)]\mathbf{E}[h(Y)].$$

- If X and Y are independent, then

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

3.6 BAYES' RULE AND APPLICATIONS IN INFERENCE

Bayes' Rule Relations for Random Variables

Let X and Y be two random variables.

- If X and Y are discrete, we have for all x, y with $p_X(x) \neq 0, p_Y(y) \neq 0$,

$$p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y),$$

and the terms on the two sides in this relation are both equal to

$$p_{X,Y}(x,y).$$

- If X is discrete and Y is continuous, we have for all x, y with $p_X(x) \neq 0, f_Y(y) \neq 0$,

$$p_X(x)f_{Y|X}(y|x) = f_Y(y)p_{X|Y}(x|y),$$

and the terms on the two sides in this relation are both equal to

$$\lim_{\delta \rightarrow 0} \frac{\mathbf{P}(X = x, y \leq Y \leq y + \delta)}{\delta}.$$

- If X and Y are continuous, we have for all x, y with $f_X(x) \neq 0, f_Y(y) \neq 0$,

$$f_X(x)f_{Y|X}(y|x) = f_Y(y)f_{X|Y}(x|y),$$

and the terms on the two sides in this relation are both equal to

$$\lim_{\delta \rightarrow 0} \frac{\mathbf{P}(x \leq X \leq x + \delta, y \leq Y \leq y + \delta)}{\delta^2}.$$

3.7 SUMMARY AND DISCUSSION

Summary of Results for Special Random Variables

Continuous Uniform Over $[a, b]$:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)^2}{12}.$$

Exponential with Parameter λ :

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases} \quad F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{1}{\lambda}, \quad \text{var}(X) = \frac{1}{\lambda^2}.$$

Normal with Parameters μ and $\sigma^2 > 0$:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2},$$

$$\mathbf{E}[X] = \mu, \quad \text{var}(X) = \sigma^2.$$

4

Further Topics on Random Variables

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

4.1. Derived Distributions	p. 202
4.2. Covariance and Correlation	p. 217
4.3. Conditional Expectation and Variance Revisited	p. 222
4.4. Transforms	p. 229
4.5. Sum of a Random Number of Independent Random Variables	p. 240
4.6. Summary and Discussion	p. 244
Problems	p. 246

4.1 DERIVED DISTRIBUTIONS

Calculation of the PDF of a Function $Y = g(X)$ of a Continuous Random Variable X

1. Calculate the CDF F_Y of Y using the formula

$$F_Y(y) = \mathbf{P}(g(X) \leq y) = \int_{\{x \mid g(x) \leq y\}} f_X(x) dx.$$

2. Differentiate to obtain the PDF of Y :

$$f_Y(y) = \frac{dF_Y}{dy}(y).$$

The PDF of a Linear Function of a Random Variable

Let X be a continuous random variable with PDF f_X , and let

$$Y = aX + b,$$

where a and b are scalars, with $a \neq 0$. Then,

$$f_Y(y) = \frac{1}{|a|} f_X \left(\frac{y - b}{a} \right).$$

PDF Formula for a Strictly Monotonic Function of a Continuous Random Variable

Suppose that g is strictly monotonic and that for some function h and all x in the range of X we have

$$y = g(x) \quad \text{if and only if} \quad x = h(y).$$

Assume that h is differentiable. Then, the PDF of Y in the region where $f_Y(y) > 0$ is given by

$$f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy}(y) \right|.$$

4.2 COVARIANCE AND CORRELATION

Covariance and Correlation

- The **covariance** of X and Y is given by

$$\text{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y].$$

- If $\text{cov}(X, Y) = 0$, we say that X and Y are **uncorrelated**.
- If X and Y are independent, they are uncorrelated. The converse is not always true.
- We have

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y).$$

- The **correlation coefficient** $\rho(X, Y)$ of two random variables X and Y with positive variances is defined by

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}},$$

and satisfies

$$-1 \leq \rho(X, Y) \leq 1.$$

4.3 CONDITIONAL EXPECTATION AND VARIANCE REVISITED

Law of Iterated Expectations: $\mathbf{E}[\mathbf{E}[X | Y]] = \mathbf{E}[X].$

Law of Total Variance: $\text{var}(X) = \mathbf{E}[\text{var}(X | Y)] + \text{var}(\mathbf{E}[X | Y]).$

Properties of the Conditional Expectation and Variance

- $\mathbf{E}[X | Y = y]$ is a number whose value depends on y .
- $\mathbf{E}[X | Y]$ is a function of the random variable Y , hence a random variable. Its value is $\mathbf{E}[X | Y = y]$ whenever the value of Y is y .
- $\mathbf{E}[\mathbf{E}[X | Y]] = \mathbf{E}[X]$ (law of iterated expectations).
- $\mathbf{E}[X | Y = y]$ may be viewed as an estimate of X given $Y = y$. The corresponding error $\mathbf{E}[X | Y] - X$ is a zero mean random variable that is uncorrelated with $\mathbf{E}[X | Y]$.
- $\text{var}(X | Y)$ is a random variable whose value is $\text{var}(X | Y = y)$ whenever the value of Y is y .
- $\text{var}(X) = \mathbf{E}[\text{var}(X | Y)] + \text{var}(\mathbf{E}[X | Y])$ (law of total variance).

4.4 TRANSFORMS**Summary of Transforms and their Properties**

- The transform associated with a random variable X is given by

$$M_X(s) = \mathbf{E}[e^{sX}] = \begin{cases} \sum_x e^{sx} p_X(x), & X \text{ discrete}, \\ \int_{-\infty}^{\infty} e^{sx} f_X(x) dx, & X \text{ continuous}. \end{cases}$$

- The distribution of a random variable is completely determined by the corresponding transform.
- Moment generating properties:

$$M_X(0) = 1, \quad \left. \frac{d}{ds} M_X(s) \right|_{s=0} = \mathbf{E}[X], \quad \left. \frac{d^n}{ds^n} M_X(s) \right|_{s=0} = \mathbf{E}[X^n].$$

- If $Y = aX + b$, then $M_Y(s) = e^{sb} M_X(as)$.
- If X and Y are independent, then $M_{X+Y}(s) = M_X(s)M_Y(s)$.

Transforms for Common Discrete Random Variables

Bernoulli(p) ($k = 0, 1$)

$$p_X(k) = \begin{cases} p, & \text{if } k = 1, \\ 1 - p, & \text{if } k = 0, \end{cases} \quad M_X(s) = 1 - p + pe^s.$$

Binomial(n, p) ($k = 0, 1, \dots, n$)

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad M_X(s) = (1 - p + pe^s)^n.$$

Geometric(p) ($k = 1, 2, \dots$)

$$p_X(k) = p(1-p)^{k-1}, \quad M_X(s) = \frac{pe^s}{1 - (1-p)e^s}.$$

Poisson(λ) ($k = 0, 1, \dots$)

$$p_X(k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad M_X(s) = e^{\lambda(e^s - 1)}.$$

Uniform(a, b) ($k = a, a+1, \dots, b$)

$$p_X(k) = \frac{1}{b-a+1}, \quad M_X(s) = \frac{e^{sa}(e^{s(b-a+1)} - 1)}{(b-a+1)(e^s - 1)}.$$

Transforms for Common Continuous Random Variables

Uniform(a, b) ($a \leq x \leq b$)

$$f_X(x) = \frac{1}{b-a}, \quad M_X(s) = \frac{e^{sb} - e^{sa}}{s(b-a)}.$$

Exponential(λ) ($x \geq 0$)

$$f_X(x) = \lambda e^{-\lambda x}, \quad M_X(s) = \frac{\lambda}{\lambda - s}, \quad (s < \lambda).$$

Normal(μ, σ^2) ($-\infty < x < \infty$)

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad M_X(s) = e^{(\sigma^2 s^2/2) + \mu s}.$$

4.5 SUM OF A RANDOM NUMBER OF INDEPENDENT RANDOM VARIABLES

Properties of the Sum of a Random Number of Independent Random Variables

Let X_1, X_2, \dots be identically distributed random variables with mean $\mathbf{E}[X]$ and variance $\text{var}(X)$. Let N be a random variable that takes nonnegative integer values. We assume that all of these random variables are independent, and we consider the sum

$$Y = X_1 + \dots + X_N.$$

Then:

- $\mathbf{E}[Y] = \mathbf{E}[N] \mathbf{E}[X]$.
- $\text{var}(Y) = \mathbf{E}[N] \text{var}(X) + (\mathbf{E}[X])^2 \text{var}(N)$.
- We have

$$M_Y(s) = M_N(\log M_X(s)).$$

Equivalently, the transform $M_Y(s)$ is found by starting with the transform $M_N(s)$ and replacing each occurrence of e^s with $M_X(s)$.

4.6 SUMMARY AND DISCUSSION

5

Limit Theorems

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

5.1. Markov and Chebyshev Inequalities	p. 265
5.2. The Weak Law of Large Numbers	p. 269
5.3. Convergence in Probability	p. 271
5.4. The Central Limit Theorem	p. 273
5.5. The Strong Law of Large Numbers	p. 280
5.6. Summary and Discussion	p. 282
Problems	p. 284

5.1 MARKOV AND CHEBYSHEV INEQUALITIES

Markov Inequality

If a random variable X can only take nonnegative values, then

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}, \quad \text{for all } a > 0.$$

Chebyshev Inequality

If X is a random variable with mean μ and variance σ^2 , then

$$\mathbf{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}, \quad \text{for all } c > 0.$$

5.2 THE WEAK LAW OF LARGE NUMBERS

The Weak Law of Large Numbers

Let X_1, X_2, \dots be independent identically distributed random variables with mean μ . For every $\epsilon > 0$, we have

$$\mathbf{P}(|M_n - \mu| \geq \epsilon) = \mathbf{P}\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| \geq \epsilon\right) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

5.3 CONVERGENCE IN PROBABILITY**Convergence of a Deterministic Sequence**

Let a_1, a_2, \dots be a sequence of real numbers, and let a be another real number. We say that the sequence a_n converges to a , or $\lim_{n \rightarrow \infty} a_n = a$, if for every $\epsilon > 0$ there exists some n_0 such that

$$|a_n - a| \leq \epsilon, \quad \text{for all } n \geq n_0.$$

Convergence in Probability

Let Y_1, Y_2, \dots be a sequence of random variables (not necessarily independent), and let a be a real number. We say that the sequence Y_n **converges to a in probability**, if for every $\epsilon > 0$, we have

$$\lim_{n \rightarrow \infty} \mathbf{P}(|Y_n - a| \geq \epsilon) = 0.$$

5.4 THE CENTRAL LIMIT THEOREM

The Central Limit Theorem

Let X_1, X_2, \dots be a sequence of independent identically distributed random variables with common mean μ and variance σ^2 , and define

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Then, the CDF of Z_n converges to the standard normal CDF

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx,$$

in the sense that

$$\lim_{n \rightarrow \infty} \mathbf{P}(Z_n \leq z) = \Phi(z), \quad \text{for every } z.$$

Normal Approximation Based on the Central Limit Theorem

Let $S_n = X_1 + \cdots + X_n$, where the X_i are independent identically distributed random variables with mean μ and variance σ^2 . If n is large, the probability $\mathbf{P}(S_n \leq c)$ can be approximated by treating S_n as if it were normal, according to the following procedure.

1. Calculate the mean $n\mu$ and the variance $n\sigma^2$ of S_n .
2. Calculate the normalized value $z = (c - n\mu)/\sigma\sqrt{n}$.
3. Use the approximation

$$\mathbf{P}(S_n \leq c) \approx \Phi(z),$$

where $\Phi(z)$ is available from standard normal CDF tables.

De Moivre-Laplace Approximation to the Binomial

If S_n is a binomial random variable with parameters n and p , n is large, and k, l are nonnegative integers, then

$$\mathbf{P}(k \leq S_n \leq l) \approx \Phi\left(\frac{l + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

5.5 THE STRONG LAW OF LARGE NUMBERS**The Strong Law of Large Numbers**

Let X_1, X_2, \dots be a sequence of independent identically distributed random variables with mean μ . Then, the sequence of sample means $M_n = (X_1 + \dots + X_n)/n$ converges to μ , **with probability 1**, in the sense that

$$\mathbf{P}\left(\lim_{n \rightarrow \infty} \frac{X_1 + \dots + X_n}{n} = \mu\right) = 1.$$

Convergence with Probability 1

Let Y_1, Y_2, \dots be a sequence of random variables (not necessarily independent). Let c be a real number. We say that Y_n converges to c **with probability 1** (or **almost surely**) if

$$\mathbf{P}\left(\lim_{n \rightarrow \infty} Y_n = c\right) = 1.$$

5.6 SUMMARY AND DISCUSSION

6

The Bernoulli and Poisson Processes

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

6.1. The Bernoulli Process	p. 297
6.2. The Poisson Process	p. 309
6.3. Summary and Discussion	p. 324
Problems	p. 326

6.1 THE BERNOULLI PROCESS

Some Random Variables Associated with the Bernoulli Process and their Properties

- **The binomial with parameters p and n .** This is the number S of successes in n independent trials. Its PMF, mean, and variance are

$$ps(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n,$$

$$\mathbf{E}[S] = np, \quad \text{var}(S) = np(1-p).$$

- **The geometric with parameter p .** This is the number T of trials up to (and including) the first success. Its PMF, mean, and variance are

$$pt(t) = (1-p)^{t-1} p, \quad t = 1, 2, \dots,$$

$$\mathbf{E}[T] = \frac{1}{p}, \quad \text{var}(T) = \frac{1-p}{p^2}.$$

Independence Properties of the Bernoulli Process

- For any given time n , the sequence of random variables X_{n+1}, X_{n+2}, \dots (the future of the process) is also a Bernoulli process, and is independent from X_1, \dots, X_n (the past of the process).
- Let n be a given time and let \bar{T} be the time of the first success after time n . Then, $\bar{T} - n$ has a geometric distribution with parameter p , and is independent of the random variables X_1, \dots, X_n .

Alternative Description of the Bernoulli Process

1. Start with a sequence of independent geometric random variables T_1, T_2, \dots , with common parameter p , and let these stand for the interarrival times.
2. Record a success (or arrival) at times $T_1, T_1 + T_2, T_1 + T_2 + T_3$, etc.

Properties of the k th Arrival Time

- The k th arrival time is equal to the sum of the first k interarrival times

$$Y_k = T_1 + T_2 + \cdots + T_k,$$

and the latter are independent geometric random variables with common parameter p .

- The mean and variance of Y_k are given by

$$\mathbf{E}[Y_k] = \mathbf{E}[T_1] + \cdots + \mathbf{E}[T_k] = \frac{k}{p},$$

$$\text{var}(Y_k) = \text{var}(T_1) + \cdots + \text{var}(T_k) = \frac{k(1-p)}{p^2}.$$

- The PMF of Y_k is given by

$$p_{Y_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \quad t = k, k+1, \dots,$$

and is known as the **Pascal PMF of order k** .

Poisson Approximation to the Binomial

- A Poisson random variable Z with parameter λ takes nonnegative integer values and is described by the PMF

$$p_Z(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

Its mean and variance are given by

$$\mathbf{E}[Z] = \lambda, \quad \text{var}(Z) = \lambda.$$

- For any fixed nonnegative integer k , the binomial probability

$$p_S(k) = \frac{n!}{(n-k)! k!} \cdot p^k (1-p)^{n-k}$$

converges to $p_Z(k)$, when we take the limit as $n \rightarrow \infty$ and $p = \lambda/n$, while keeping λ constant.

- In general, the Poisson PMF is a good approximation to the binomial as long as $\lambda = np$, n is very large, and p is very small.

6.2 THE POISSON PROCESS

Definition of the Poisson Process

An arrival process is called a Poisson process with rate λ if it has the following properties:

- (a) **(Time-homogeneity)** The probability $P(k, \tau)$ of k arrivals is the same for all intervals of the same length τ .
- (b) **(Independence)** The number of arrivals during a particular interval is independent of the history of arrivals outside this interval.
- (c) **(Small interval probabilities)** The probabilities $P(k, \tau)$ satisfy

$$\begin{aligned} P(0, \tau) &= 1 - \lambda\tau + o(\tau), \\ P(1, \tau) &= \lambda\tau + o_1(\tau), \\ P(k, \tau) &= o_k(\tau), \quad \text{for } k = 2, 3, \dots \end{aligned}$$

Here, $o(\tau)$ and $o_k(\tau)$ are functions of τ that satisfy

$$\lim_{\tau \rightarrow 0} \frac{o(\tau)}{\tau} = 0, \quad \lim_{\tau \rightarrow 0} \frac{o_k(\tau)}{\tau} = 0.$$

Random Variables Associated with the Poisson Process and their Properties

- **The Poisson with parameter $\lambda\tau$.** This is the number N_τ of arrivals in a Poisson process with rate λ , over an interval of length τ . Its PMF, mean, and variance are

$$p_{N_\tau}(k) = P(k, \tau) = e^{-\lambda\tau} \frac{(\lambda\tau)^k}{k!}, \quad k = 0, 1, \dots,$$

$$\mathbf{E}[N_\tau] = \lambda\tau, \quad \text{var}(N_\tau) = \lambda\tau.$$

- **The exponential with parameter λ .** This is the time T until the first arrival. Its PDF, mean, and variance are

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0, \quad \mathbf{E}[T] = \frac{1}{\lambda}, \quad \text{var}(T) = \frac{1}{\lambda^2}.$$

Independence Properties of the Poisson Process

- For any given time $t > 0$, the history of the process after time t is also a Poisson process, and is independent from the history of the process until time t .
- Let t be a given time and let \bar{T} be the time of the first arrival after time t . Then, $\bar{T} - t$ has an exponential distribution with parameter λ , and is independent of the history of the process until time t .

Alternative Description of the Poisson Process

1. Start with a sequence of independent exponential random variables T_1, T_2, \dots , with common parameter λ , and let these represent the interarrival times.
2. Record an arrival at times $T_1, T_1 + T_2, T_1 + T_2 + T_3$, etc.

Properties of the k th Arrival Time

- The k th arrival time is equal to the sum of the first k interarrival times

$$Y_k = T_1 + T_2 + \cdots + T_k,$$

and the latter are independent exponential random variables with common parameter λ .

- The mean and variance of Y_k are given by

$$\mathbf{E}[Y_k] = \mathbf{E}[T_1] + \cdots + \mathbf{E}[T_k] = \frac{k}{\lambda},$$

$$\text{var}(Y_k) = \text{var}(T_1) + \cdots + \text{var}(T_k) = \frac{k}{\lambda^2}.$$

- The PDF of Y_k is given by

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0,$$

and is known as the **Erlang PDF of order k** .

Properties of Sums of a Random Number of Random Variables

Let N, X_1, X_2, \dots be independent random variables, where N takes nonnegative integer values. Let $Y = X_1 + \dots + X_N$ for positive values of N , and let $Y = 0$ when $N = 0$.

- If X_i is Bernoulli with parameter p , and N is binomial with parameters m and q , then Y is binomial with parameters m and pq .
- If X_i is Bernoulli with parameter p , and N is Poisson with parameter λ , then Y is Poisson with parameter λp .
- If X_i is geometric with parameter p , and N is geometric with parameter q , then Y is geometric with parameter pq .
- If X_i is exponential with parameter λ , and N is geometric with parameter q , then Y is exponential with parameter λq .

6.3 SUMMARY AND DISCUSSION

7

Markov Chains

Excerpts from Introduction to Probability: Second Edition

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

7.1. Discrete-Time Markov Chains	p. 340
7.2. Classification of States	p. 346
7.3. Steady-State Behavior	p. 352
7.4. Absorption Probabilities and Expected Time to Absorption . .	p. 362
7.5. Continuous-Time Markov Chains	p. 369
7.6. Summary and Discussion	p. 378
Problems	p. 380

7.1 DISCRETE-TIME MARKOV CHAINS

Specification of Markov Models

- A Markov chain model is specified by identifying:
 - (a) the set of states $\S = \{1, \dots, m\}$,
 - (b) the set of possible transitions, namely, those pairs (i, j) for which $p_{ij} > 0$, and,
 - (c) the numerical values of those p_{ij} that are positive.
- The Markov chain specified by this model is a sequence of random variables X_0, X_1, X_2, \dots , that take values in \S , and which satisfy

$$\mathbf{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = p_{ij},$$

for all times n , all states $i, j \in \S$, and all possible sequences i_0, \dots, i_{n-1} of earlier states.

Chapman-Kolmogorov Equation for the n -Step Transition Probabilities

The n -step transition probabilities can be generated by the recursive formula

$$r_{ij}(n) = \sum_{k=1}^m r_{ik}(n-1)p_{kj}, \quad \text{for } n > 1, \text{ and all } i, j,$$

starting with

$$r_{ij}(1) = p_{ij}.$$

7.2 CLASSIFICATION OF STATES

Markov Chain Decomposition

- A Markov chain can be decomposed into one or more recurrent classes, plus possibly some transient states.
- A recurrent state is accessible from all states in its class, but is not accessible from recurrent states in other classes.
- A transient state is not accessible from any recurrent state.
- At least one, possibly more, recurrent states are accessible from a given transient state.

Periodicity

Consider a recurrent class R .

- The class is called **periodic** if its states can be grouped in $d > 1$ disjoint subsets S_1, \dots, S_d , so that all transitions from S_k lead to S_{k+1} (or to S_1 if $k = d$).
- The class is **aperiodic** (not periodic) if and only if there exists a time n such that $r_{ij}(n) > 0$, for all $i, j \in R$.

7.3 STEADY-STATE BEHAVIOR

Steady-State Convergence Theorem

Consider a Markov chain with a single recurrent class, which is aperiodic. Then, the states j are associated with steady-state probabilities π_j that have the following properties.

- (a) For each j , we have

$$\lim_{n \rightarrow \infty} r_{ij}(n) = \pi_j, \quad \text{for all } i.$$

- (b) The π_j are the unique solution to the system of equations below:

$$\begin{aligned}\pi_j &= \sum_{k=1}^m \pi_k p_{kj}, \quad j = 1, \dots, m, \\ 1 &= \sum_{k=1}^m \pi_k.\end{aligned}$$

- (c) We have

$$\begin{aligned}\pi_j &= 0, \quad \text{for all transient states } j, \\ \pi_j &> 0, \quad \text{for all recurrent states } j.\end{aligned}$$

Steady-State Probabilities as Expected State Frequencies

For a Markov chain with a single class which is aperiodic, the steady-state probabilities π_j satisfy

$$\pi_j = \lim_{n \rightarrow \infty} \frac{v_{ij}(n)}{n},$$

where $v_{ij}(n)$ is the expected value of the number of visits to state j within the first n transitions, starting from state i .

Expected Frequency of a Particular Transition

Consider n transitions of a Markov chain with a single class which is aperiodic, starting from a given initial state. Let $q_{jk}(n)$ be the expected number of such transitions that take the state from j to k . Then, regardless of the initial state, we have

$$\lim_{n \rightarrow \infty} \frac{q_{jk}(n)}{n} = \pi_j p_{jk}.$$

7.4 ABSORPTION PROBABILITIES AND EXPECTED TIME TO ABSORPTION

Absorption Probability Equations

Consider a Markov chain where each state is either transient or absorbing, and fix a particular absorbing state s . Then, the probabilities a_i of eventually reaching state s , starting from i , are the unique solution to the equations

$$\begin{aligned} a_s &= 1, \\ a_i &= 0, \quad \text{for all absorbing } i \neq s, \\ a_i &= \sum_{j=1}^m p_{ij} a_j, \quad \text{for all transient } i. \end{aligned}$$

Equations for the Expected Times to Absorption

Consider a Markov chain where all states are transient, except for a single absorbing state. The expected times to absorption, μ_1, \dots, μ_m , are the unique solution to the equations

$$\begin{aligned}\mu_i &= 0, && \text{if } i \text{ is the absorbing state,} \\ \mu_i &= 1 + \sum_{j=1}^m p_{ij} \mu_j, && \text{if } i \text{ is transient.}\end{aligned}$$

Equations for Mean First Passage and Recurrence Times

Consider a Markov chain with a single recurrent class, and let s be a particular recurrent state.

- The mean first passage times μ_i to reach state s starting from i , are the unique solution to the system of equations

$$\mu_s = 0, \quad \mu_i = 1 + \sum_{j=1}^m p_{ij} \mu_j, \quad \text{for all } i \neq s.$$

- The mean recurrence time μ_s^* of state s is given by

$$\mu_s^* = 1 + \sum_{j=1}^m p_{sj} \mu_j.$$

7.5 CONTINUOUS-TIME MARKOV CHAINS**Continuous-Time Markov Chain Assumptions**

- If the current state is i , the time until the next transition is exponentially distributed with a given parameter ν_i , independent of the past history of the process and of the next state.
- If the current state is i , the next state will be j with a given probability p_{ij} , independent of the past history of the process and of the time until the next transition.

Alternative Description of a Continuous-Time Markov Chain

Given the current state i of a continuous-time Markov chain, and for any $j \neq i$, the state δ time units later is equal to j with probability

$$q_{ij}\delta + o(\delta),$$

independent of the past history of the process.

Steady-State Convergence Theorem

Consider a continuous-time Markov chain with a single recurrent class. Then, the states j are associated with steady-state probabilities π_j that have the following properties.

- (a) For each j , we have

$$\lim_{t \rightarrow \infty} \mathbf{P}(X(t) = j \mid X(0) = i) = \pi_j, \quad \text{for all } i.$$

- (b) The π_j are the unique solution to the system of equations below:

$$\begin{aligned} \pi_j \sum_{k \neq j} q_{jk} &= \sum_{k \neq j} \pi_k q_{kj}, \quad j = 1, \dots, m, \\ 1 &= \sum_{k=1}^m \pi_k. \end{aligned}$$

- (c) We have

$$\begin{aligned} \pi_j &= 0, && \text{for all transient states } j, \\ \pi_j &> 0, && \text{for all recurrent states } j. \end{aligned}$$

7.6 SUMMARY AND DISCUSSION

Bayesian Statistical Inference

Excerpts from **Introduction to Probability: Second Edition**

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

8.1. Bayesian Inference and the Posterior Distribution	p. 412
8.2. Point Estimation, Hypothesis Testing, and the MAP Rule	p. 420
8.3. Bayesian Least Mean Squares Estimation	p. 430
8.4. Bayesian Linear Least Mean Squares Estimation	p. 437
8.5. Summary and Discussion	p. 444
Problems	p. 446

Major Terms, Problems, and Methods in this Chapter

- **Bayesian statistics** treats unknown parameters as random variables with known prior distributions.
- In **parameter estimation**, we want to generate estimates that are close to the true values of the parameters in some probabilistic sense.
- In **hypothesis testing**, the unknown parameter takes one of a finite number of values, corresponding to competing hypotheses; we want to choose one of the hypotheses, aiming to achieve a small probability of error.
- Principal Bayesian inference methods:
 - (a) **Maximum a posteriori probability** (MAP) rule: Out of the possible parameter values/hypotheses, select one with maximum conditional/posterior probability given the data (Section 8.2).
 - (b) **Least mean squares** (LMS) estimation: Select an estimator/function of the data that minimizes the mean squared error between the parameter and its estimate (Section 8.3).
 - (c) **Linear least mean squares** estimation: Select an estimator which is a linear function of the data and minimizes the mean squared error between the parameter and its estimate (Section 8.4). This may result in higher mean squared error, but requires simple calculations, based only on the means, variances, and covariances of the random variables involved.

8.1 BAYESIAN INFERENCE AND THE POSTERIOR DISTRIBUTION

Summary of Bayesian Inference

- We start with a prior distribution p_{Θ} or f_{Θ} for the unknown random variable Θ .
- We have a model $p_{X|\Theta}$ or $f_{X|\Theta}$ of the observation vector X .
- After observing the value x of X , we form the posterior distribution of Θ , using the appropriate version of Bayes' rule.

The Four Versions of Bayes' Rule

- Θ discrete, X discrete:

$$p_{\Theta|X}(\theta | x) = \frac{p_\Theta(\theta)p_{X|\Theta}(x | \theta)}{\sum_{\theta'} p_\Theta(\theta')p_{X|\Theta}(x | \theta')}.$$

- Θ discrete, X continuous:

$$p_{\Theta|X}(\theta | x) = \frac{p_\Theta(\theta)f_{X|\Theta}(x | \theta)}{\sum_{\theta'} p_\Theta(\theta')f_{X|\Theta}(x | \theta')}.$$

- Θ continuous, X discrete:

$$f_{\Theta|X}(\theta | x) = \frac{f_\Theta(\theta)p_{X|\Theta}(x | \theta)}{\int f_\Theta(\theta')p_{X|\Theta}(x | \theta') d\theta'}.$$

- Θ continuous, X continuous:

$$f_{\Theta|X}(\theta | x) = \frac{f_\Theta(\theta)f_{X|\Theta}(x | \theta)}{\int f_\Theta(\theta')f_{X|\Theta}(x | \theta') d\theta'}.$$

8.2 POINT ESTIMATION, HYPOTHESIS TESTING, AND THE MAP RULE

The Maximum a Posteriori Probability (MAP) Rule

- Given the observation value x , the MAP rule selects a value $\hat{\theta}$ that maximizes over θ the posterior distribution $p_{\Theta|X}(\theta|x)$ (if Θ is discrete) or $f_{\Theta|X}(\theta|x)$ (if Θ is continuous).
- Equivalently, it selects $\hat{\theta}$ that maximizes over θ :

$$p_{\Theta}(\theta)p_{X|\Theta}(x|\theta) \quad (\text{if } \Theta \text{ and } X \text{ are discrete}),$$

$$p_{\Theta}(\theta)f_{X|\Theta}(x|\theta) \quad (\text{if } \Theta \text{ is discrete and } X \text{ is continuous}),$$

$$f_{\Theta}(\theta)p_{X|\Theta}(x|\theta) \quad (\text{if } \Theta \text{ is continuous and } X \text{ is discrete}),$$

$$f_{\Theta}(\theta)f_{X|\Theta}(x|\theta) \quad (\text{if } \Theta \text{ and } X \text{ are continuous}).$$

- If Θ takes only a finite number of values, the MAP rule minimizes (over all decision rules) the probability of selecting an incorrect hypothesis. This is true for both the unconditional probability of error and the conditional one, given any observation value x .

Point Estimates

- An **estimator** is a random variable of the form $\hat{\Theta} = g(X)$, for some function g . Different choices of g correspond to different estimators.
- An **estimate** is the value $\hat{\theta}$ of an estimator, as determined by the realized value x of the observation X .
- Once the value x of X is observed, the **Maximum a Posteriori Probability (MAP)** estimator, sets the estimate $\hat{\theta}$ to a value that maximizes the posterior distribution over all possible values of θ .
- Once the value x of X is observed, the **Conditional Expectation (LMS)** estimator sets the estimate $\hat{\theta}$ to $\mathbf{E}[\Theta | X = x]$.

The MAP Rule for Hypothesis Testing

- Given the observation value x , the MAP rule selects a hypothesis H_i for which the value of the posterior probability $\mathbf{P}(\Theta = \theta_i | X = x)$ is largest.
- Equivalently, it selects a hypothesis H_i for which $p_{\Theta}(\theta_i)p_{X|\Theta}(x | \theta_i)$ (if X is discrete) or $p_{\Theta}(\theta_i)f_{X|\Theta}(x | \theta_i)$ (if X is continuous) is largest.
- The MAP rule minimizes the probability of selecting an incorrect hypothesis for any observation value x , as well as the probability of error over all decision rules.

8.3 BAYESIAN LEAST MEAN SQUARES ESTIMATION

Key Facts About Least Mean Squares Estimation

- In the absence of any observations, $\mathbf{E}[(\Theta - \hat{\theta})^2]$ is minimized when $\hat{\theta} = \mathbf{E}[\Theta]$:

$$\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta])^2\right] \leq \mathbf{E}\left[(\Theta - \hat{\theta})^2\right], \quad \text{for all } \hat{\theta}.$$

- For any given value x of X , $\mathbf{E}[(\Theta - \hat{\theta})^2 | X = x]$ is minimized when $\hat{\theta} = \mathbf{E}[\Theta | X = x]$:

$$\mathbf{E}\left[\left(\Theta - \mathbf{E}[\Theta | X = x]\right)^2 | X = x\right] \leq \mathbf{E}\left[(\Theta - \hat{\theta})^2 | X = x\right], \quad \text{for all } \hat{\theta}.$$

- Out of all estimators $g(X)$ of Θ based on X , the mean squared estimation error $\mathbf{E}\left[(\Theta - g(X))^2\right]$ is minimized when $g(X) = \mathbf{E}[\Theta | X]$:

$$\mathbf{E}\left[\left(\Theta - \mathbf{E}[\Theta | X]\right)^2\right] \leq \mathbf{E}\left[\left(\Theta - g(X)\right)^2\right], \quad \text{for all estimators } g(X).$$

Properties of the Estimation Error

- The estimation error $\tilde{\Theta}$ is **unbiased**, i.e., it has zero unconditional and conditional mean:

$$\mathbf{E}[\tilde{\Theta}] = 0, \quad \mathbf{E}[\tilde{\Theta} | X = x] = 0, \quad \text{for all } x.$$

- The estimation error $\tilde{\Theta}$ is uncorrelated with the estimate $\hat{\Theta}$:

$$\text{cov}(\hat{\Theta}, \tilde{\Theta}) = 0.$$

- The variance of Θ can be decomposed as

$$\text{var}(\Theta) = \text{var}(\hat{\Theta}) + \text{var}(\tilde{\Theta}).$$

8.4 BAYESIAN LINEAR LEAST MEAN SQUARES ESTIMATION

Linear LMS Estimation Formulas

- The linear LMS estimator $\hat{\Theta}$ of Θ based on X is

$$\hat{\Theta} = \mathbf{E}[\Theta] + \frac{\text{cov}(\Theta, X)}{\text{var}(X)}(X - \mathbf{E}[X]) = \mathbf{E}[\Theta] + \rho \frac{\sigma_\Theta}{\sigma_X}(X - \mathbf{E}[X]),$$

where

$$\rho = \frac{\text{cov}(\Theta, X)}{\sigma_\Theta \sigma_X}$$

is the correlation coefficient.

- The resulting mean squared estimation error is equal to

$$(1 - \rho^2)\sigma_\Theta^2.$$

8.5 SUMMARY AND DISCUSSION

9

Classical Statistical Inference

Excerpts from **Introduction to Probability: Second Edition**

by Dimitri P. Bertsekas and John N. Tsitsiklis ©

Massachusetts Institute of Technology

Contents

9.1. Classical Parameter Estimation	p. 460
9.2. Linear Regression	p. 475
9.3. Binary Hypothesis Testing	p. 485
9.4. Significance Testing	p. 495
9.5. Summary and Discussion	p. 506
Problems	p. 507

Major Terms, Problems, and Methods in this Chapter

- **Classical statistics** treats unknown parameters as constants to be determined. A separate probabilistic model is assumed for each possible value of the unknown parameter.
- In **parameter estimation**, we want to generate estimates that are nearly correct under any possible value of the unknown parameter.
- In **hypothesis testing**, the unknown parameter takes a finite number m of values ($m \geq 2$), corresponding to competing hypotheses; we want to choose one of the hypotheses, aiming to achieve a small probability of error under any of the possible hypotheses.
- In **significance testing**, we want to accept or reject a single hypothesis, while keeping the probability of false rejection suitably small.
- Principal classical inference methods in this chapter:
 - (a) **Maximum likelihood (ML) estimation:** Select the parameter that makes the observed data “most likely,” i.e., maximizes the probability of obtaining the data at hand (Section 9.1).
 - (b) **Linear regression:** Find the linear relation that matches best a set of data pairs, in the sense that it minimizes the sum of the squares of the discrepancies between the model and the data (Section 9.2).
 - (c) **Likelihood ratio test:** Given two hypotheses, select one based on the ratio of their “likelihoods,” so that certain error probabilities are suitably small (Section 9.3).
 - (d) **Significance testing:** Given a hypothesis, reject it if and only if the observed data falls within a certain rejection region. This region is specially designed to keep the probability of false rejection below some threshold (Section 9.4).

9.1 CLASSICAL PARAMETER ESTIMATION

Terminology Regarding Estimators

Let $\hat{\Theta}_n$ be an **estimator** of an unknown parameter θ , that is, a function of n observations X_1, \dots, X_n whose distribution depends on θ .

- The **estimation error**, denoted by $\tilde{\Theta}_n$, is defined by $\tilde{\Theta}_n = \hat{\Theta}_n - \theta$.
- The **bias** of the estimator, denoted by $b_\theta(\hat{\Theta}_n)$, is the expected value of the estimation error:

$$b_\theta(\hat{\Theta}_n) = \mathbf{E}_\theta[\hat{\Theta}_n] - \theta.$$

- The expected value, the variance, and the bias of $\hat{\Theta}_n$ depend on θ , while the estimation error depends in addition on the observations X_1, \dots, X_n .
- We call $\hat{\Theta}_n$ **unbiased** if $\mathbf{E}_\theta[\hat{\Theta}_n] = \theta$, for every possible value of θ .
- We call $\hat{\Theta}_n$ **asymptotically unbiased** if $\lim_{n \rightarrow \infty} \mathbf{E}_\theta[\hat{\Theta}_n] = \theta$, for every possible value of θ .
- We call $\hat{\Theta}_n$ **consistent** if the sequence $\hat{\Theta}_n$ converges to the true value of the parameter θ , in probability, for every possible value of θ .

Maximum Likelihood Estimation

- We are given the realization $x = (x_1, \dots, x_n)$ of a random vector $X = (X_1, \dots, X_n)$, distributed according to a PMF $p_X(x; \theta)$ or PDF $f_X(x; \theta)$.
- The maximum likelihood (ML) estimate is a value of θ that maximizes the likelihood function, $p_X(x; \theta)$ or $f_X(x; \theta)$, over all θ .
- The ML estimate of a one-to-one function $h(\theta)$ of θ is $h(\hat{\theta}_n)$, where $\hat{\theta}_n$ is the ML estimate of θ (the invariance principle).
- When the random variables X_i are i.i.d., and under some mild additional assumptions, each component of the ML estimator is consistent and asymptotically normal.

Estimates of the Mean and Variance of a Random Variable

Let the observations X_1, \dots, X_n be i.i.d., with mean θ and variance v that are unknown.

- The sample mean

$$M_n = \frac{X_1 + \dots + X_n}{n}$$

is an unbiased estimator of θ , and its mean squared error is v/n .

- Two variance estimators are

$$\bar{S}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - M_n)^2, \quad \hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M_n)^2.$$

- The estimator \bar{S}_n^2 coincides with the ML estimator if the X_i are normal. It is biased but asymptotically unbiased. The estimator \hat{S}_n^2 is unbiased. For large n , the two variance estimators essentially coincide.

Confidence Intervals

- A **confidence interval** for a scalar unknown parameter θ is an interval whose endpoints $\hat{\Theta}_n^-$ and $\hat{\Theta}_n^+$ bracket θ with a given high probability.
- $\hat{\Theta}_n^-$ and $\hat{\Theta}_n^+$ are random variables that depend on the observations X_1, \dots, X_n .
- A $1 - \alpha$ confidence interval is one that satisfies

$$\mathbf{P}_\theta(\hat{\Theta}_n^- \leq \theta \leq \hat{\Theta}_n^+) \geq 1 - \alpha,$$

for all possible values of θ .

9.2 LINEAR REGRESSION**Linear Regression**

Given n data pairs (x_i, y_i) , the estimates that minimize the sum of the squared residuals are given by

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x},$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Bayesian Linear Regression

- **Model:**

- (a) We assume a linear relation $Y_i = \Theta_0 + \Theta_1 x_i + W_i$.
- (b) The x_i are modeled as known constants.
- (c) The random variables $\Theta_0, \Theta_1, W_1, \dots, W_n$ are normal and independent.
- (d) The random variables Θ_0 and Θ_1 have mean zero and variances σ_0^2, σ_1^2 , respectively.
- (e) The random variables W_i have mean zero and variance σ^2 .

- **Estimation Formulas:**

Given the data pairs (x_i, y_i) , the MAP estimates of Θ_0 and Θ_1 are

$$\hat{\theta}_1 = \frac{\sigma_1^2}{\sigma^2 + \sigma_1^2 \sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

$$\hat{\theta}_0 = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2} (\bar{y} - \hat{\theta}_1 \bar{x}),$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

9.3 BINARY HYPOTHESIS TESTING

Likelihood Ratio Test (LRT)

- Start with a target value α for the false rejection probability.
- Choose a value for ξ such that the false rejection probability is equal to α :

$$\mathbf{P}(L(X) > \xi; H_0) = \alpha.$$

- Once the value x of X is observed, reject H_0 if $L(x) > \xi$.

Neyman-Pearson Lemma

Consider a particular choice of ξ in the LRT, which results in error probabilities

$$\mathbf{P}(L(X) > \xi; H_0) = \alpha, \quad \mathbf{P}(L(X) \leq \xi; H_1) = \beta.$$

Suppose that some other test, with rejection region R , achieves a smaller or equal false rejection probability:

$$\mathbf{P}(X \in R; H_0) \leq \alpha.$$

Then,

$$\mathbf{P}(X \notin R; H_1) \geq \beta,$$

with strict inequality $\mathbf{P}(X \notin R; H_1) > \beta$ when $\mathbf{P}(X \in R; H_0) < \alpha$.

9.4 SIGNIFICANCE TESTING

Significance Testing Methodology

A statistical test of a hypothesis H_0 is to be performed, based on the observations X_1, \dots, X_n .

- The following steps are carried out before the data are observed.
 - (a) Choose a **statistic** S , that is, a scalar random variable that will summarize the data to be obtained. Mathematically, this involves the choice of a function $h : \Re^n \rightarrow \Re$, resulting in the statistic $S = h(X_1, \dots, X_n)$.
 - (b) Determine the **shape of the rejection region** by specifying the set of values of S for which H_0 will be rejected as a function of a yet undetermined critical value ξ .
 - (c) Choose the **significance level**, i.e., the desired probability α of a false rejection of H_0 .
 - (d) Choose the **critical value** ξ so that the probability of false rejection is equal (or approximately equal) to α . At this point, the rejection region is completely determined.
- Once the values x_1, \dots, x_n of X_1, \dots, X_n are observed:
 - (i) Calculate the value $s = h(x_1, \dots, x_n)$ of the statistic S .
 - (ii) Reject the hypothesis H_0 if s belongs to the rejection region.

The Chi-Square Test:

- Use the statistic

$$S = \sum_{k=1}^m N_k \log \left(\frac{N_k}{n\theta_k^*} \right)$$

(or possibly the related statistic T) and a rejection region of the form

reject H_0 if $2S > \gamma$

(or $T > \gamma$, respectively).

- The critical value γ is determined from the CDF tables for the χ^2 distribution with $m - 1$ degrees of freedom so that

$$\mathbf{P}(2S > \gamma; H_0) = \alpha,$$

where α is a given significance level.

9.5 SUMMARY AND DISCUSSION

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability

John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.