LECTURE 1: Probability models and axioms

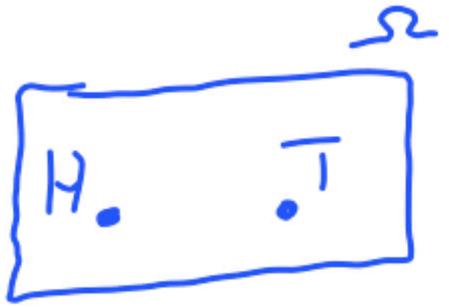
- Sample space
- Probability laws
  - Axioms
  - Properties that follow from the axioms
- Examples
  - Discrete
  - Continuous
- Discussion
  - Countable additivity
  - Mathematical subtleties
- Interpretations of probabilities

Sample space

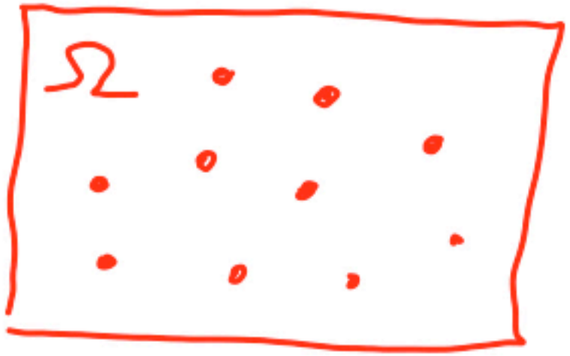
- Two steps:
  - Describe possible outcomes
  - Describe beliefs about likelihood of outcomes

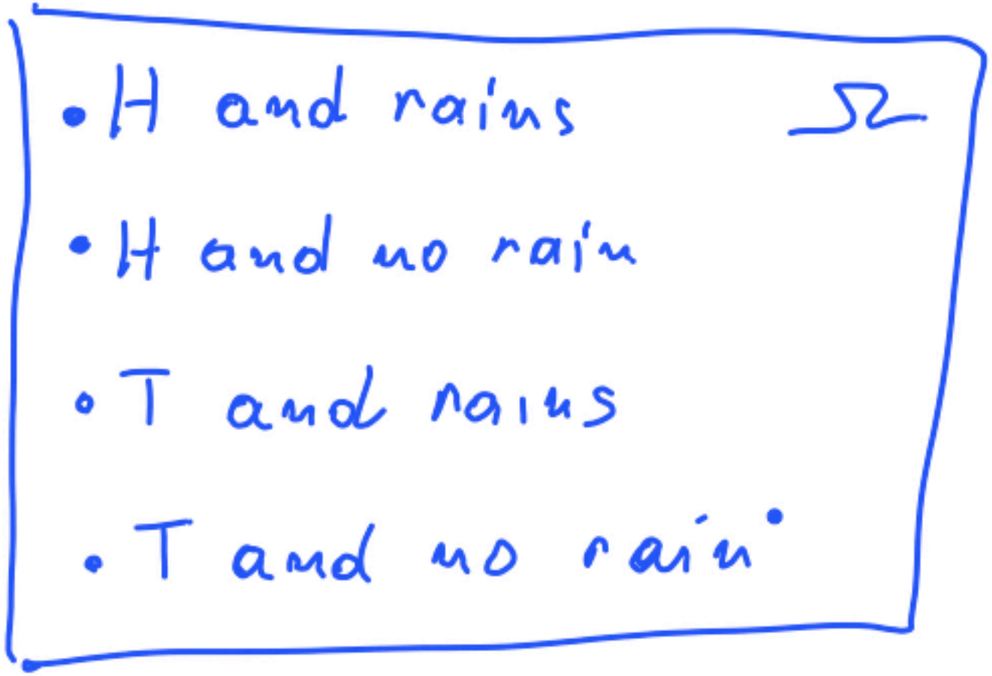
Sample space

- List (set) of possible outcomes, Ω



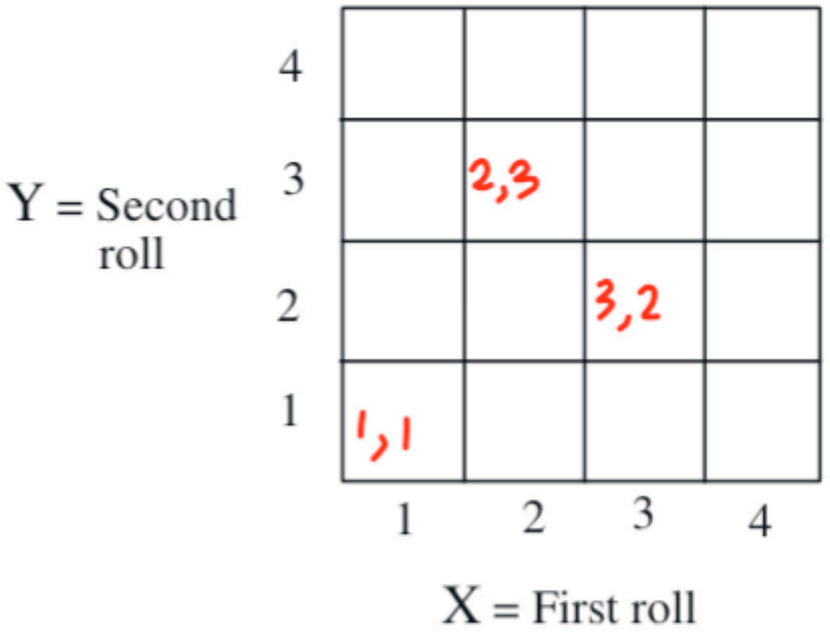
- List must be:
  - Mutually exclusive
  - Collectively exhaustive
  - At the "right" granularity



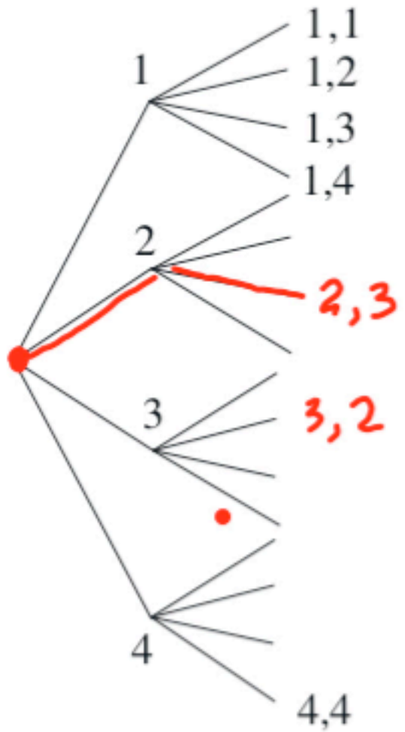


Sample space: discrete/finite example

- Two rolls of a tetrahedral die



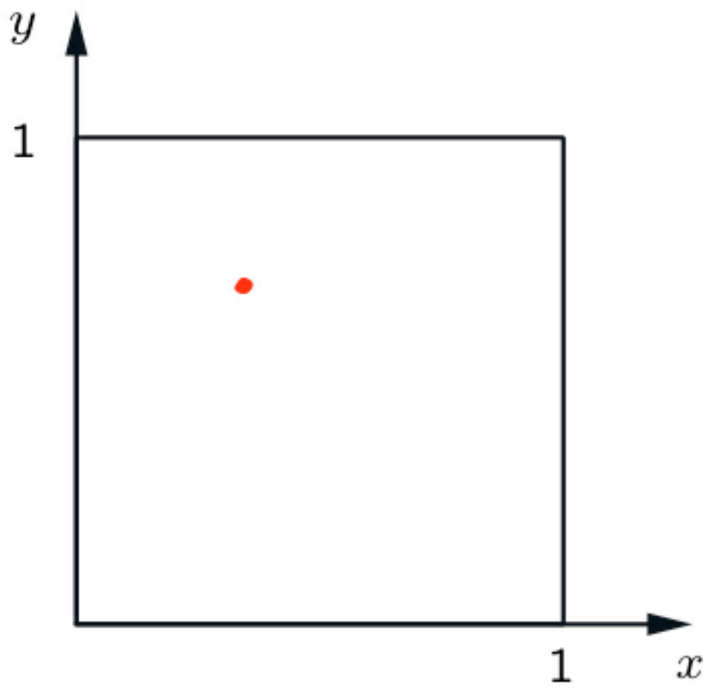
sequential description



Tree

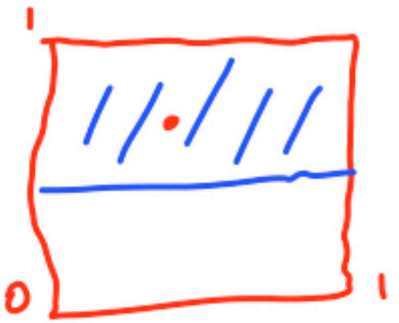
Sample space: continuous example

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

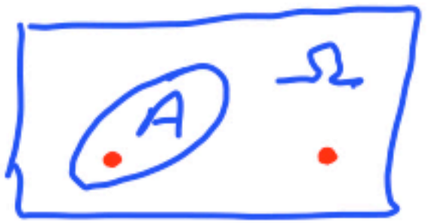


Probability axioms

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



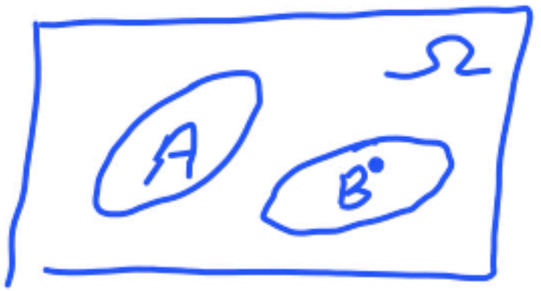
$\mathbb{P}(A)$



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

$A \cap B$

$A \cup B$



∅
empty set

Some simple consequences of the axioms

Axioms

$\mathbf{P}(A) \geq 0$

$\mathbf{P}(\Omega) = 1$

For disjoint events:

$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$

Consequences

$\mathbf{P}(A) \leq 1$

$\mathbf{P}(\varnothing) = 0$

$\mathbf{P}(A) + \mathbf{P}(A^c) = 1$

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

and similarly for k disjoint events

$$\mathbf{P}(\{s_1, s_2, \ldots, s_k\}) = \mathbf{P}(\{s_1\}) + \cdots + \mathbf{P}(\{s_k\})$$

$$= \mathbf{P}(s_1) + \cdots + \mathbf{P}(s_k)$$

Some simple consequences of the axioms

Axioms

(a) $\mathrm{P}(A) \geq 0$

(b) $\mathrm{P}(\Omega) = 1$

For disjoint events:

(c) $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$

$A \cup A^c = \Omega$

$A \cap A^c = \emptyset$

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

$\overset{(c)}{=} P(A) + P(A^c)$

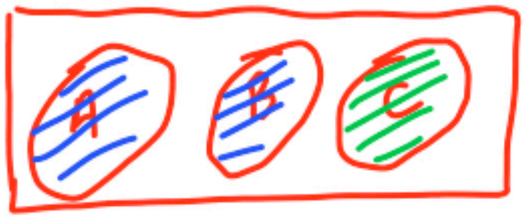
$P(A) = 1 - \underbrace{P(A^c)}_{\geq 0} \overset{(a)}{\leq} 1$

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

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

Some simple consequences of the axioms

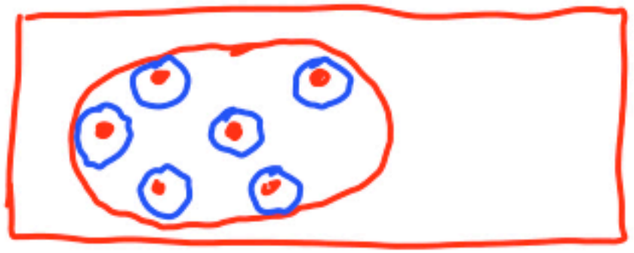
- A, B, C disjoint: $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C)$



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

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

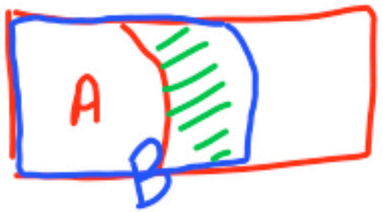
- $\mathbf{P}(\{s_1, s_2, \ldots, s_k\}) =$



$$P(\{s_1\} \cup \{s_2\} \cup \cdots \cup \{s_k\})$$
$$= P(\{s_1\}) + \cdots + P(\{s_k\})$$
$$= P(s_1) + \cdots + P(s_k)$$

More consequences of the axioms

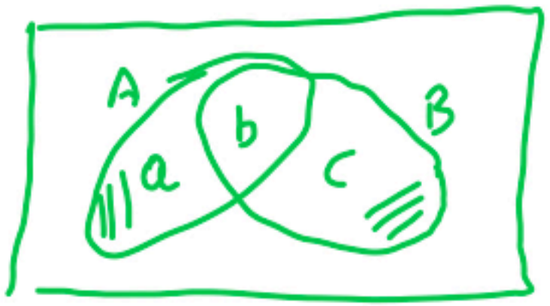
- If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$



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

$P(B) = P(A) + \underbrace{P(B \cap A^c)}_{\geq 0} \geq P(A)$

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



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

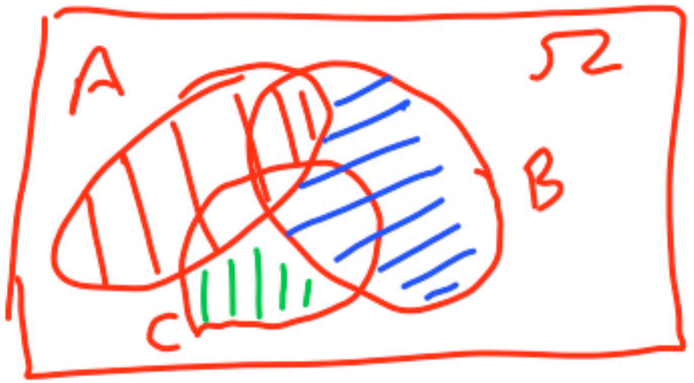
$P(A \cup B) = a + b + c$

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

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

More consequences of the axioms

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

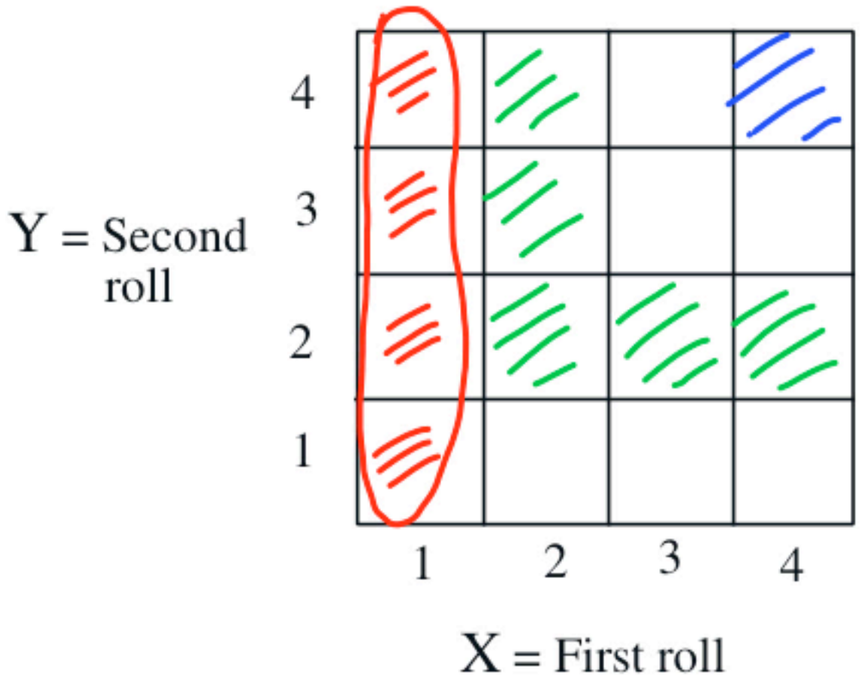


$$P(A \cup B \cup C) =$$

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

Probability calculation: discrete/finite example

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



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

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

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

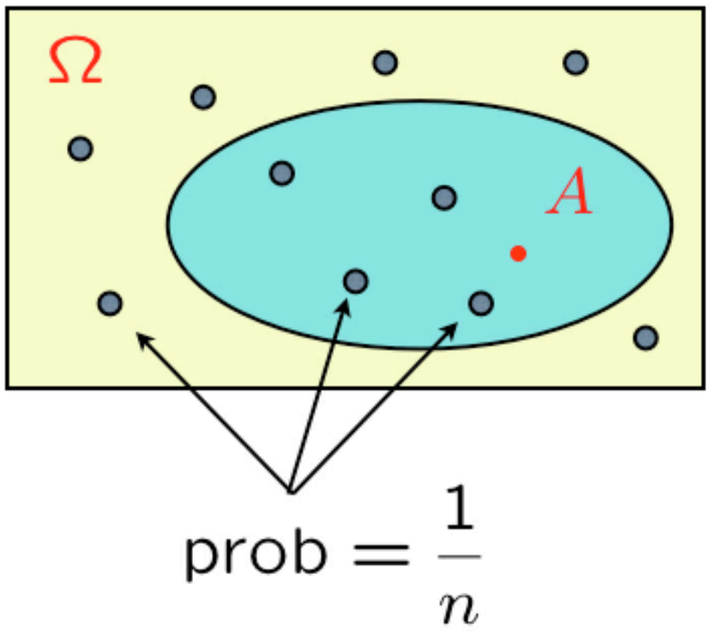
- $\mathbf{P}(Z = 4) = 1/16$
- $\mathbf{P}(Z = 2) = 5 \cdot \frac{1}{16}$.

Discrete uniform law

finite

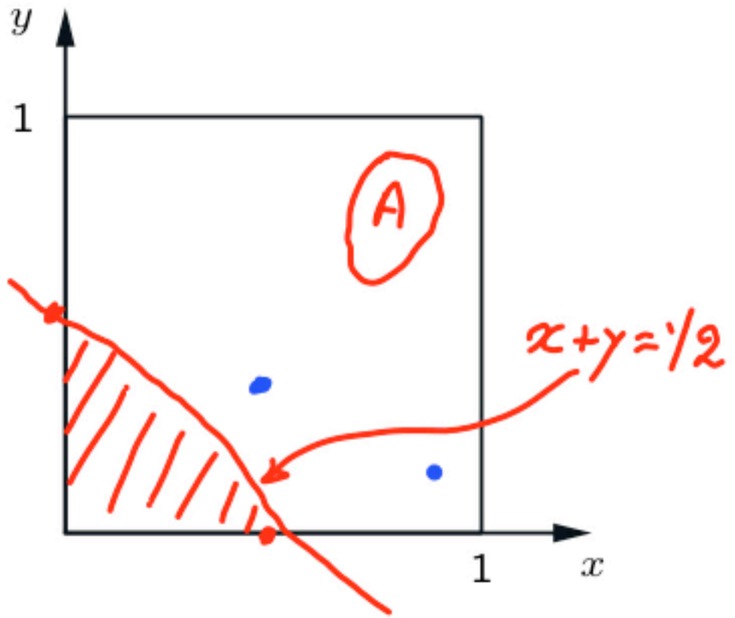
- Assume Ω consists of n equally likely elements
- Assume A consists of k elements

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



Probability calculation: continuous example

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



- **Uniform** probability law: Probability = Area

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

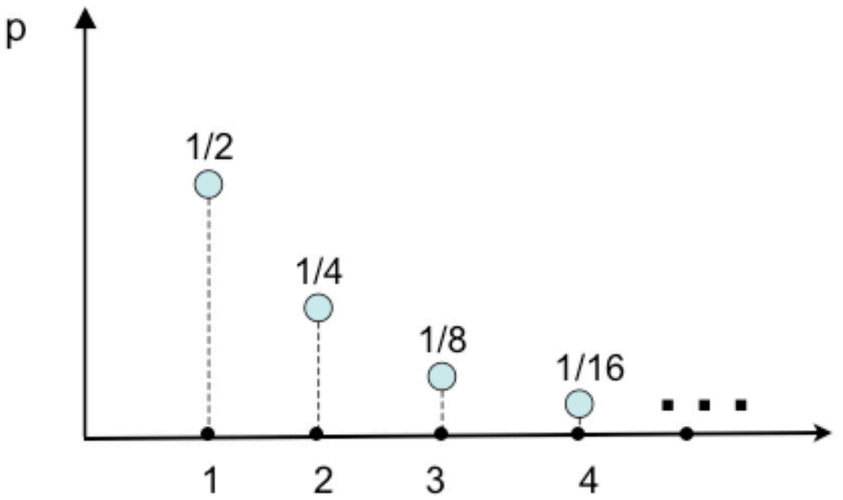
$$\mathbf{P}\big(\{(0.5, 0.3)\}\big) = 0$$

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: $\{1, 2, \ldots\}$
  - We are given $\mathbf{P}(n) = \frac{1}{2^n}, \quad n = 1, 2, \ldots$



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

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

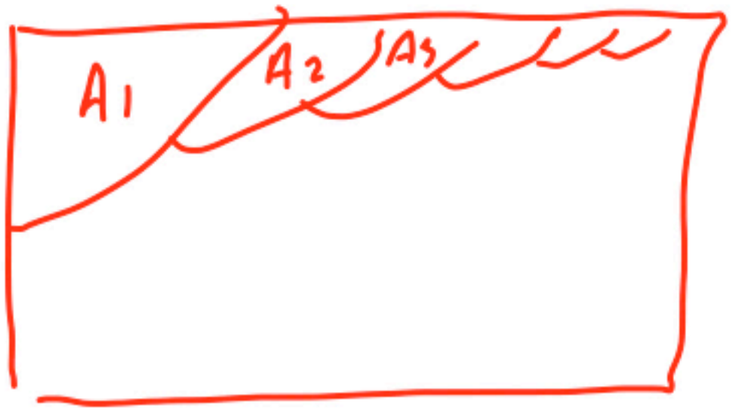
$$= P(\{2\} \cup \{4\} \cup \{6\} \cup \ldots) = P(2) + P(4) + P(6) + \ldots$$

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

Countable additivity axiom

- Strengthens the finite additivity axiom

Countable Additivity Axiom:
If A_1, A_2, A_3,... is an infinite **sequence** of disjoint events,
then $\mathbf{P}(A_1 \cup A_2 \cup A_3 \cup \cdots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3) + \cdots$



Mathematical subtleties

Countable Additivity Axiom:
If A_1, A_2, A_3,... is an infinite **sequence** of disjoint events,
then $\mathbf{P}(A_1 \cup A_2 \cup A_3 \cup \cdots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3) + \cdots$

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

- Additivity holds only for "countable" sequences of events
- The unit square (simlarly, 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

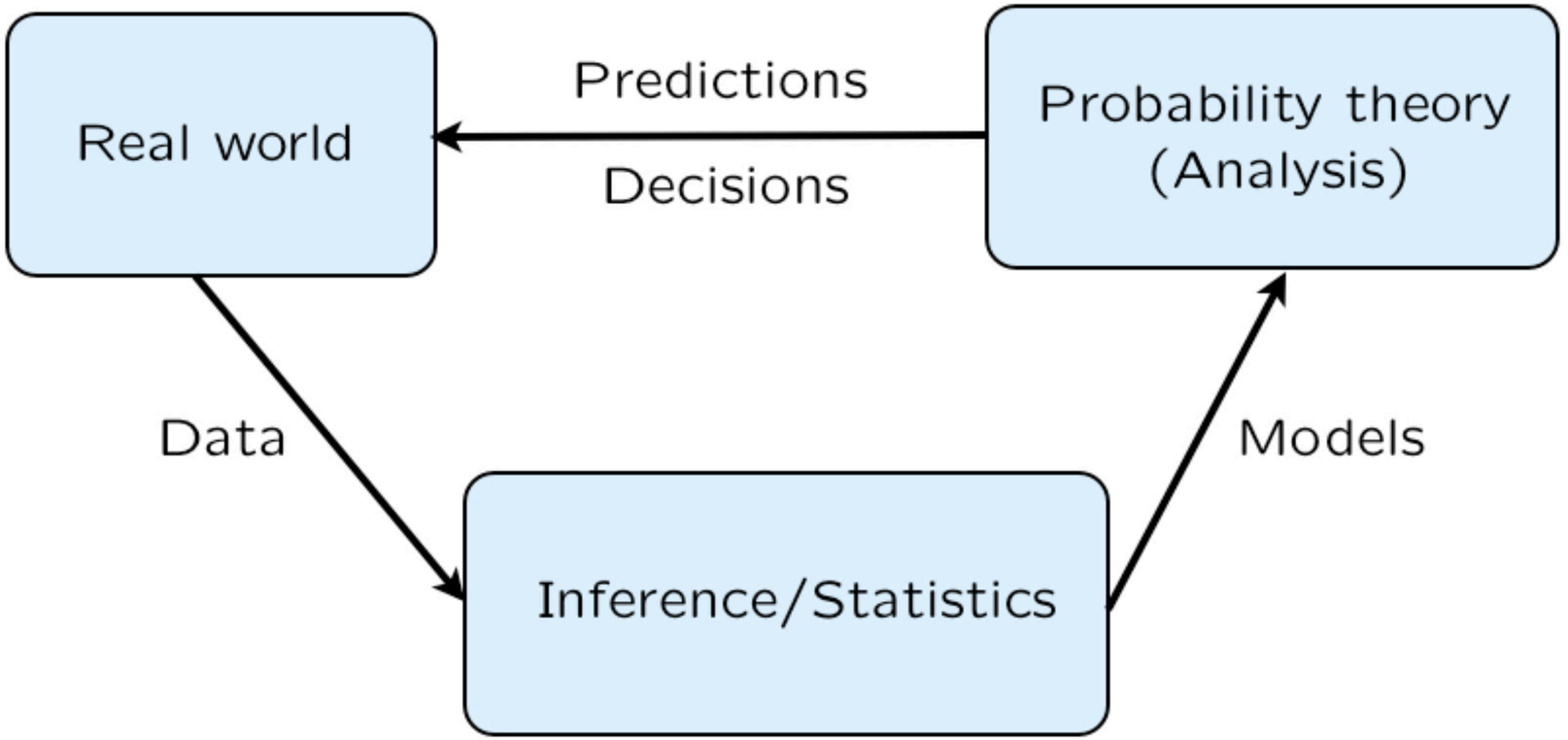
"Thm:" "Frequency" of event A "is" $\mathbf{P}(A)$

- Are probabilities frequencies?
  - $\mathbf{P}(\text{coin toss yields heads}) = 1/2$
  - $\mathbf{P}(\text{the president of } \ldots \text{ will be reelected}) = 0.7$

- Probabilities are often intepreted 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



LECTURE 2: Conditioning and Bayes' rule

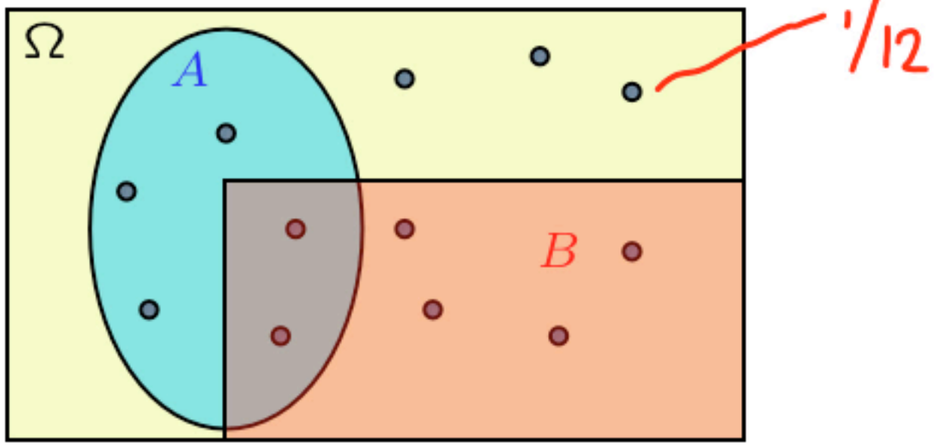
- Conditional probability

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

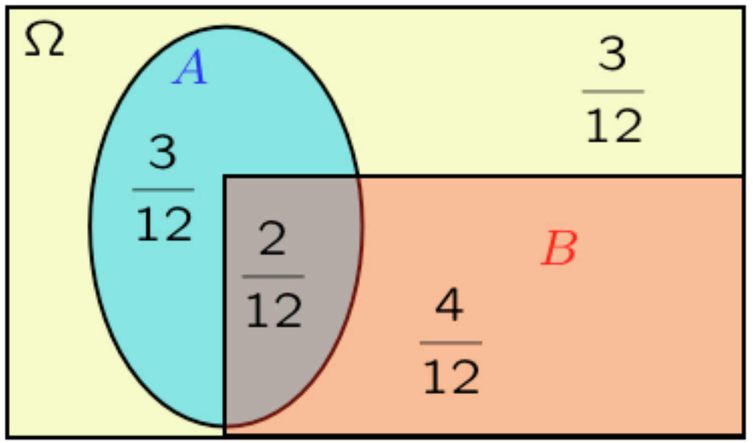
The idea of conditioning

Use new information to revise a model

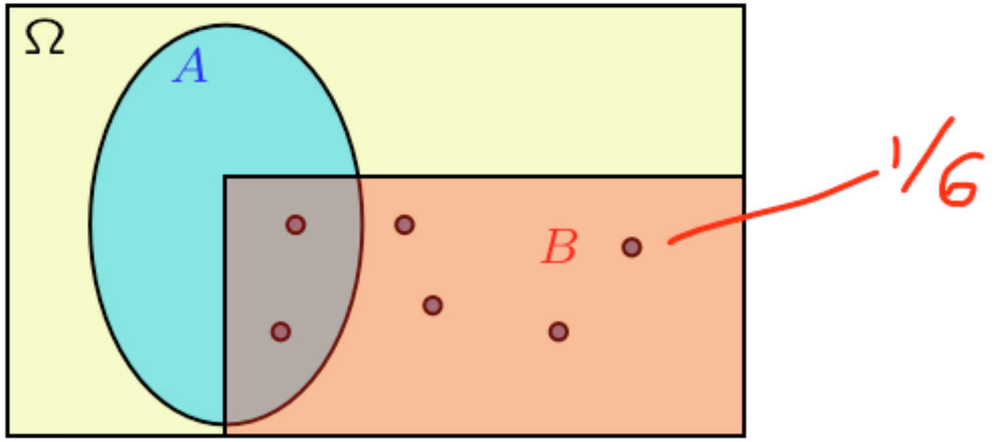
Assume 12 equally likely outcomes



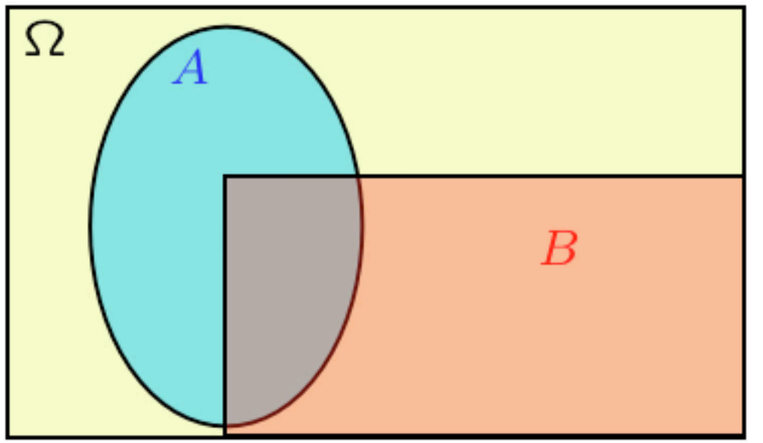
$$\mathbf{P}(A) = \frac{5}{12} \qquad \mathbf{P}(B) = \frac{6}{12}$$



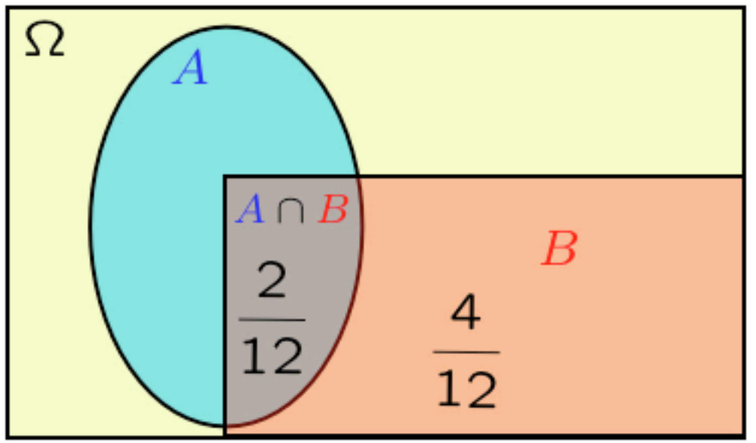
If told B occurred:



$$\mathbf{P}(A \mid B) = \frac{2}{6} = \frac{1}{3} \qquad \mathbf{P}(B \mid B) = 1$$



Definition of conditional probability



- $\mathbf{P}(A \mid B)$ = "probability of A, given that B occurred"

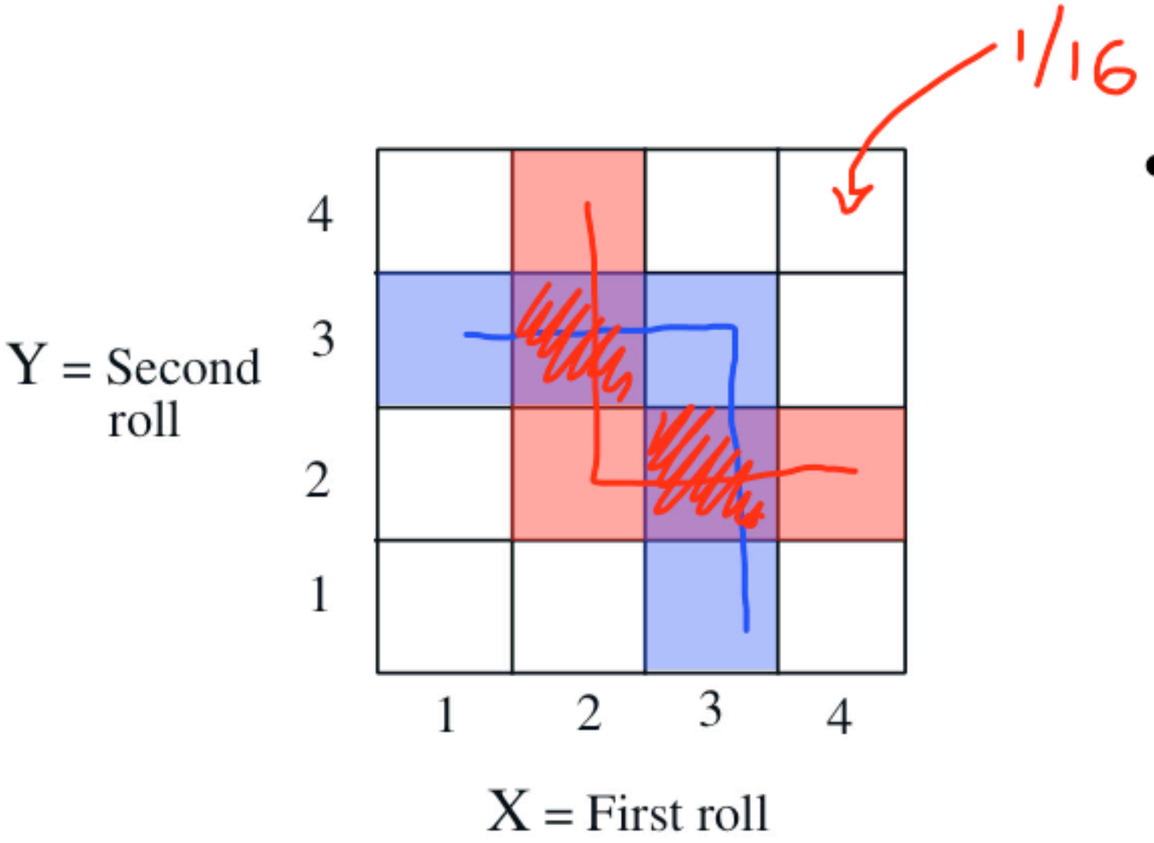
Def. →

$$\mathbf{P}(A \mid B) \triangleq \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$$

defined only when $\mathbf{P}(B) > 0$

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

Example: two rolls of a 4-sided die



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

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

$$\mathbf{P}(M = 1 \mid B) = 0$$

$$\mathbf{P}(M = 3 \mid B) = \frac{P(M=3 \text{ and } B)}{P(B)}$$

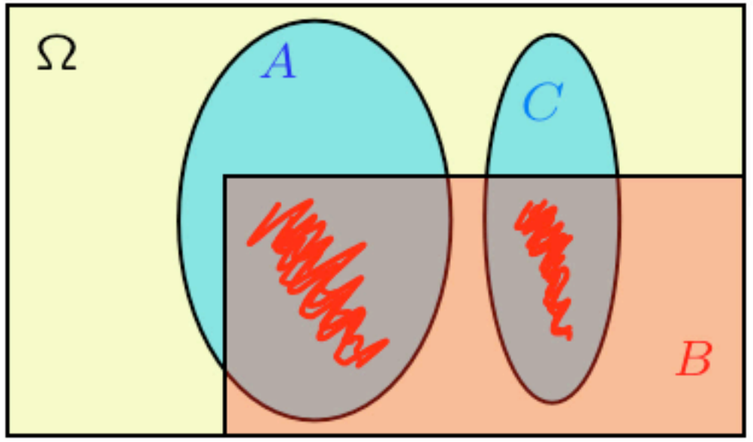
$$= \frac{2/16}{5/16} = \frac{2}{5}$$

Conditional probabilities share properties of ordinary probabilities

$\mathbf{P}(A \mid B) \geq 0$ assuming $\mathbf{P}(B) > 0$

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

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



If $A \cap C = \emptyset$, then $\mathbf{P}(A \cup C \mid B) = \mathbf{P}(A \mid B) + \mathbf{P}(C \mid B)$

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

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

also finite

countable

additivity

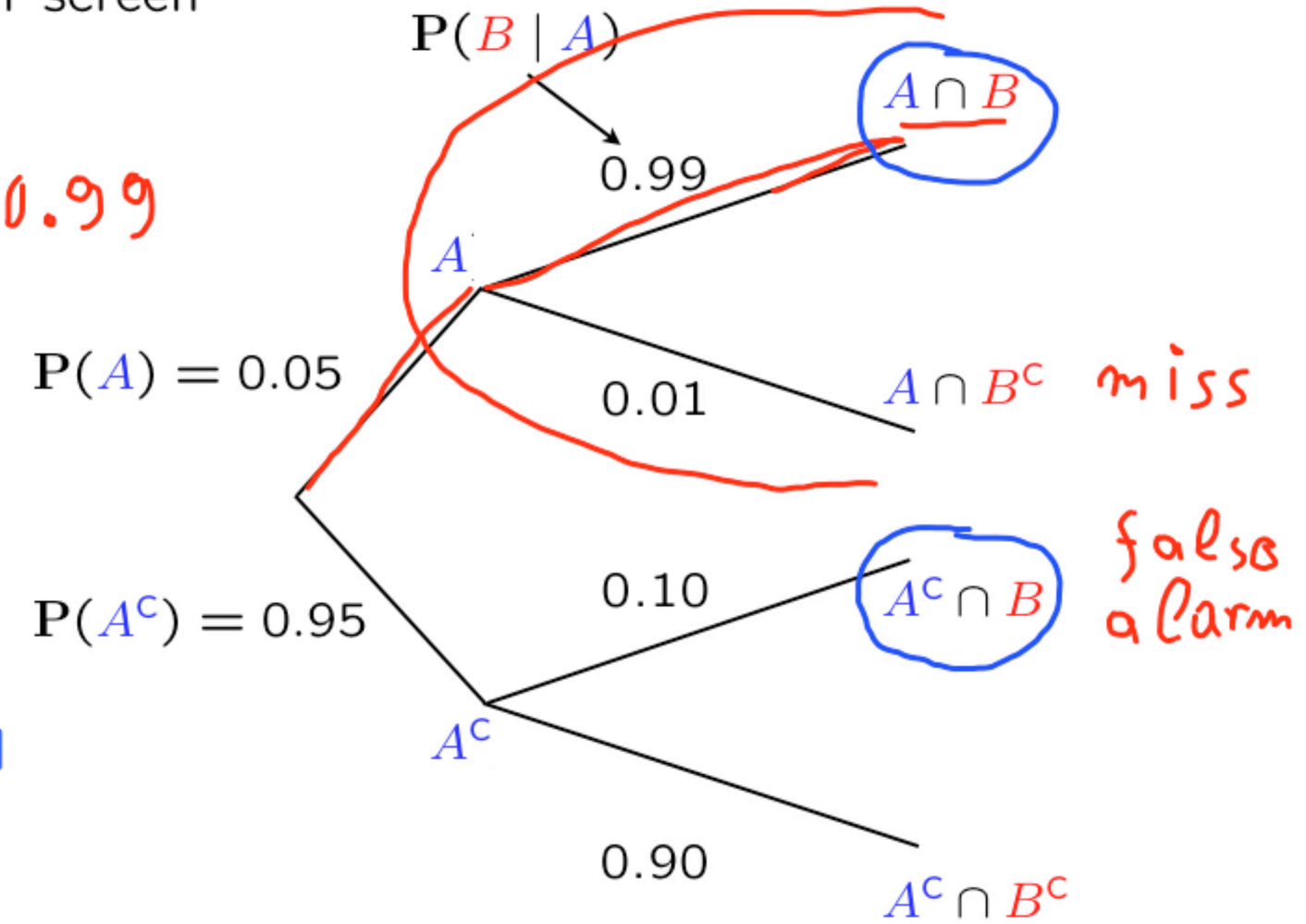
Models based on conditional probabilities

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \qquad \mathbf{P}(B \mid A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)}$$

Event A: Airplane is flying above

Event B: Something registers on radar screen

- $\mathbf{P}(A \cap B) =$ $P(A) \cdot P(B|A) = 0.05 \cdot 0.99$
- $\mathbf{P}(B) =$ $0.05 \cdot 0.99 + 0.95 \cdot 0.1 = 0.1445$
- $\mathbf{P}(A \mid B) =$ $\dfrac{0.05 \cdot 0.99}{0.1445} = 0.34$

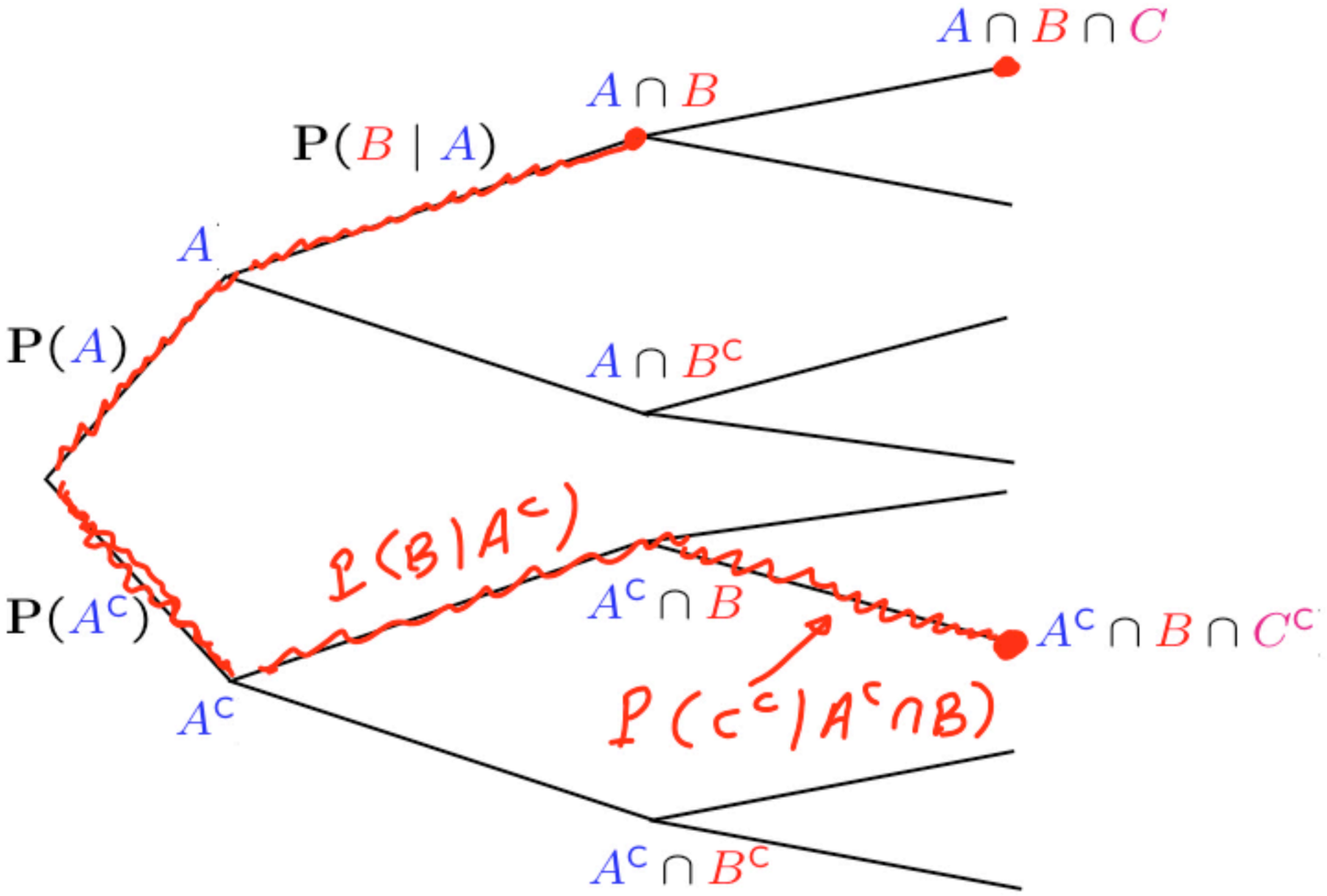


The multiplication rule

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

$$\mathbf{P}(A \cap B) = \mathbf{P}(B)\,\mathbf{P}(A \mid B)$$

$$= \mathbf{P}(A)\,\mathbf{P}(B \mid A)$$



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

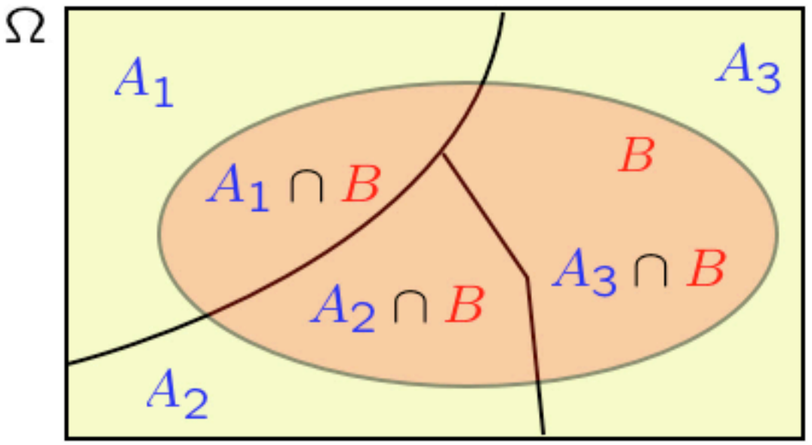
$= \mathbf{P}(A^C \cap B)\,\mathbf{P}(C^C \mid A^C \cap B)$

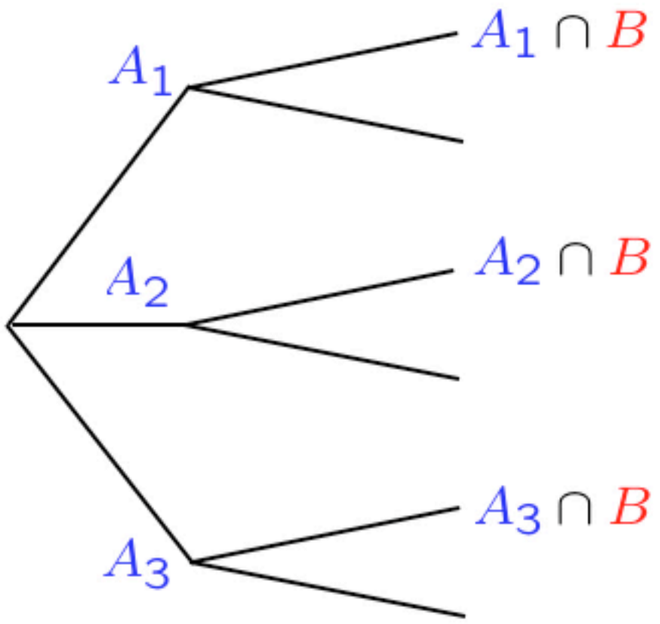
$= \mathbf{P}(A^C) \cdot \mathbf{P}(B \mid A^C)\,\mathbf{P}(C^C \mid A^C \cap B)$

$\mathbf{P}(A_1 \cap A_2 \cap \cdots \cap A_n)$

$= \mathbf{P}(A_1) \prod_{i=2}^{n} \mathbf{P}(A_i \mid A_1 \cap \cdots \cap A_{i-1})$

Total probability theorem





- Partition of sample space into $A_1, A_2, A_3, \ldots$
- Have $\mathbf{P}(A_i)$, for every i
- Have $\mathbf{P}(B \mid A_i)$, for every i

$$\mathbf{P}(B) = \mathbf{P}(B \cap A_1) + \mathbf{P}(B \cap A_2) + \mathbf{P}(B \cap A_3)$$

$$= \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \cdots$$

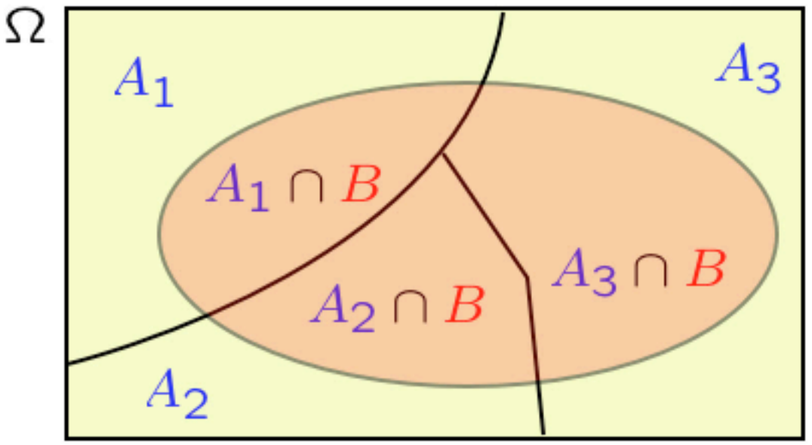
$$\sum_i \mathbf{P}(A_i) = 1$$

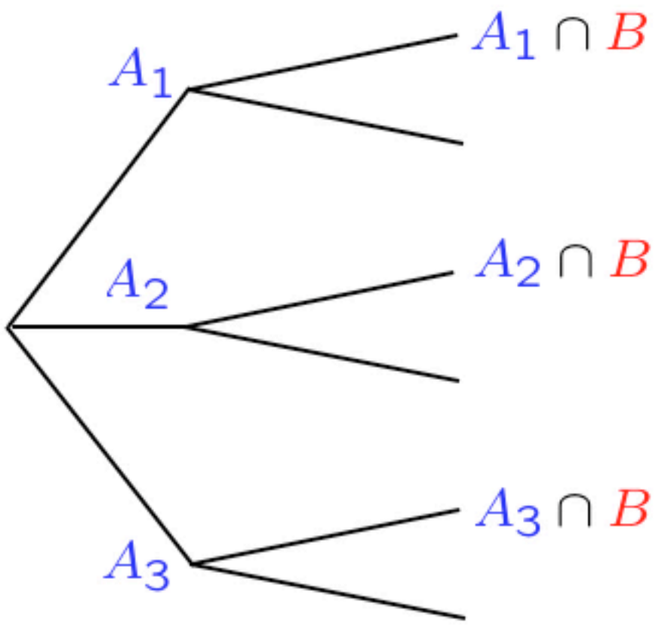
weights

$$\mathbf{P}(B) = \sum_i \mathbf{P}(A_i)\,\mathbf{P}(B \mid A_i)$$

weighted average of $\mathbf{P}(B \mid A_i)$

Bayes' rule





- Partition of sample space into A_1, A_2, A_3
- Have $\mathbf{P}(A_i)$, for every i initial "beliefs"
- Have $\mathbf{P}(B \mid A_i)$, for every i

revised "beliefs," given that B occurred:

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

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

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 $\mathbf{P}(A_i)$ on possible causes of an observed event B
  - model of the world under each A_i: $\mathbf{P}(B \mid A_i)$

$$A_i \xrightarrow[\mathbf{P}(B \mid A_i)]{\text{model}} B$$

  - draw conclusions about causes

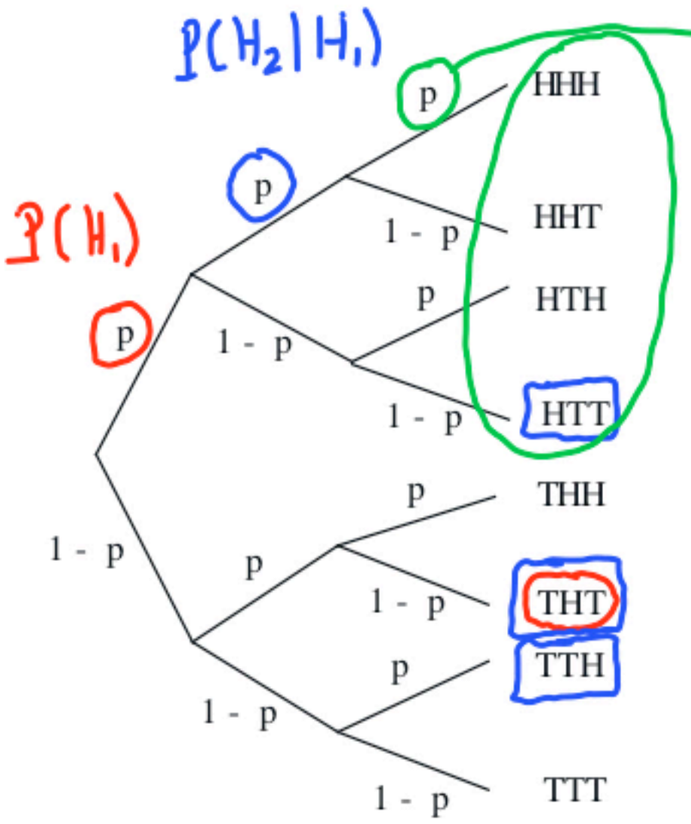
$$B \xrightarrow[\mathbf{P}(A_i \mid B).]{\text{inference}} A_i$$

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: $\mathbf{P}(H) = p$, $\mathbf{P}(T) = 1 - p$



$P(H_2 \mid H_1)$ $P(H_1)$ $P(H_3 \mid H_1 \cap H_2)$

$P(H_2 \mid H_1) = p = P(H_2 \mid T_1)$

$P(H_2) = P(H_1)\,P(H_2 \mid H_1) + P(T_1)\,P(H_2 \mid T_1) = p$

- Multiplication rule: $\mathbf{P}(THT) = (1-p)\,p\,(1-p)$
- Total probability:

  $\mathbf{P}(1 \text{ head}) = 3\,p(1-p)^2$
- Bayes rule:

  $\mathbf{P}(\text{first toss is H} \mid \text{1 head}) = \dfrac{P(H_1 \cap \text{1 Head})}{P(\text{1 Head})}$

$= \dfrac{p(1-p)^2}{3\,p(1-p)^2} = \dfrac{1}{3}$

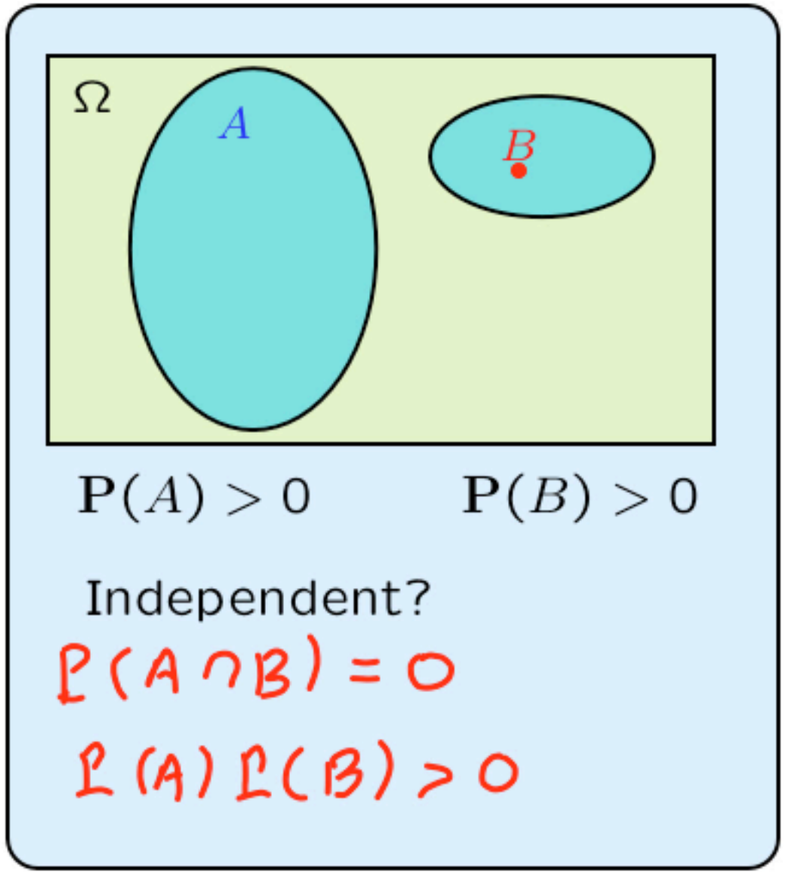
Independence of two events

- **Intuitive "definition":** $\mathbf{P}(B \mid A) = \mathbf{P}(B)$
  - 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: $\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B)$

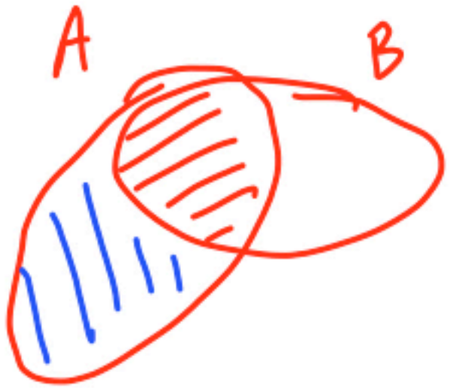
- Symmetric with respect to A and B
- implies $\mathbf{P}(A \mid B) = \mathbf{P}(A)$
- applies even if $\mathbf{P}(A) = 0$



Independence of event complements

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

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



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

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

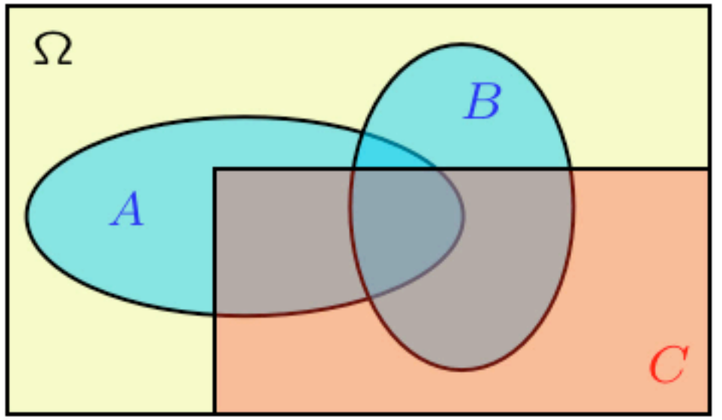
$$= P(A)P(B) + P(A \cap B^c)$$

$$P(A \cap B^c) = P(A) - P(A)P(B) = P(A)(1 - P(B))$$

$$= P(A)P(B^c)$$

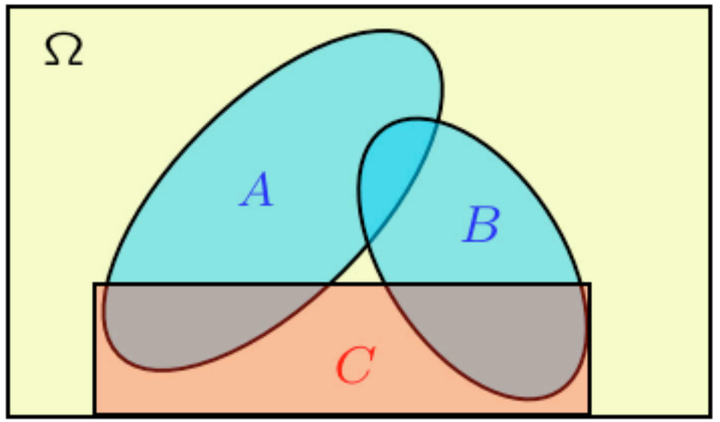
Conditional independence

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



$$P(A \cap B \mid C) = P(A \mid C)\, P(B \mid C)$$

Assume A and B are independent



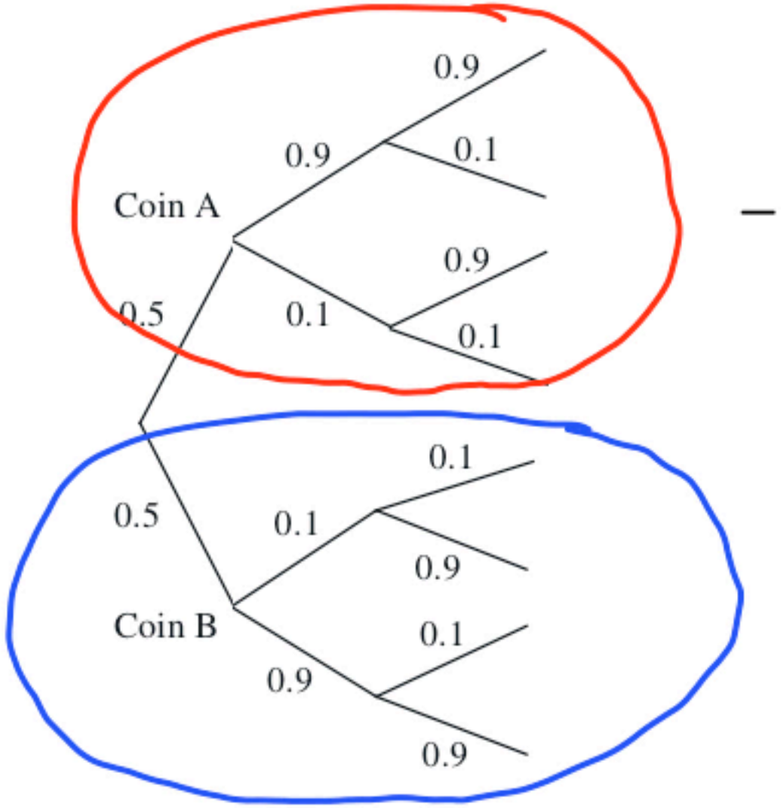
- If we are told that C occurred,
  are A and B independent? No

Conditioning may affect independence

- Two unfair coins, A and B:
  $\mathbf{P}(H \mid \text{coin } A) = 0.9$, $\mathbf{P}(H \mid \text{coin } B) = 0.1$

- choose either coin with equal probability

given a coin: independent tosses



- Are coin tosses independent?

No!

– Compare:

$\mathbf{P}(\text{toss } 11 = H) = P(A)P(H_{11}|A) + P(B)P(H_{11}|B)$

$= 0.5 \times 0.9 + 0.5 \times 0.1 = 0.5$

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

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

Independence of a collection of events

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

$A_1, A_2, \ldots$ indep $\Rightarrow \mathbb{P}(A_3 \cap A_4^c) = \mathbb{P}(A_3 \cap A_4^c \mid A_1 \cup (A_2 \cap A_5^c))$.

$\mathbb{P}(A_3) = \mathbb{P}(A_3 \mid A_1 \cap A_2) = \mathbb{P}(A_3 \mid A_1 \cap A_2^c) = \mathbb{P}(A_3 \mid A_1^c \cap A_2)$

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

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

$n = 3$:

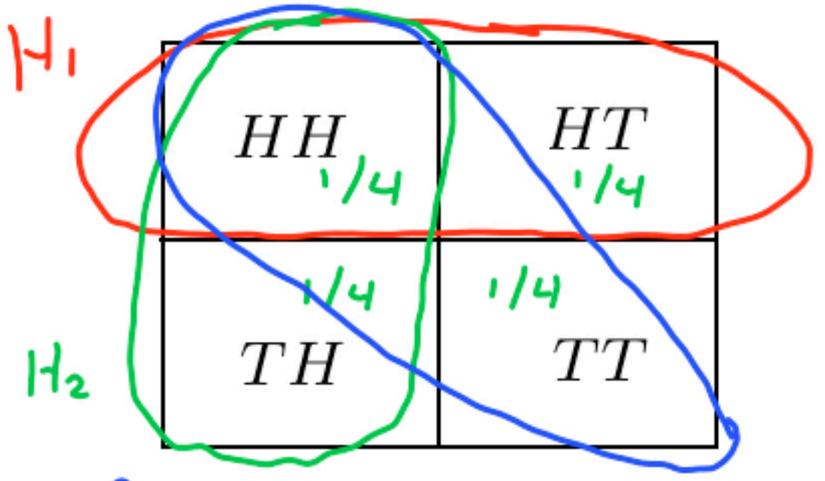
$$\left.\begin{aligned} \mathbf{P}(A_1 \cap A_2) &= \mathbf{P}(A_1) \cdot \mathbf{P}(A_2) \\ \mathbf{P}(A_1 \cap A_3) &= \mathbf{P}(A_1) \cdot \mathbf{P}(A_3) \\ \mathbf{P}(A_2 \cap A_3) &= \mathbf{P}(A_2) \cdot \mathbf{P}(A_3) \end{aligned}\right\}$$ pairwise independence

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1) \cdot \mathbf{P}(A_2) \cdot \mathbf{P}(A_3)$$

Independence vs. pairwise independence

- Two independent fair coin tosses
  - H_1: First toss is H
  - H_2: Second toss is H

$$\mathbf{P}(H_1) = \mathbf{P}(H_2) = 1/2$$



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

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

H_2, C: indep.

$P(H_1 \cap H_2 \cap C) = P(HH) = 1/4$

$P(H_1)P(H_2)P(C) = 1/8$ $\Big\rangle$ diff.

$P(C \mid H_1) = P(H_2 \mid H_1) = P(H_2) = 1/2 = P(C)$

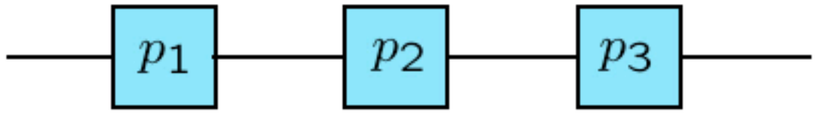
$P(C \mid H_1 \cap H_2) = 1 \neq P(C) = 1/2$

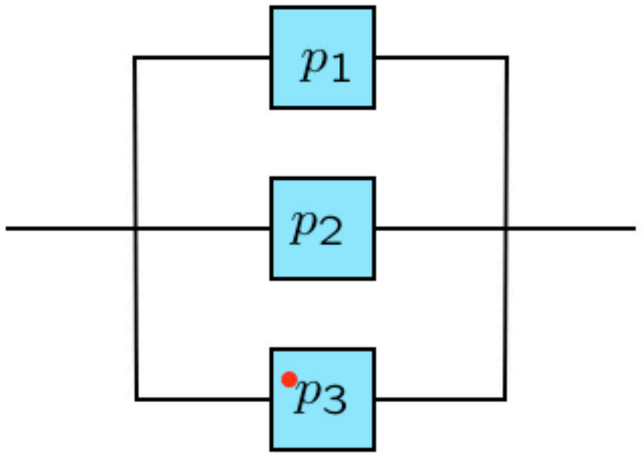
H_1, H_2, and C are pairwise independent, but not independent

Reliability

p_i: probability that unit i is "up"

independent units





U_i : ith unit up

$U_1, U_2, \ldots, U_n$ independent

F_i : ith unit down

$\Rightarrow F_i$ independent

probability that system is "up"?

$P(\text{system up}) = P(U_1 \cap U_2 \cap U_3)$

$= P(U_1)\,P(U_2)\,P(U_3) = p_1 p_2 p_3$

$P(\text{system is up}) = P(U_1 \cup U_2 \cup U_3)$

$= 1 - P(F_1 \cap F_2 \cap F_3)$

$= 1 - P(F_1)\,P(F_2)\,P(F_3)$

$= 1 - (1-p_1)(1-p_2)(1-p_3)$

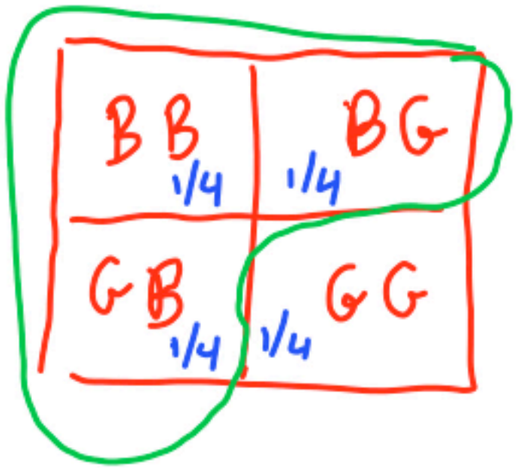
The king's sibling

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

1/2 ?

boys have precedence

P(boy) = P(girl) = 1/2
independent



2/3

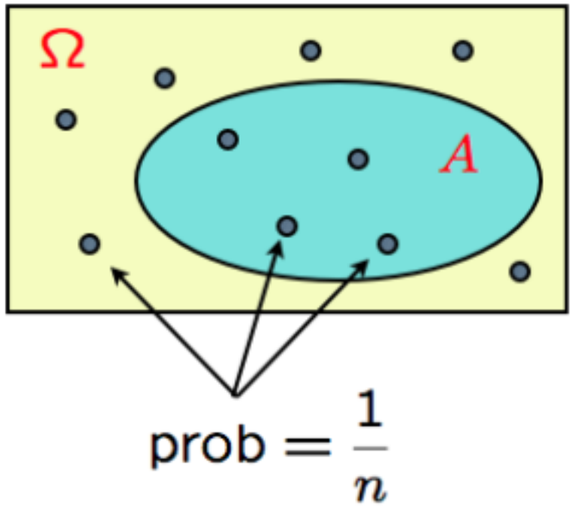
- till 1 boy ⇒ P(G) = 1
- till 2 boys ⇒ P(G) = 0

LECTURE 4: Counting

Discrete uniform law

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

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



- Basic counting principle

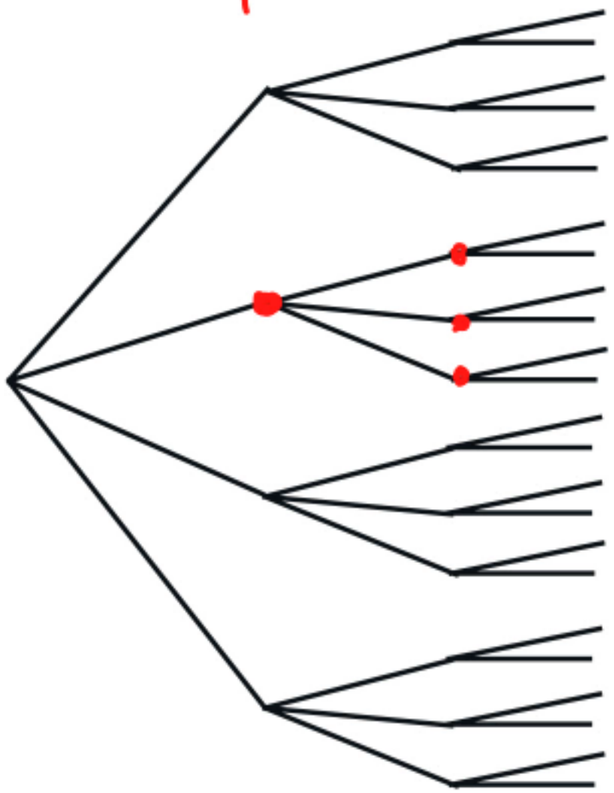
- Applications

permutations	number of subsets
combinations	binomial probabilities
partitions	

Basic counting principle

4 shirts
3 ties
2 jackets

Number of possible attires?



4 12 24 $= 4 \cdot 3 \cdot 2$

$r = 3$

$n_1 = 4$

$n_2 = 3$

$n_3 = 2$

- 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:

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

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

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

$n = 1 \quad \{1\} \quad 2^1 = 2$

$\{1\} \quad \phi$

Example

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

  (Assume all outcomes equally likely.)

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

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

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

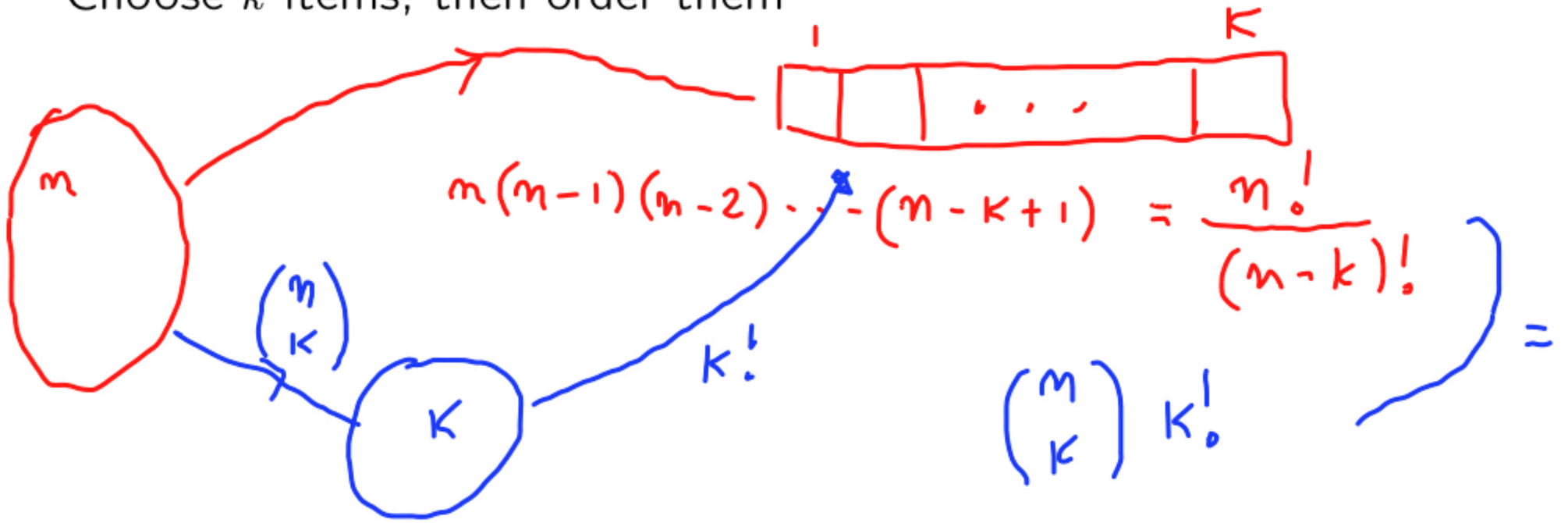
Combinations

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

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

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

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



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

$$\binom{n}{n} = 1 \qquad \frac{n!}{n!\,0!}$$

$0! = 1$ convention

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

$\emptyset$

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

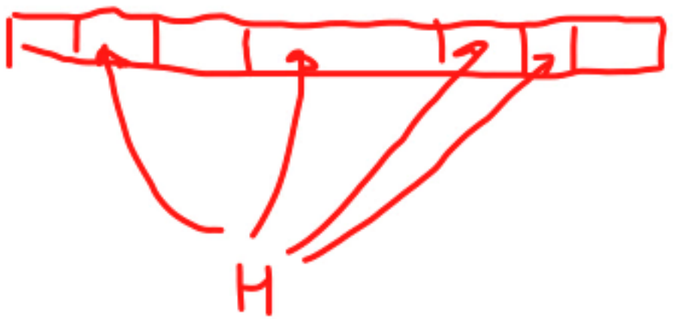
Binomial coefficient $\binom{n}{k}$ $\longrightarrow$ Binomial probabilities

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

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

- $\mathbf{P}(HTTHHH) =$ $p(1-p)(1-p)ppp = p^4(1-p)^2$ $\quad n=6$
- $\mathbf{P}(\text{particular sequence}) =$ $p^{\#\text{heads}}(1-p)^{\#\text{tails}}$
- $\mathbf{P}(\text{particular } k\text{–head sequence}) = p^k(1-p)^{n-k}$

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



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

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
- $\mathbf{P}(H) = p$

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

- First solution:

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{P(H_1 H_2 \text{ and one H in tosses } 3, \ldots, 10)}{P(B)}$$

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

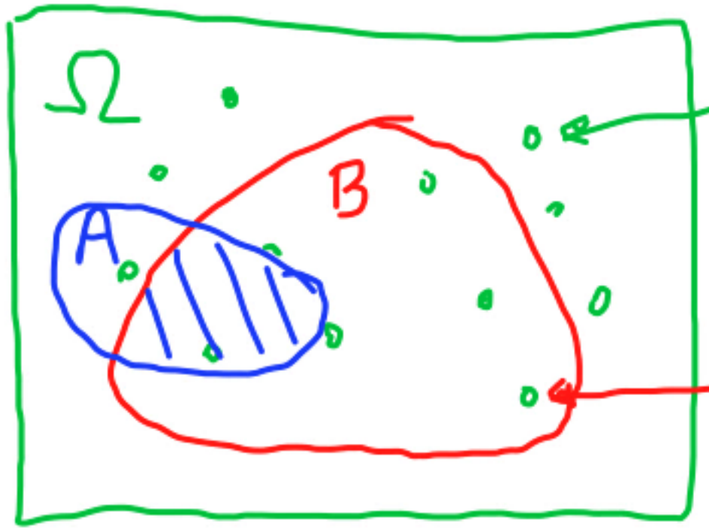
A coin tossing problem

- 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
- $\mathbf{P}(H) = p$

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

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



length 10 seq.

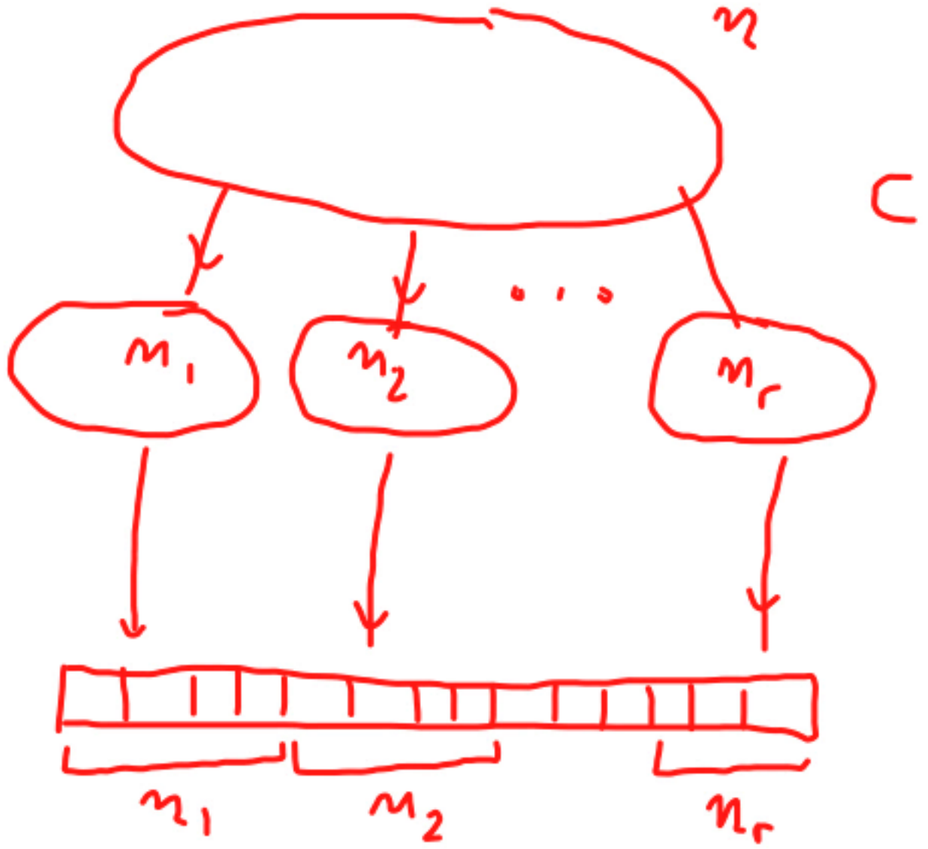
3-head seqs

$p^3(1-p)^7$

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

Partitions

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



$$C \; n_1! \, n_2! \cdots n_r! = n!$$

$r = 2 \qquad n_1 = k \qquad n_2 = n - k$

number of partitions $= \dfrac{n!}{n_1!\, n_2! \cdots n_r!}$ (multinomial coefficient)

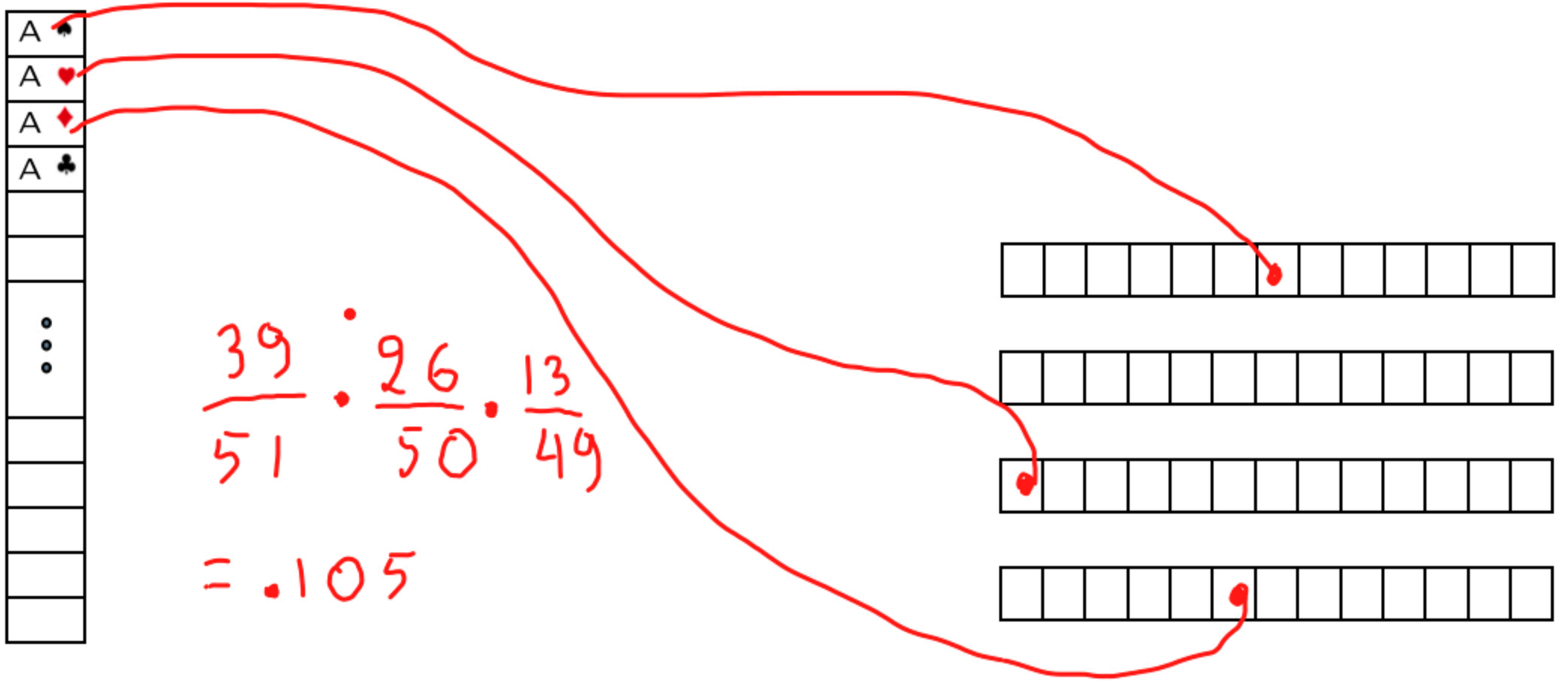
Example: 52-card deck, dealt (fairly) to four players.
Find $\mathbf{P}$(each player gets an ace)

- 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: $$\frac{4 \cdot 3 \cdot 2 \cdot \dfrac{48!}{12!\,12!\,12!\,12!}}{\dfrac{52!}{13!\,13!\,13!\,13!}}$$

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

A smart solution

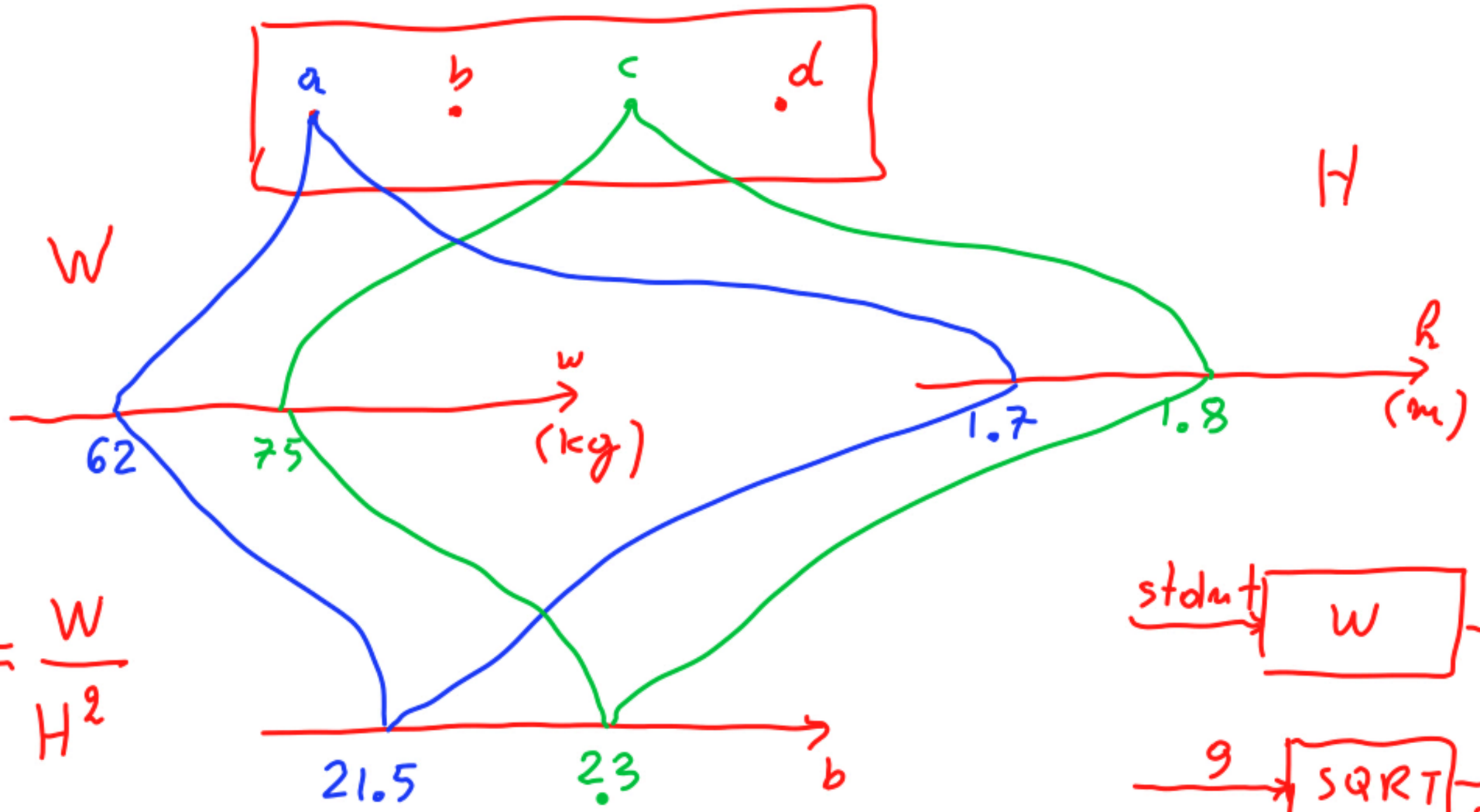
Stack the deck, aces on top



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

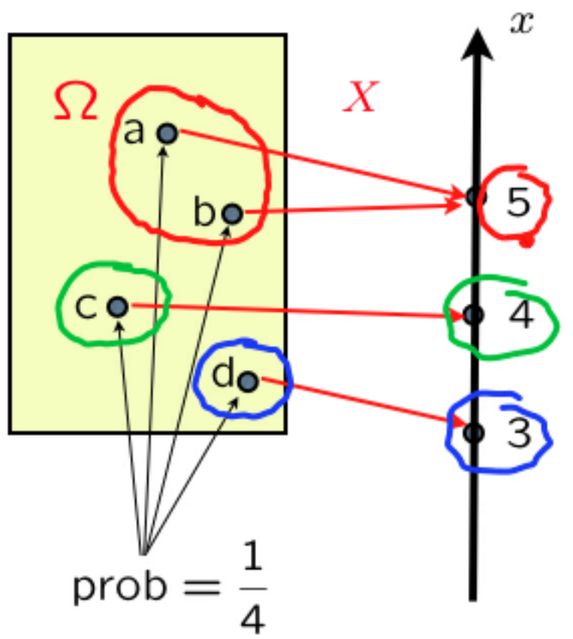
- We can have several random variables defined on the same sample space
- A function of one or several random variables is also a random variable
  - meaning of $X + Y$: r.v takes value $x+y$, when X takes value x, Y takes value y

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

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

$p_X(5) = 1/2$



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

$P_Y(y)$

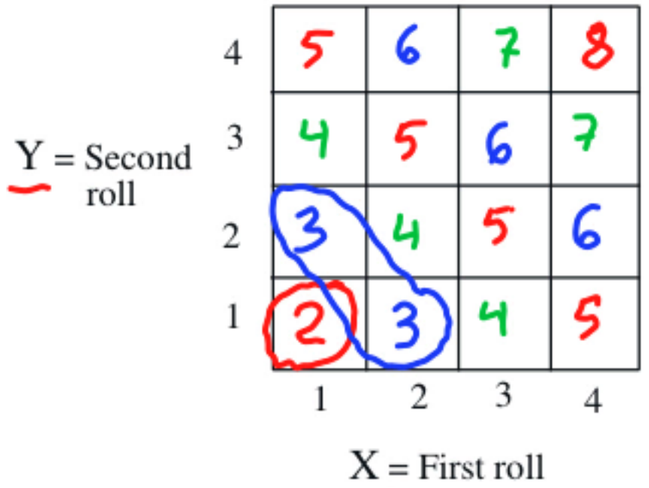
- **Properties:** $p_X(x) \geq 0 \qquad \sum_x p_X(x) = 1$

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

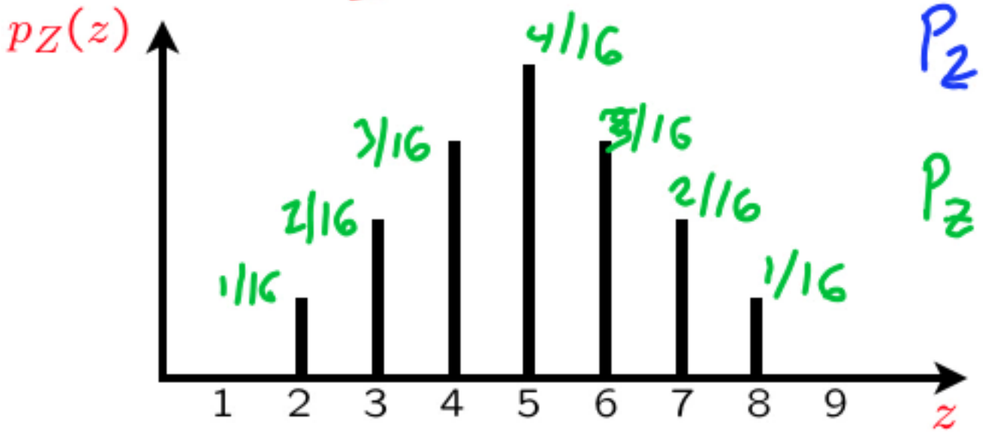


- repeat for all z:
  - collect all possible outcomes for which Z is equal to z
  - add their probabilities

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

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

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

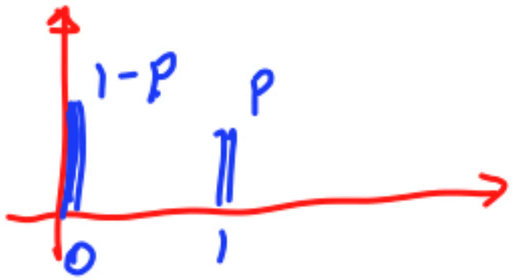


The simplest random variable: Bernoulli with parameter $p \in [0,1]$

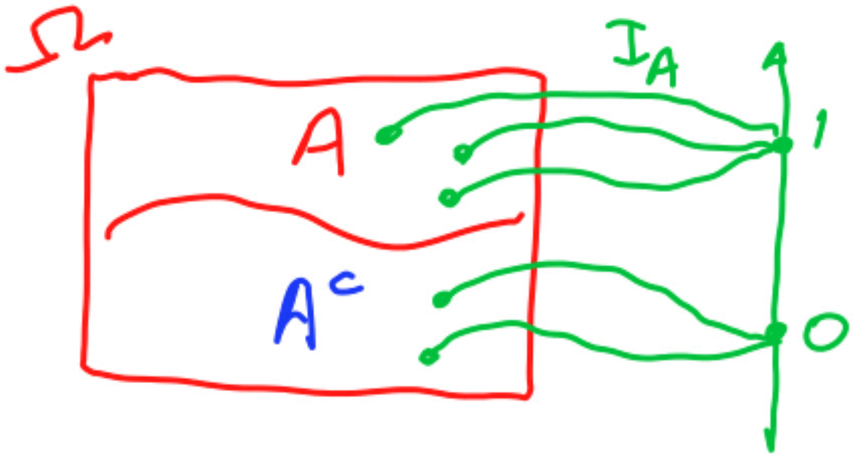
$$X = \begin{cases} 1, & \text{w.p. } p \\ 0, & \text{w.p. } 1-p \end{cases}$$

$p_X(0) = 1 - p$

$p_X(1) = p$



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



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

$P(A) = p$

Discrete uniform random variable; parameters $a,\ b$

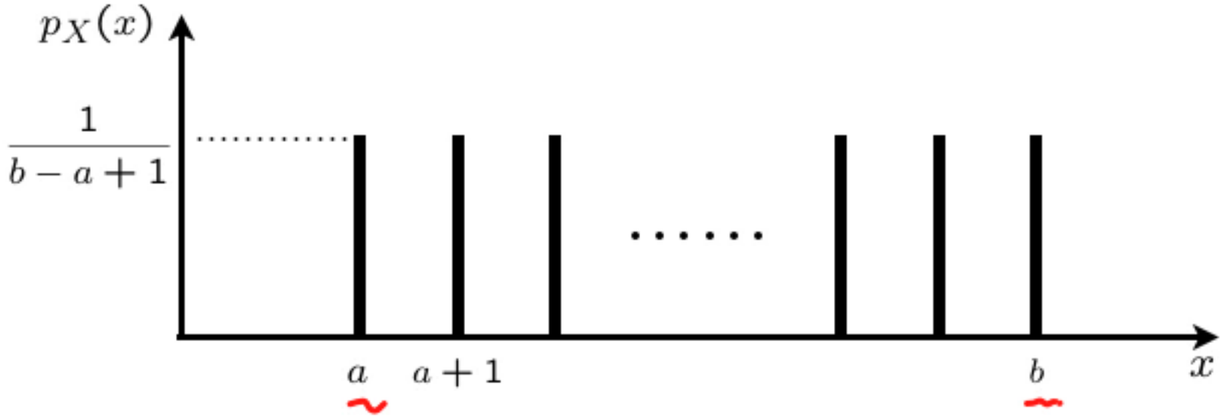
- **Parameters:** integers $a,\ b$; $a \leq b$
- **Experiment:** Pick one of $a, a+1, \ldots, b$ at random; all equally likely
- **Sample space:** $\{a, a+1, \ldots, b\}$
- **Random variable** X: $X(\omega) = \omega$
- **Model of:** complete ignorance

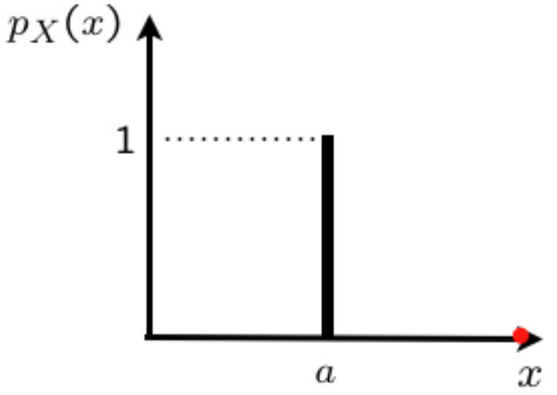
$b - a + 1$ possible values

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

Special case: $a = b$

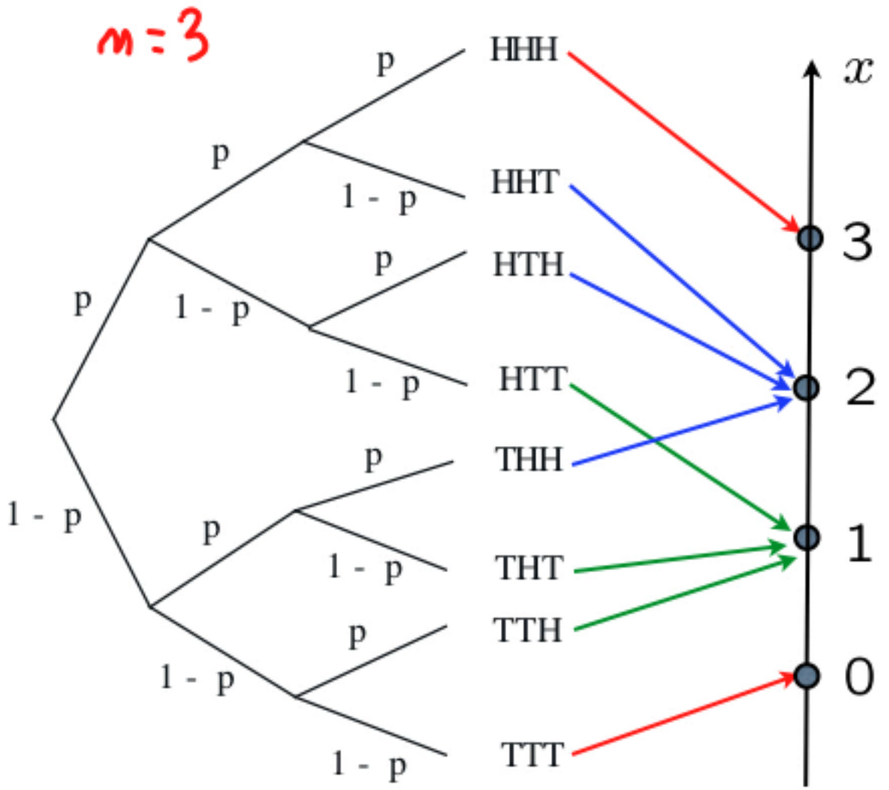
constant/deterministic r.v.





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

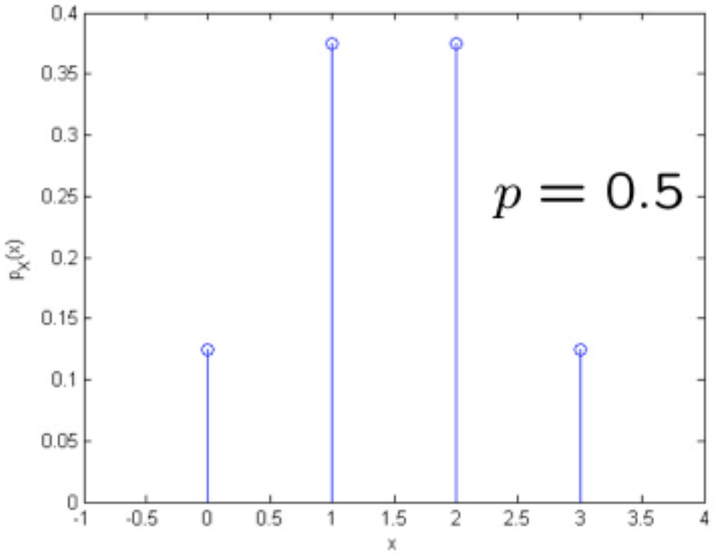
- **Experiment:** n independent tosses of a coin with $\mathbf{P}(\text{Heads}) = p$
- **Sample space:** Set of sequences of H and T, of length n
- **Random variable X:** number of Heads observed
- **Model of:** number of successes in a given number of independent trials

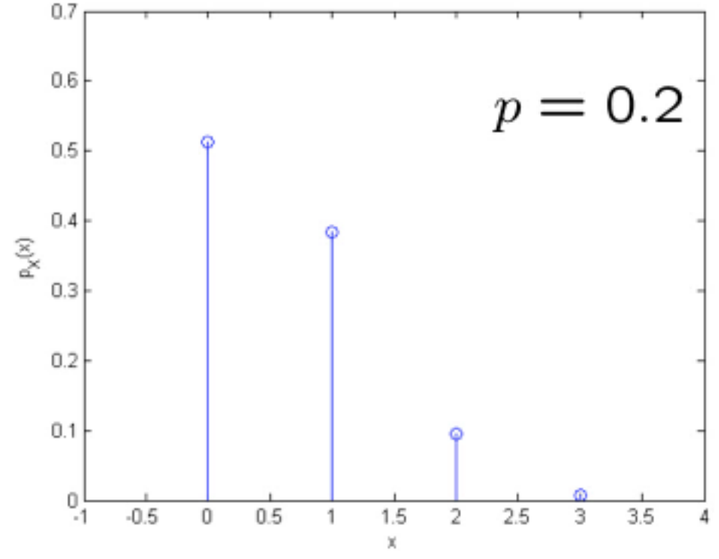


$$P_X(2) = P(X=2)$$
$$= P(HHT) + P(HTH) + P(THH)$$
$$= 3p^2(1-p) = \binom{3}{2} p^2(1-p)$$

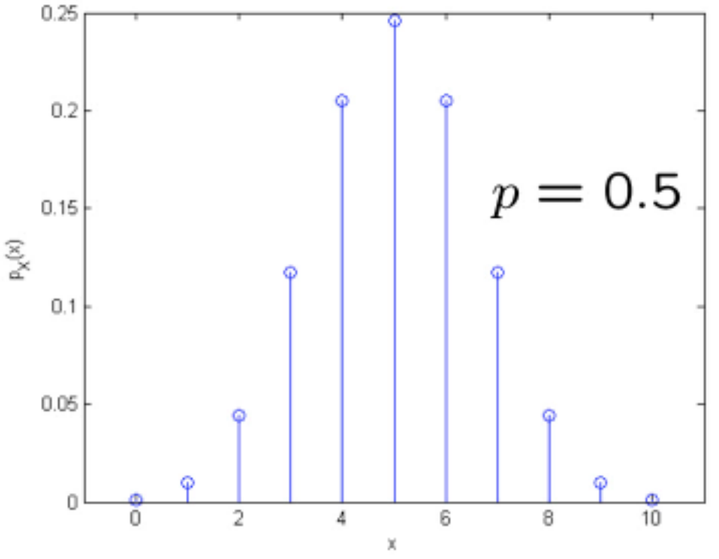
$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k = 0, 1, \ldots, n$$

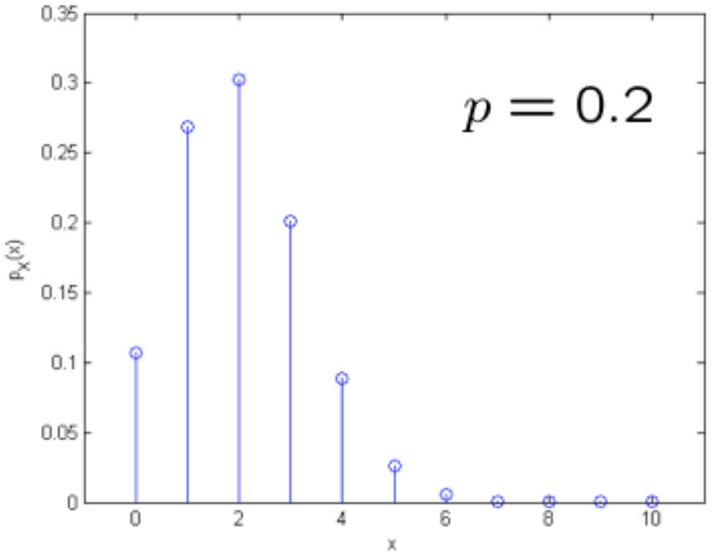
$n = 3$



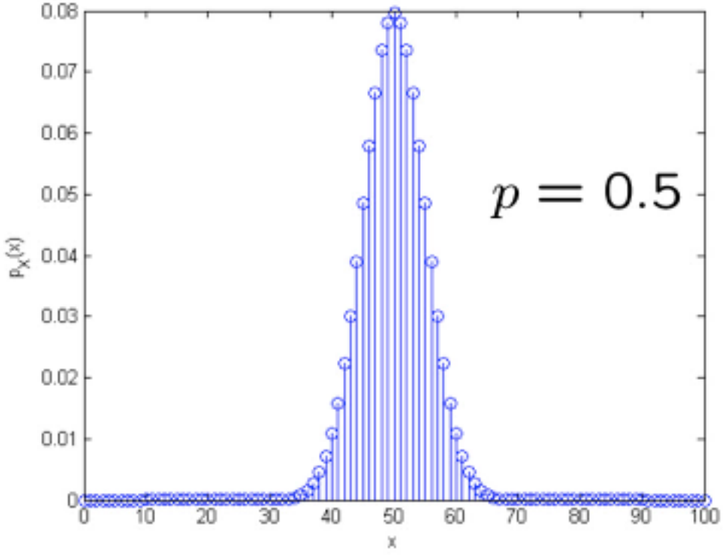


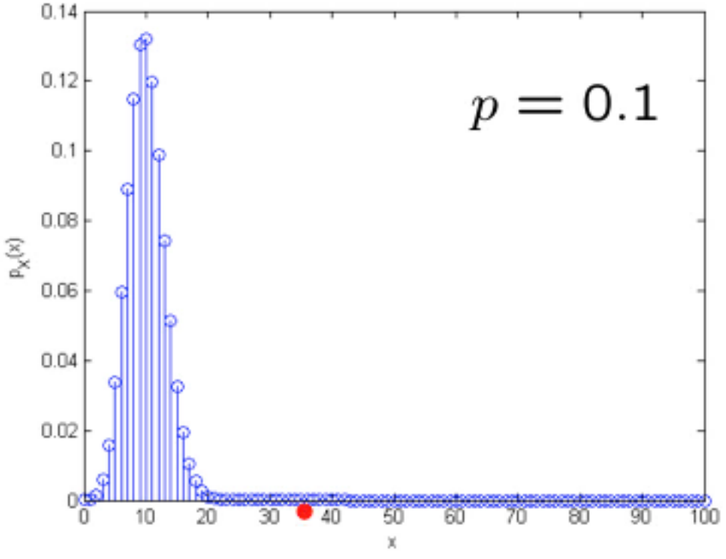
$n = 10$





$n = 100$





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

- **Experiment:** infinitely many independent tosses of a coin; $\mathbf{P}(\text{Heads}) = p$
- **Sample space:** Set of infinite sequences of H and T

TTTTHHT...

$X = 5$

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

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

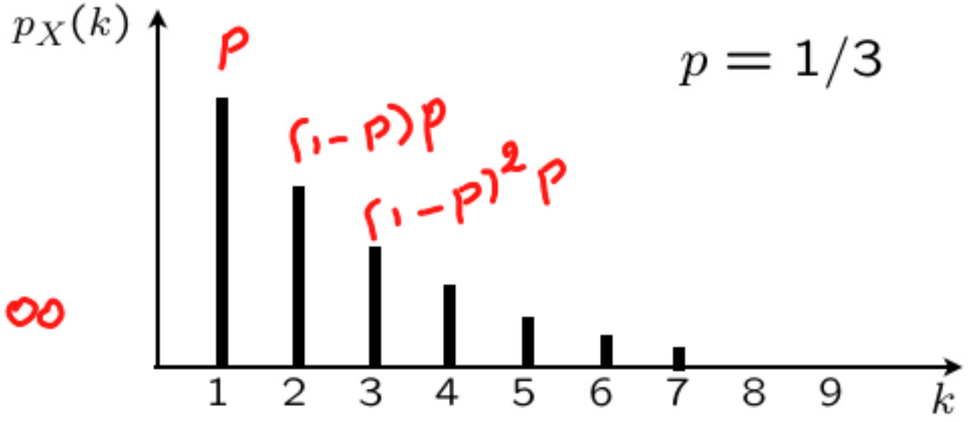
$$p_X(k) = P(X = k) = P(\underbrace{T \cdots T}_{k-1}\, H) = (1-p)^{k-1} p \qquad k = 1, 2, 3, \ldots$$

$\mathbf{P}(\text{no Heads ever}) \leq P(\underbrace{T \cdots T}_{k}) = (1-p)^k$

(TTT...) "$X = \infty$" $= 0$

$k \to \infty$

0

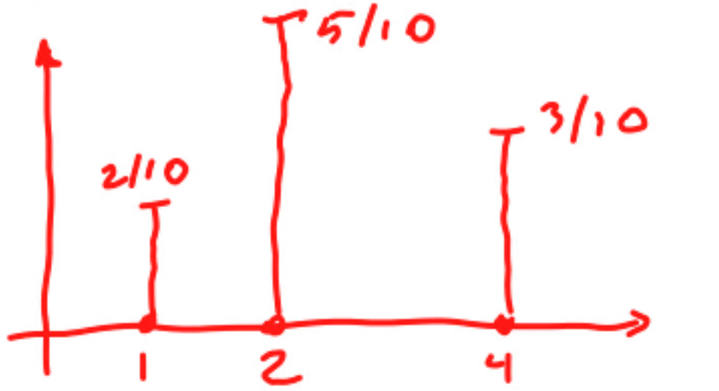


Expectation/mean of a random variable

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

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

~200
~500
~300



- "Average" gain:

$$\frac{1 \cdot 200 + 2 \cdot 500 + 4 \cdot 300}{1000}$$

$$= 1 \cdot \frac{2}{10} + 2 \cdot \frac{5}{10} + 4 \cdot \frac{3}{10}$$

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

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

- 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.

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

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

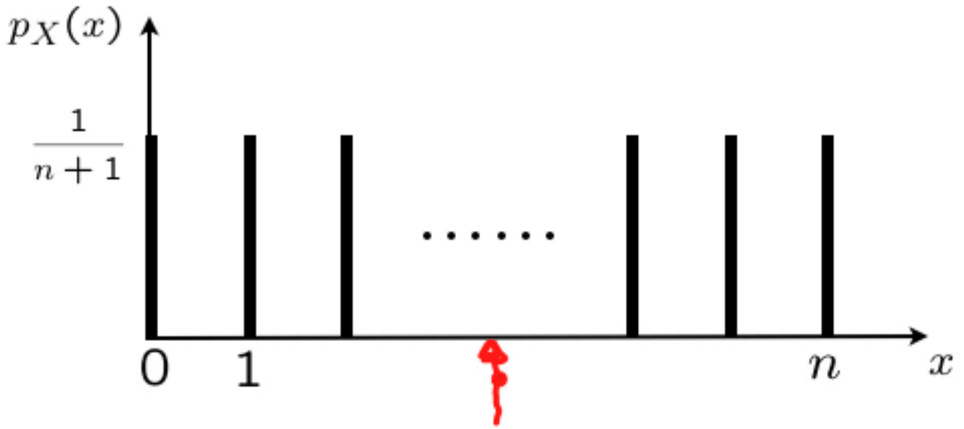
If X is the indicator of an event A, $X = I_A$:

$X = 1$ iff A occurs $\qquad p = P(A)$

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

Expectation of a uniform r.v.

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



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

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

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

Expectation as a population average

- n students
- Weight of ith 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

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

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

Elementary properties of expectations

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

$\geq 0 \quad \geq 0 \quad \geq 0$

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

for all ω: $X(\omega) \geq 0$

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

for all ω: $a \leq X(\omega) \leq b$

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

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

- If c is a constant, $\mathbf{E}[c] = c$

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

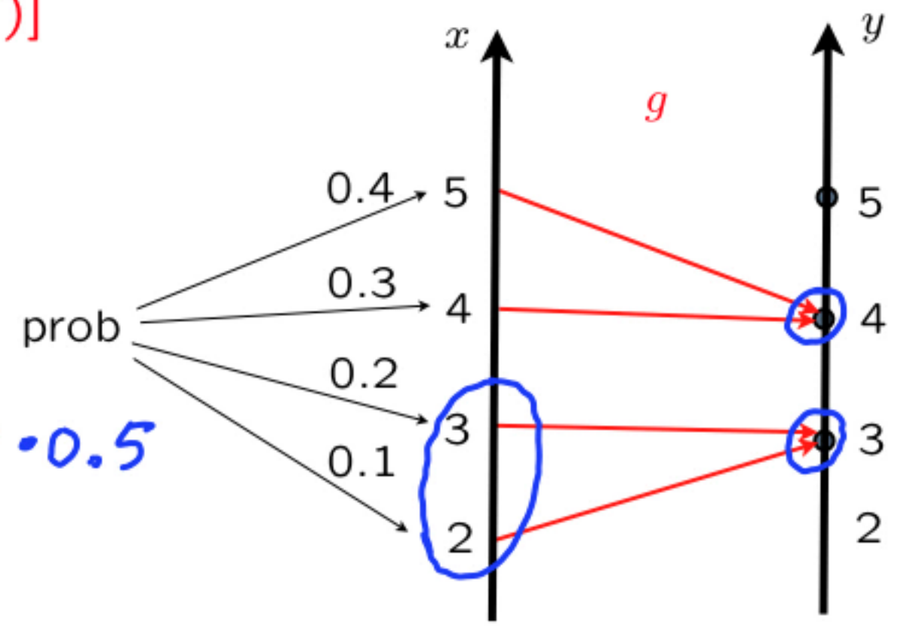
The expected value rule, for calculating $\mathbf{E}[g(X)]$

- Let X be a r.v. and let $Y = g(X)$
- Averaging over y: $\mathbf{E}[Y] = \sum_y y p_Y(y)$

$3 \cdot (0.1+0.2) + 4 \cdot (0.3+0.4)$

- Averaging over x: $3 \cdot 0.1 + 3 \cdot 0.2 + 4 \cdot 0.3 + 4 \cdot 0.5$

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



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

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

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

- $\mathbf{E}[X^2] =$ $\sum_x x^2 p_X(x)$

$g(x) = x^2$

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

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

Linearity of expectation: $\mathbf{E}[aX + b] = a\mathbf{E}[X] + b$

X = salary $E[X]$ = average salary

Y = new salary $= 2X + 100$ $E[Y] = E[2X + 100] = 2E[X] + 100$

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

$g(x) = ax + b$

$Y = g(X)$

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

$$= \sum_x (ax + b) p_X(x) = a \sum_x x\, p_X(x) + b \underbrace{\sum_x p_X(x)}_{1}$$

$$= aE[X] + b$$

$E[g(X)] = g(E[X])$ exceptional g

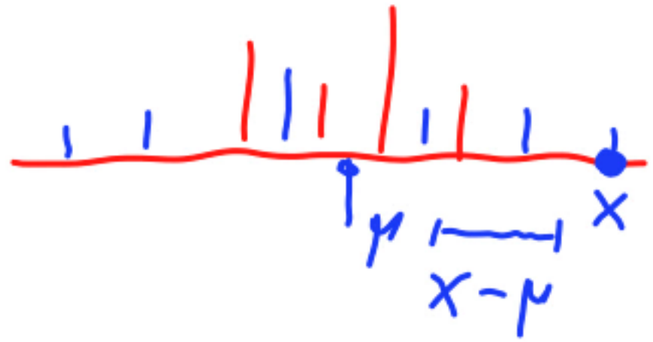
LECTURE 6: Variance; Conditioning on an event; Multiple random variables

- 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 $\mu = \mathbf{E}[X]$
- Distance from the mean: $X - \mu$
- Average distance from the mean?

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



- **Definition of variance:** $\text{var}(X) = \mathbf{E}[(X-\mu)^2]$ ≥ 0

- Calculation, using the expected value rule, $\mathbf{E}[g(X)] = \sum_x g(x) p_X(x)$

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

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

Properties of the variance

- Notation: $\mu = \mathbf{E}[X]$

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

$\text{var}(3-4X)$
$= (-4)^2 \text{var}(X)$
$= 16\ \text{var}(X)$

- Let $Y = X + b$

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

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

- Let $Y = aX$

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

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

A useful formula: $\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2$

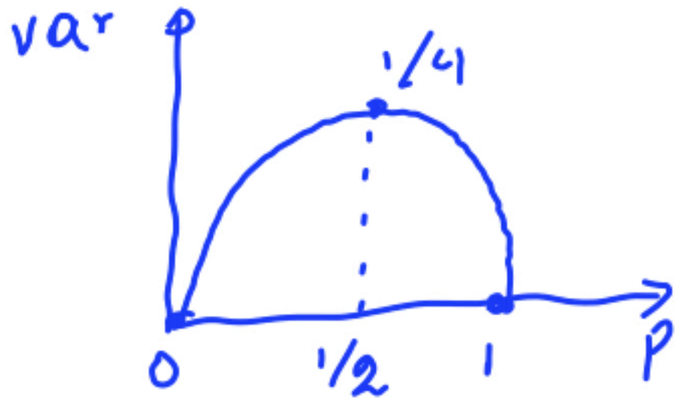
$\text{var}(X) = E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2]$

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

Variance of the Bernoulli

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

$E[X] = p$



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

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

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

$X^2 = X$

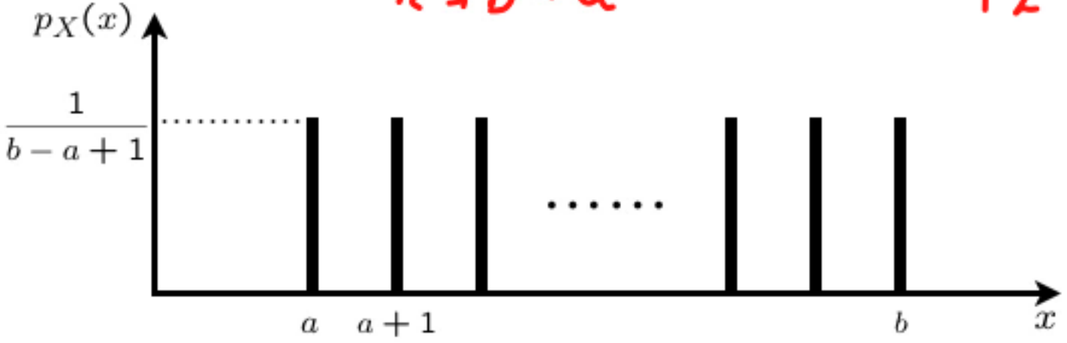
Variance of the uniform

$$\frac{1}{6}n(n+1)(2n+1)$$

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

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

$n = b-a$



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

Conditional PMF and expectation, given an event

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

assume $\mathbb{P}(A) > 0$

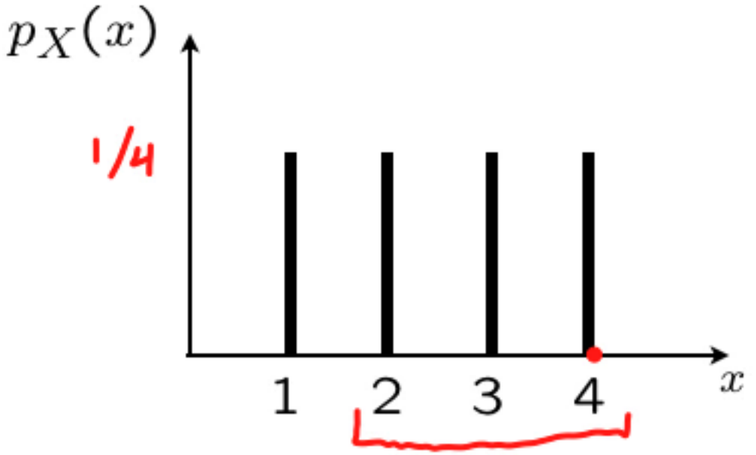
$$p_X(x) = \mathbf{P}(X = x) \qquad p_{X|A}(x) = \mathbf{P}(X = x \mid A)$$

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

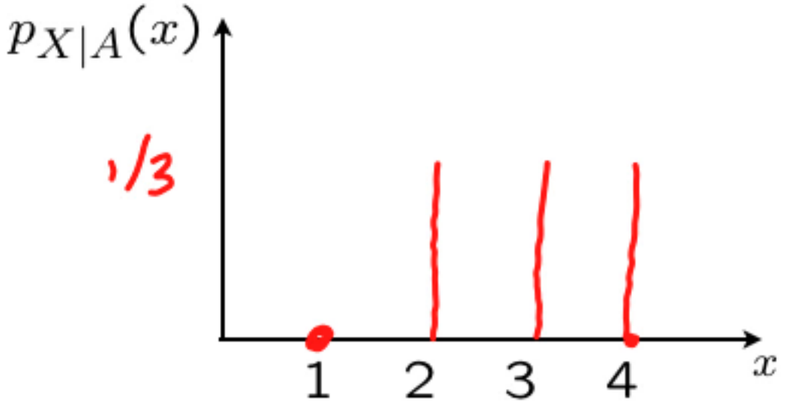
$$\mathbf{E}[X] = \sum_x x p_X(x) \qquad \mathbf{E}[X \mid A] = \sum_x x p_{X|A}(x)$$

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

Example of conditioning



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



$\mathbf{E}[X] = 2.5$

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

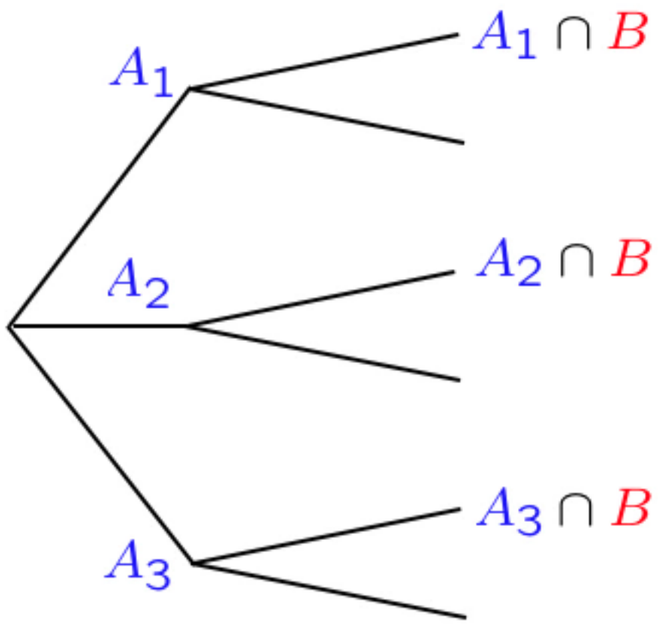
$= \frac{1}{12}\, 3 \cdot 5 = \frac{5}{4}$

$\mathbf{E}[X \mid A] = 3$

$\mathrm{var}(X \mid A) = \frac{1}{3}(4-3)^2 + \frac{1}{3}(3-3)^2$

$+ \frac{1}{3}(2-3)^2 = \frac{2}{3}$

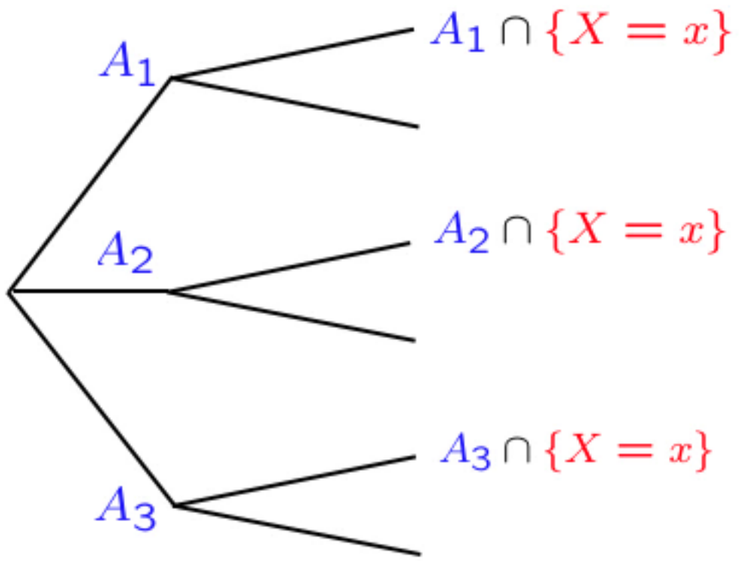
Total expectation theorem

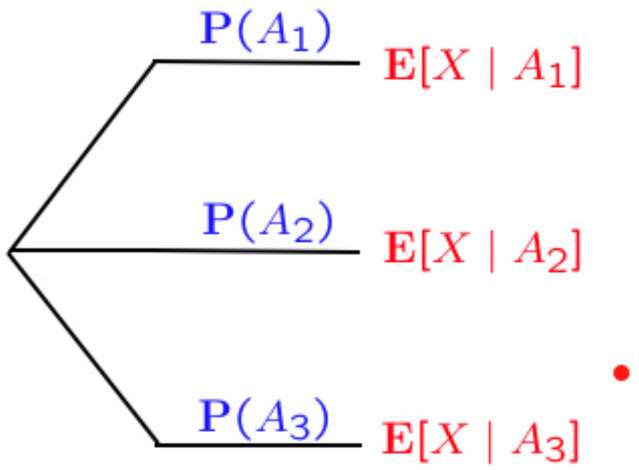


$$\mathbf{P}(B) = \mathbf{P}(A_1)\,\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\,\mathbf{P}(B \mid A_n)$$

$B = \{x = x\}$

Total expectation theorem





$$\mathbf{P}(B) = \mathbf{P}(A_1)\,\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\,\mathbf{P}(B \mid A_n)$$

$B = \{X = x\}$

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

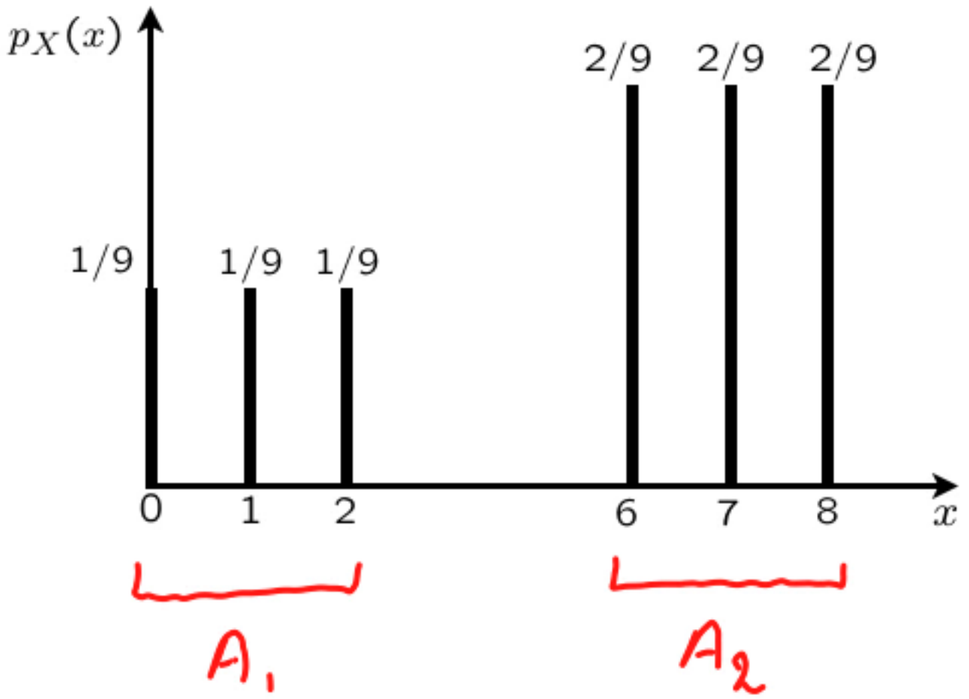
for all x

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

$E[X]$

$$\mathbf{E}[X] = \mathbf{P}(A_1)\,\mathbf{E}[X \mid A_1] + \cdots + \mathbf{P}(A_n)\,\mathbf{E}[X \mid A_n]$$

Total expectation example



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

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

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

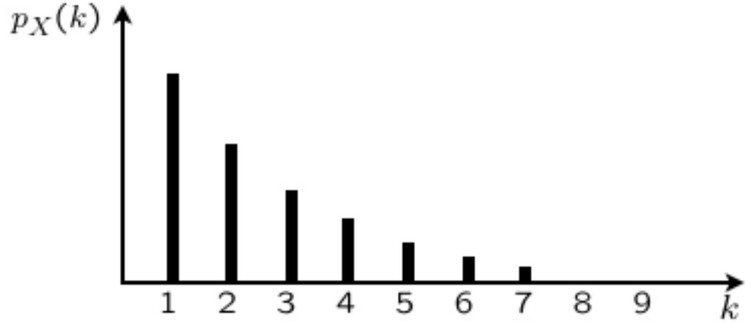
$$E[X|A_2] = 7$$

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

Conditioning a geometric random variable

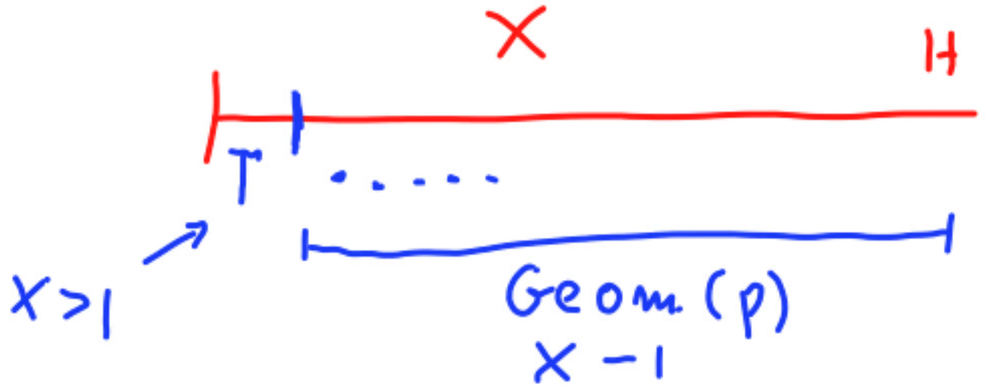
- X: number of independent coin tosses until first head; $\mathbf{P}(H) = p$

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



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

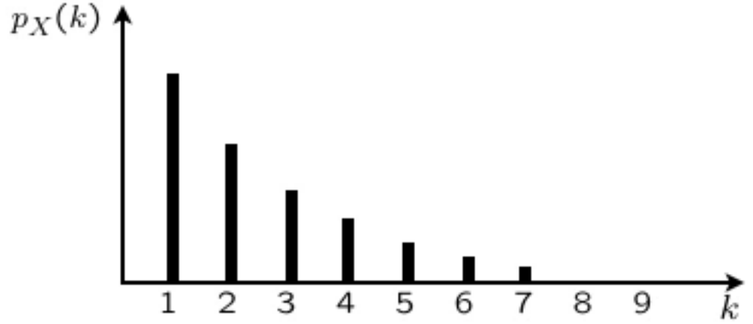
Conditioned on $X > 1$, $X - 1$ is geometric with parameter p



Conditioning a geometric random variable

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

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



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

$$p_{X-1|X>1}(3) = P(X-1=3 \mid X>1) = P(T_2\, T_3\, H_4 \mid T_1) = P(T_2\, T_3\, H_4)$$

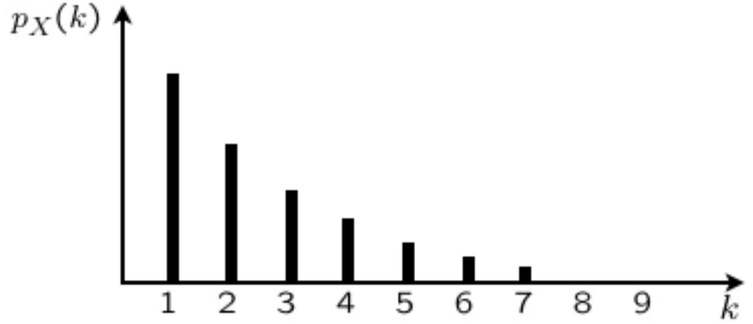
$$= (1-p)^2 p = p_X(3)$$

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

Conditioning a geometric random variable

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

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



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

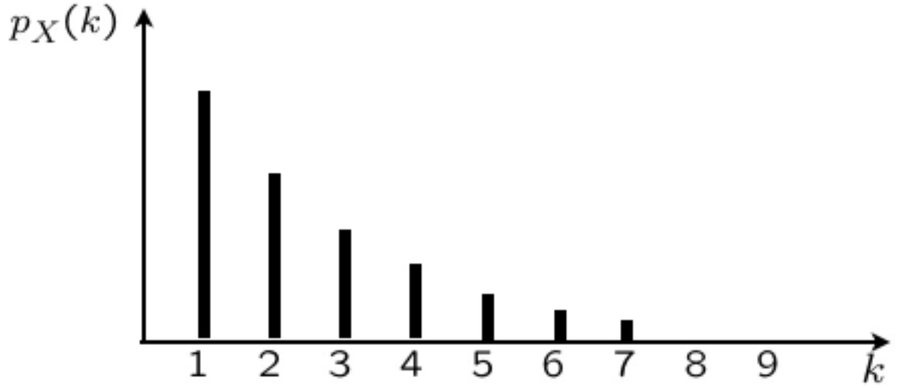
Conditioned on $X > n$, $X - n$ is geometric with parameter p

$$p_{X-1\,|\,X>1}(3) = P(X-1=3 \mid X>1) = P(T_2\,T_3\,H_4 \mid T_1) = P(T_2\,T_3\,H_4)$$

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

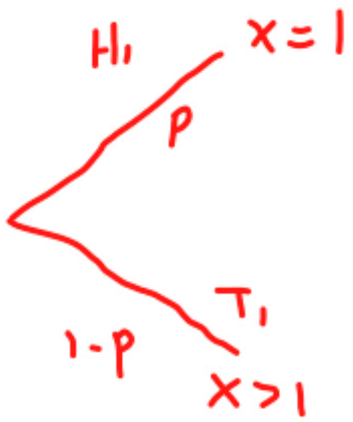
$$p_{X-1\,|\,X>1}(k) = p_X(k) = p_{X-n\,|\,X>n}(k).$$

The mean of the geometric



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

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



$$E[X] = 1 + E[X-1]$$

$$= 1 + p \cdot E[X-1 \mid X=1] + (1-p)\, E[X-1 \mid X>1]$$

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

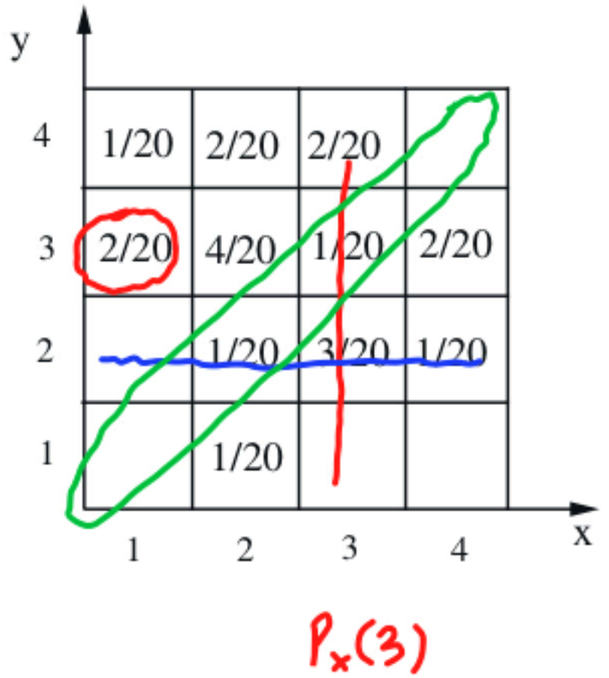
Multiple random variables and joint PMFs

marginal pmfs

$X : p_X$
$Y : p_Y$

$\mathbf{P}(X = Y) = \frac{2}{20}$

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



$P_X(3)$

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

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

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

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

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

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

More than two random variables

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

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

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

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

Functions of multiple random variables

$Z = g(X, Y)$

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

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

$E[g(X)]$

Linearity of expectations

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

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

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

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

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

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

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

$$= \sum_x x\, p_X(x) + \sum_y y\, p_Y(y) = E[X] + E[Y]$$

Linearity of expectations

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

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

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

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

The mean of the binomial

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

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

$p_X(k)$

$$\mathbf{E}[X] = np$$

$X_i = 1$ if ith trial is a success; p

$X_i = 0$ otherwise $1-p$

(indicator variable)

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

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

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

$A = \{Y = y\}$

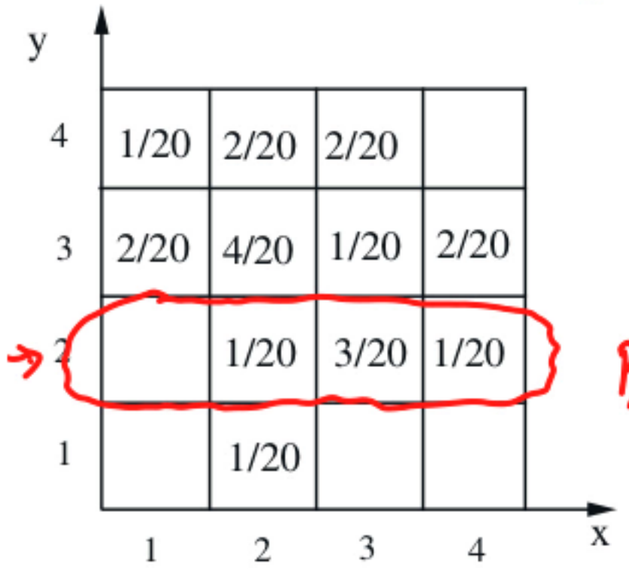
$p_{X|A}(x \mid A) = \mathbf{P}(X = x \mid A)$

$p_{X|Y}(x \mid y) = \mathbf{P}(X = x \mid Y = y) = \dfrac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)}$

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \qquad \text{defined for } y \text{ such that } p_Y(y) > 0$$

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

$Y = 2$ $\quad P_Y(2) = 5/20$



$P_{X|Y}(1|2) = 0$

$P_{X|Y}(2|2) = 1/5$

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

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

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

- Self-explanatory notation

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

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

- Multiplication rule

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

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

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

Conditional expectation

$A = \{Y = y\}$

$$\mathbf{E}[X] = \sum_x x p_X(x) \qquad \mathbf{E}[X \mid A] = \sum_x x\, p_{X|A}(x) \qquad \mathbf{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y)$$

- Expected value rule

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

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

Total probability and expectation theorems

- $A_1, \ldots, A_n$: partition of Ω

$Y = \{y_1, \ldots, y_n\}$ $A_i = \{Y = y_i\}$

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

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

- $\mathbf{E}[X] = \mathbf{P}(A_1)\, \mathbf{E}[X \mid A_1] + \cdots + \mathbf{P}(A_n)\, \mathbf{E}[X \mid A_n]$

$$\mathbf{E}[X] = \sum_y p_Y(y)\, \mathbf{E}[X \mid Y = y]$$

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

Independence

- of two events: $\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B)$ $\qquad$ $\mathbf{P}(A \mid B) = \mathbf{P}(A)$

- of a r.v. and an event: $\mathbf{P}(X = x \text{ and } A) = \mathbf{P}(X = x) \cdot \mathbf{P}(A)$, for all x

$p_{X|A}(x) = p_X(x)$, for all x $\qquad$ $P(A \mid X = x) = P(A)$, for all x

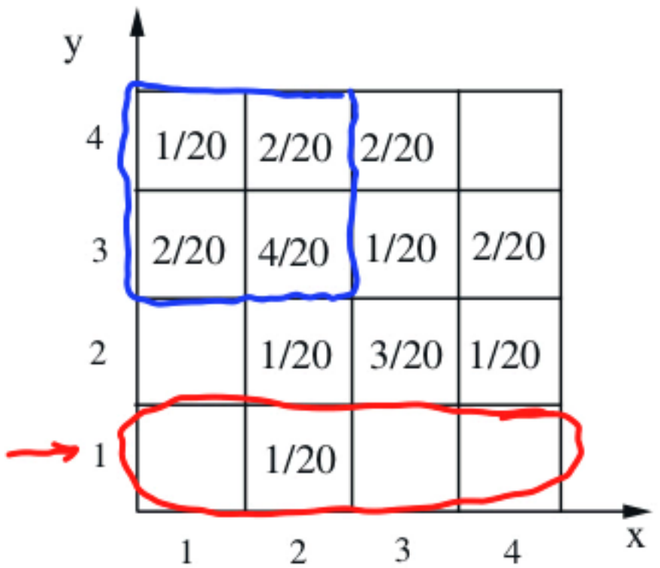
- of two r.v.'s: $\mathbf{P}(X = x \text{ and } Y = y) = \mathbf{P}(X = x) \cdot \mathbf{P}(Y = y)$, for all x, y

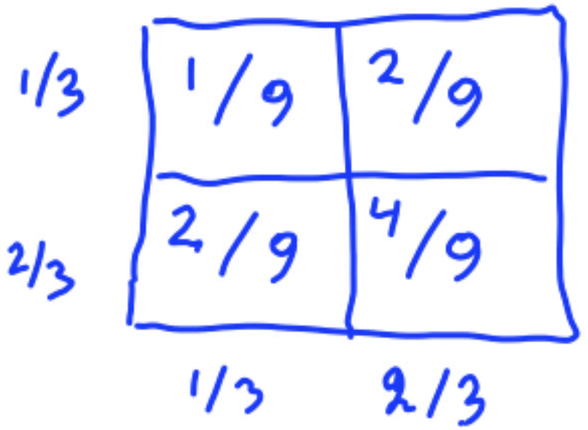
$p_{X|Y}(x|y) = p_X(x)$ $\qquad$ $p_{X,Y}(x, y) = p_X(x)\, p_Y(y)$, for all x, y

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

X, Y, Z are **independent** if:
$p_{X,Y,Z}(x, y, z) = p_X(x)\, p_Y(y)\, p_Z(z)$, 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 $Y \geq 3$?

Yes .

Independence and expectations

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

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

always true

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

If X, Y are **independent:** $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$

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

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

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

$$= \sum_x x\, p_X(x) \underbrace{\sum_y y\, p_Y(y)}_{y} = E[X]\, E[Y.]$$

Independence and variances

- Always true: $\text{var}(aX) = a^2\text{var}(X)$ $\quad$ $\text{var}(X+a) = \text{var}(X)$

- 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
$E[X] = E[Y] = 0$
$E[XY] = E[X]E[Y] = 0$

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

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

- 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

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

(indicator variable)

independent

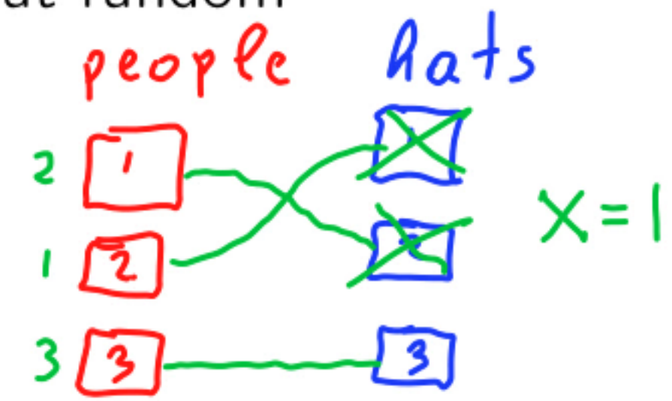
$$X = X_1 + \cdots + X_n$$

$$\text{var}(X) = \text{var}(X_1) + \cdots + \text{var}(X_n)$$

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

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 $\mathbf{E}[X]$: $E[X_1] + \cdots + E[X_n] = n \cdot \frac{1}{n} = 1$

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

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

$\frac{1}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{3!}$

$p_X(k)$: hard!

$\sum_k k\, p_X(k)$

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

The variance in the hat problem

- X: number of people who get their own hat
  - Find var(X)

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

$n = 2$

$X_1 = 1 \Rightarrow X_2 = 1$

$X_1 = 0 \Rightarrow X_2 = 0$

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

$$X^2 = \sum_i X_i^2 + \sum_{i,j:i\neq j} X_i X_j$$

(n terms; $n(n-1)$ terms)

- $\text{var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 = 2 - 1 = 1$
- $\mathbf{E}[X_i^2] = E[X_1^2] = E[X_1] = 1/n$

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

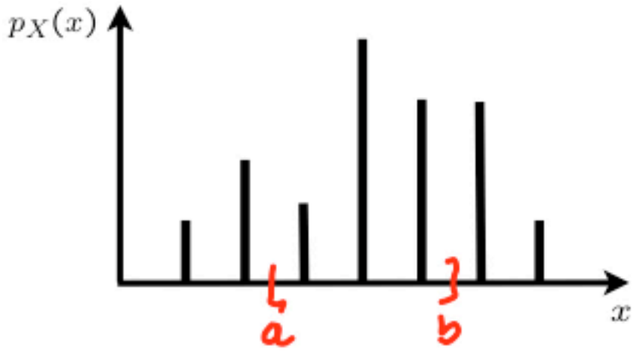
- For $i \neq j$: $\mathbf{E}[X_i X_j] = E[X_1 X_2] = P(X_1 X_2 = 1) = P(X_1 = 1, X_2 = 1)$

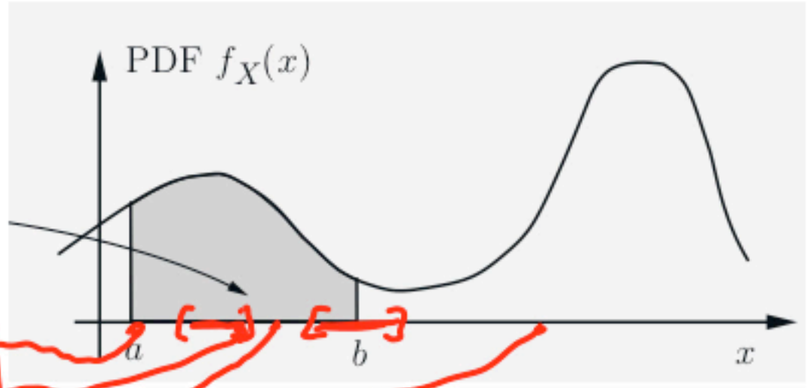
$= P(X_1 = 1)\, P(X_2 = 1 \mid X_1 = 1) = \frac{1}{n} \cdot \frac{1}{n-1}$

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)





$$\mathbf{P}(a \le X \le b) = \sum_{x:\, a \le x \le b} p_X(x)$$

$$p_X(x) \ge 0 \qquad \sum_x p_X(x) = 1$$

$$\mathbf{P}(a \le X \le b) = \int_a^b f_X(x)\, dx$$

- $f_X(x) \ge 0$
- $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$

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

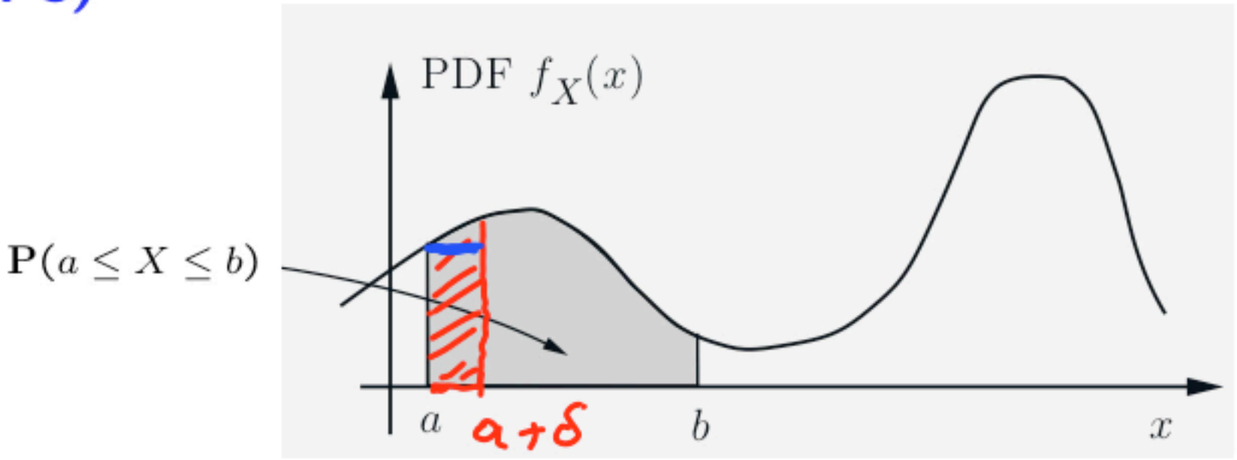
$$P(1 \le X \le 3 \text{ or } 4 \le X \le 5) = P(1 \le X \le 3) + P(4 \le X \le 5)$$

Probability density functions (PDFs)

$\delta > 0$, small

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

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



$$\mathbf{P}(a \le X \le a+\delta) \approx f_X(a) \cdot \delta$$

$$\mathbf{P}(X = a) = 0$$

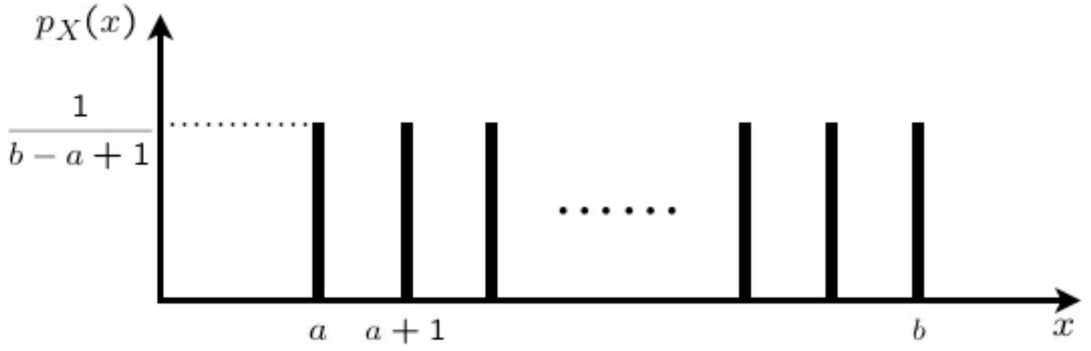
$$\mathbf{P}(a \le X \le b) = \int_a^b f_X(x)\,dx$$

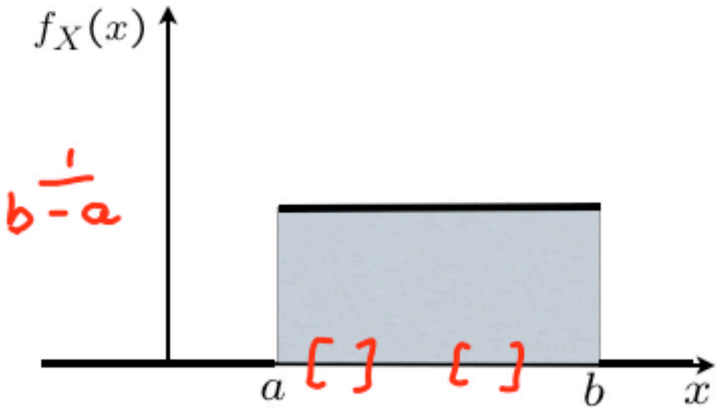
$$f_X(x) \ge 0$$

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

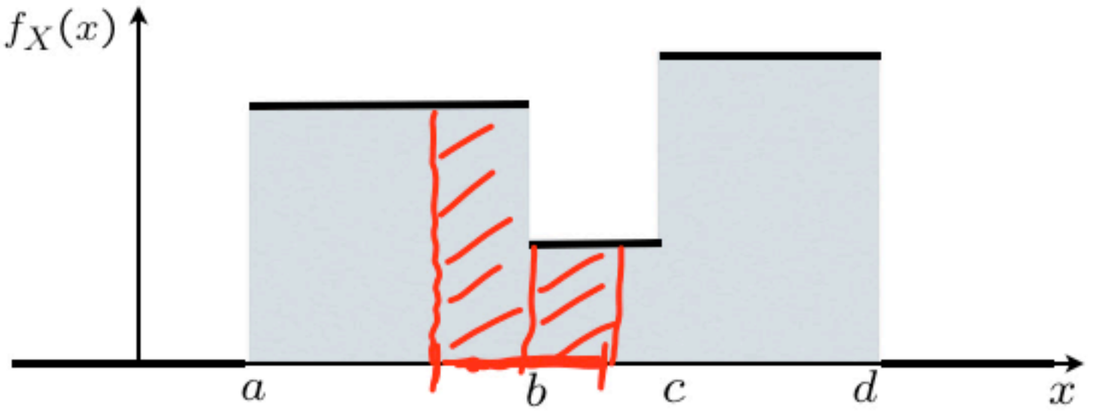
$P(a \le X \le b) = P(X = a) + P(X = b) + P(a < X < b)$

Example: continuous uniform PDF

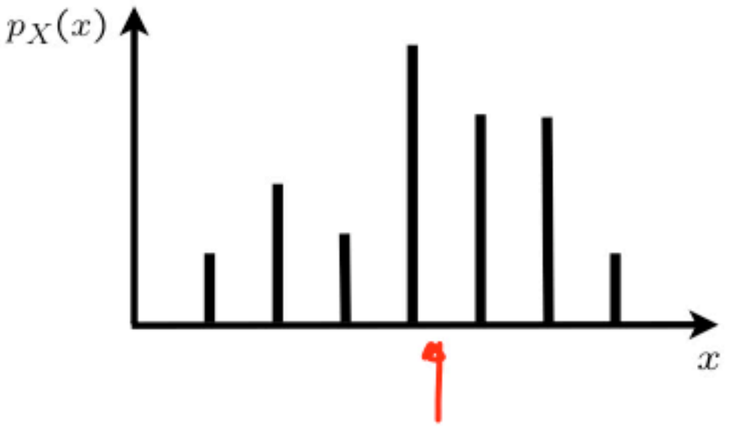




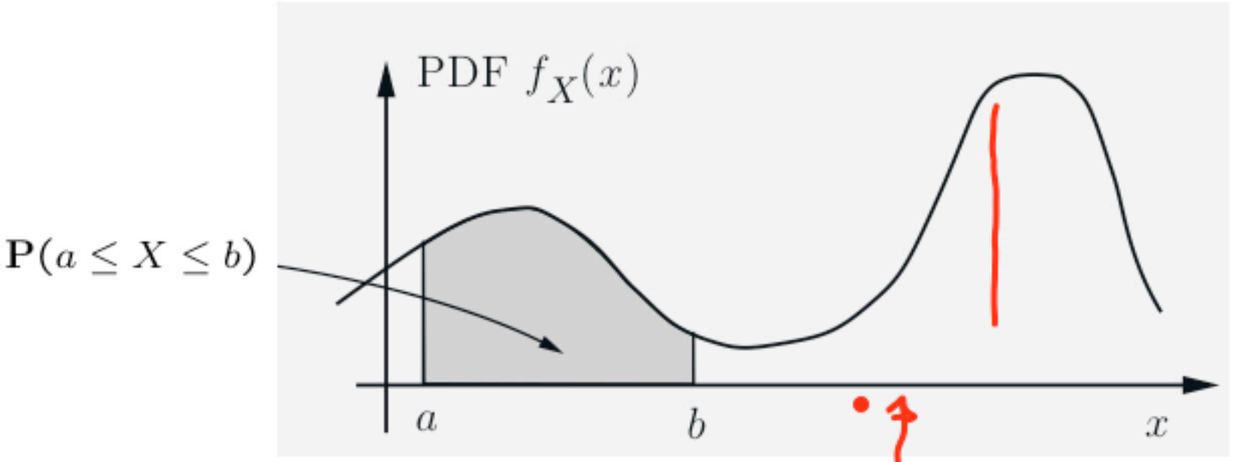
- Generalization: piecewise constant PDF



Expectation/mean of a continuous random variable



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



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

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

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

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

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

- Linearity

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

Variance and its properties

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

$\mu = E[x]$

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

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

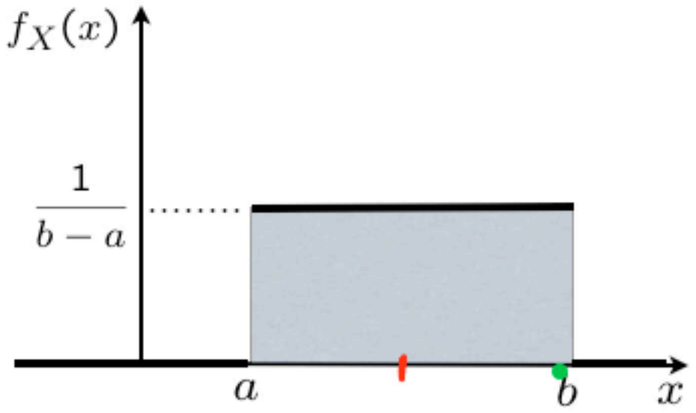
$g(x) = (x-\mu)^2$

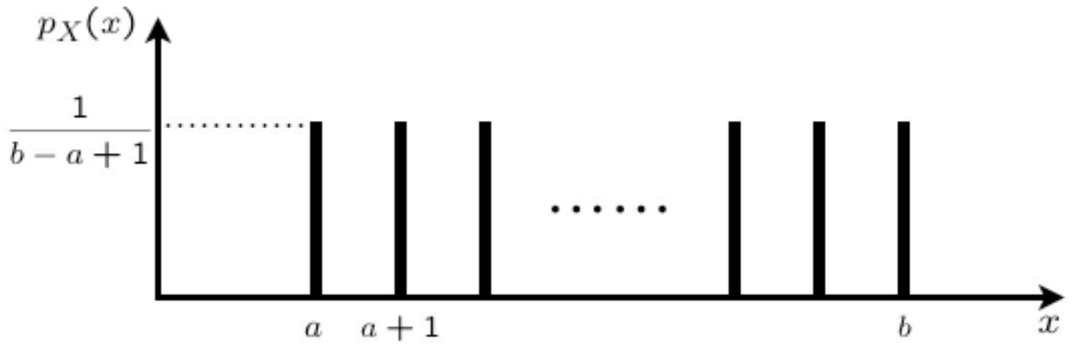
- **Standard deviation:** $\sigma_X = \sqrt{\text{var}(X)}$

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

- A useful formula: $\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2$

Continuous uniform random variable; parameters a, b





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

$$= \int_a^b x \cdot \frac{1}{b-a}\,dx = \frac{a+b}{2}$$

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

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

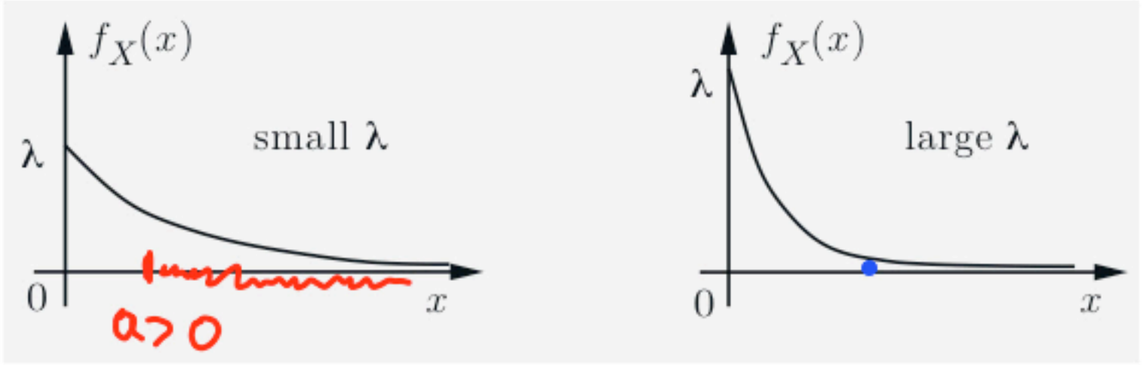
$$\mathrm{var}(X) = \frac{1}{12}(b-a)(b-a+2)$$

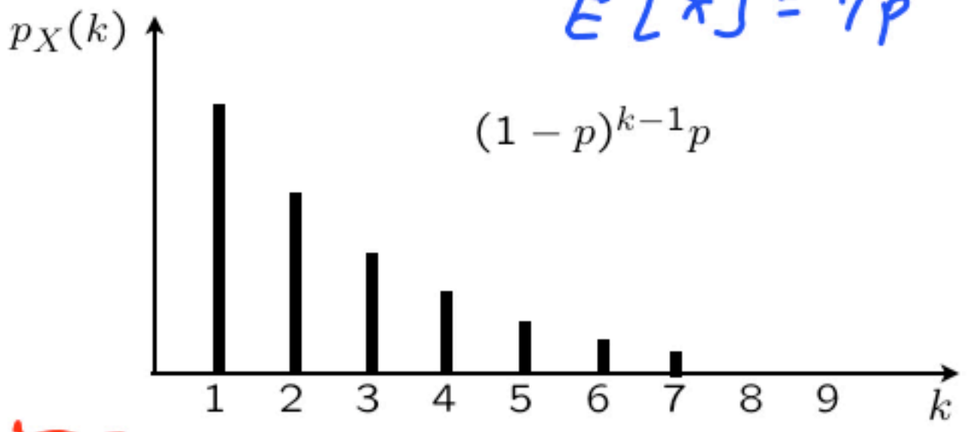
$$\mathrm{var}(X) = E[x^2] - (E[x])^2 = \boxed{(b-a)^2/12} \qquad \sigma = \frac{b-a}{\sqrt{12}}$$

Exponential random variable; parameter $\lambda > 0$

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

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





$E[X] = 1/p$

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

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

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

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

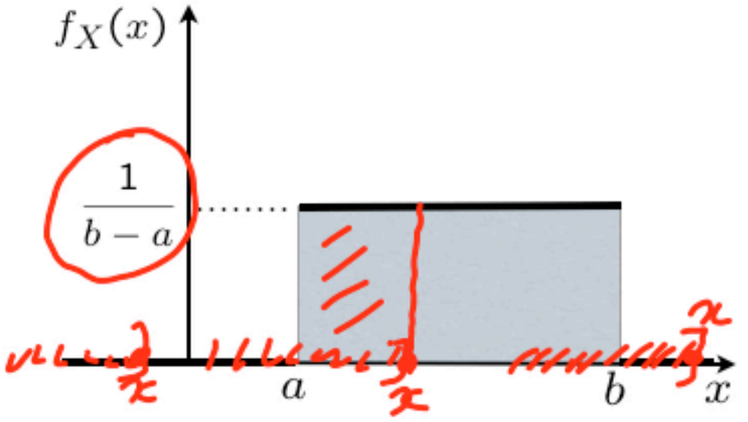
$\left[\int e^{ax}\,dx = \frac{1}{a} e^{ax} \quad a \leftrightarrow -\lambda \right]$

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

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

Cumulative distribution function (CDF)

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

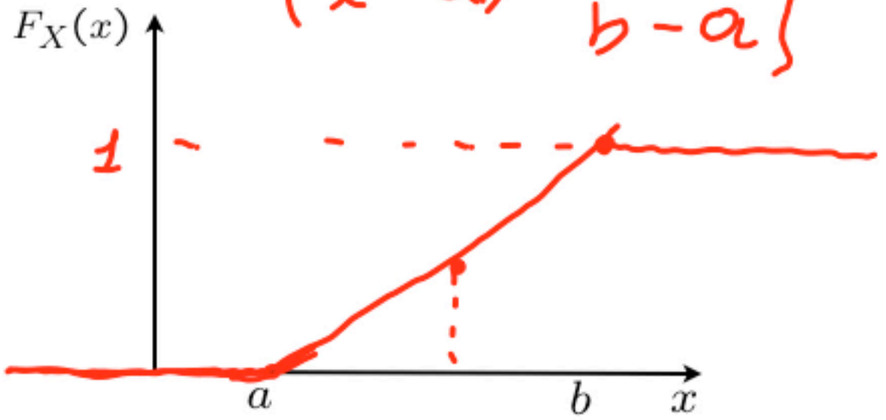


- Continuous random variables:

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

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

$$(x-a) \cdot \frac{1}{b-a}$$



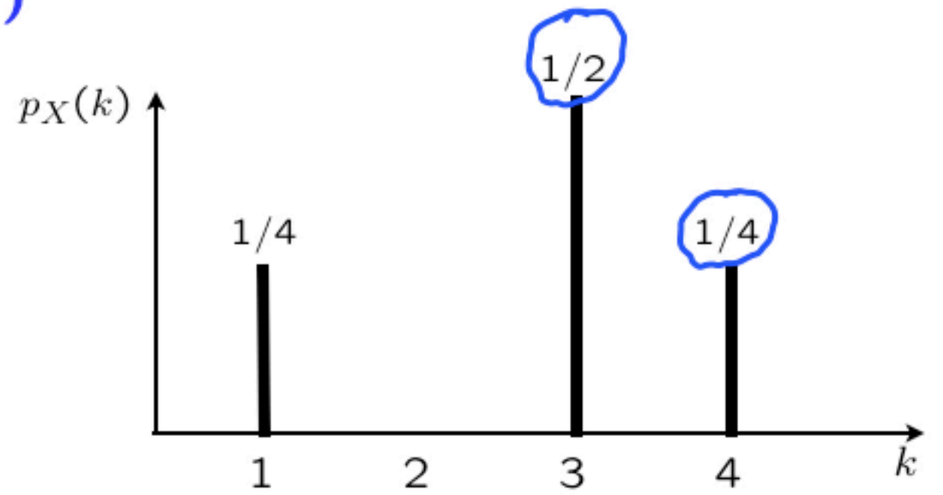
$$\frac{dF_X}{dx}(x) = f_X(x)$$

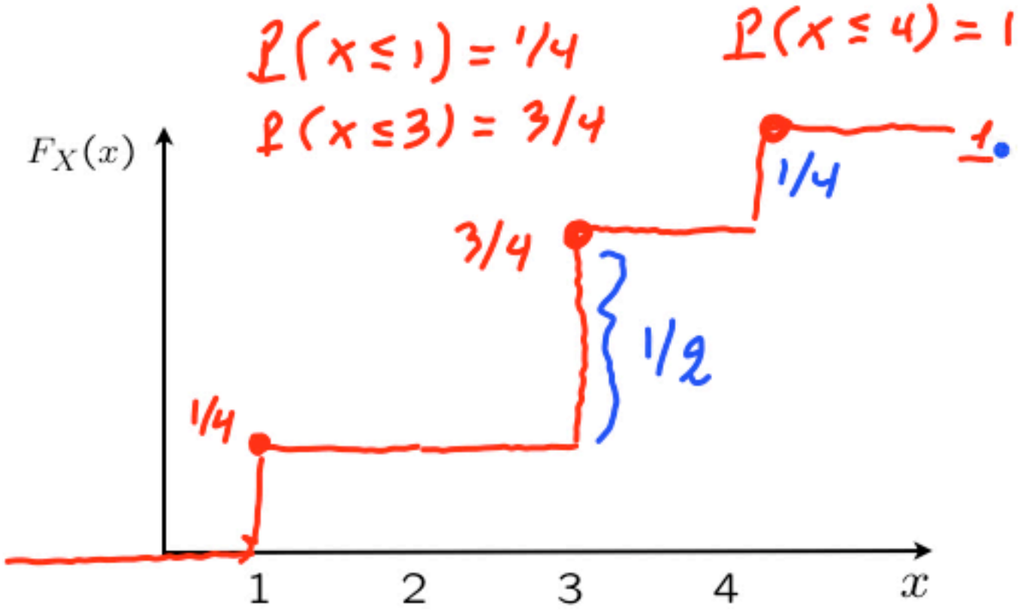
Cumulative distribution function (CDF)

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

- Discrete random variables:

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





General CDF properties

$$F_X(x) = \mathbf{P}(X \leq x)$$

- Non-decreasing

If $y \geq x \Rightarrow F_x(y) \geq F_x(x)$

- $F_X(x)$ tends to 1, as $x \to \infty$
- $F_X(x)$ tends to 0, as $x \to -\infty$

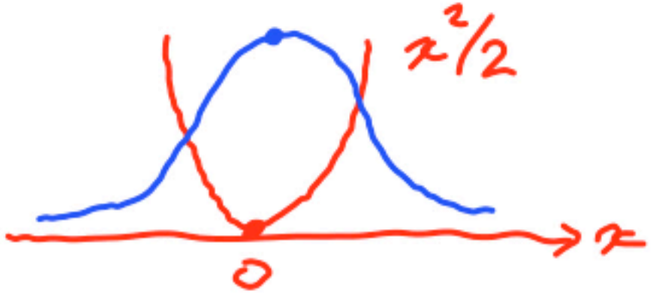
Normal (Gaussian) random variables

- 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:

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

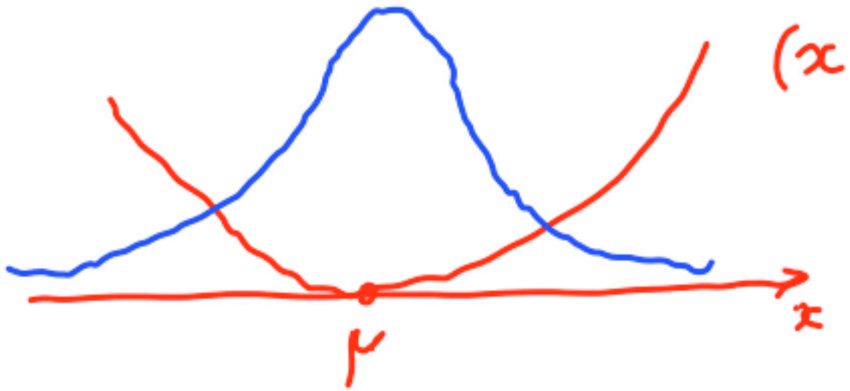
- $\mathbf{E}[X] = 0$
- $\mathrm{var}(X) = 1$

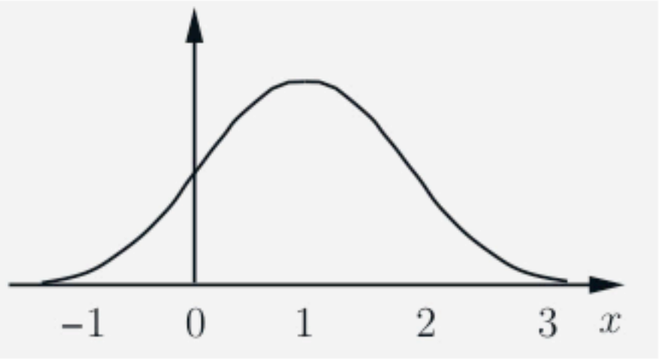
integrate by parts

General normal (Gaussian) random variables

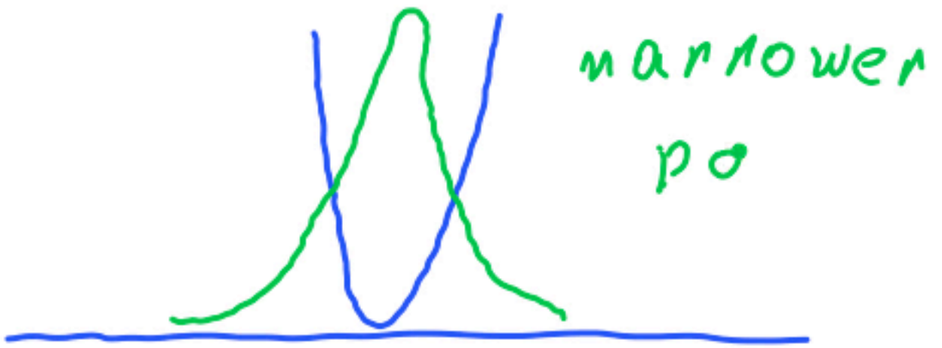
- General normal $N(\mu, \sigma^2)$: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$

$\sigma > 0$





σ small



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

Linear functions of a normal random variable

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

$\mathbf{E}[Y] = a\mu + b$

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

- Fact (will prove later in this course):

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

- Special case: $a = 0$?

$Y = b$ discrete

$N(b, 0)$

Standard normal tables

- No closed form available for CDF

but have tables, for the standard normal

$Y \sim N(0,1)$

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

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

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

$\Phi(-2) =$

$= 1 - \Phi(2)$

$= 1 - .9772$

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

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

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

Calculating normal probabilities

- Express an event of interest in terms of standard normal

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

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

(the middle term: $\frac{X-6}{2}$ → st. normal)

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

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

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

	.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

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

$\mathbf{P}(A) > 0$

$$p_X(x) = \mathbf{P}(X = x)$$

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

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

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

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

$$\mathbf{P}(X \in B) = \int_B f_X(x)\, dx$$

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

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

Def

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

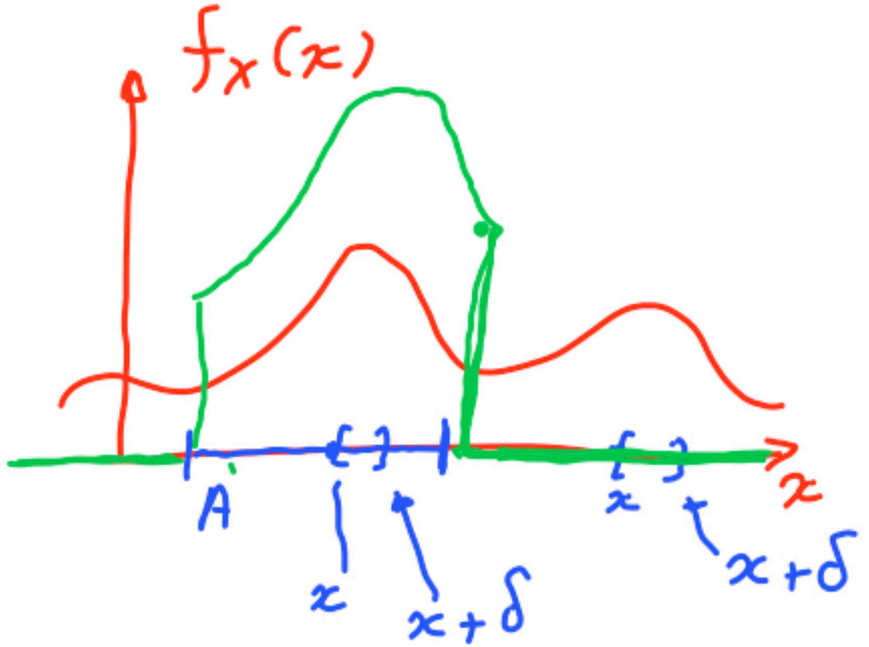
$$\int f_{X|A}(x)\, dx = 1$$

Conditional PDF of X, given that $X \in A$

$$\mathbf{P}(x \le X \le x + \delta \mid X \in A) \approx f_{X|X\in A}(x) \cdot \delta$$

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

$$= \frac{\mathbb{P}(x \le X \le x+\delta)}{\mathbb{P}(A)} \approx \frac{f_X(x)\,\delta}{\mathbb{P}(A)}$$



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

Conditional expectation of X, given an event

$$\mathbf{E}[X] = \sum_x x p_X(x) \qquad \mathbf{E}[X] = \int x f_X(x)\, dx$$

$$\mathbf{E}[X \mid A] = \sum_x x p_{X|A}(x) \qquad \mathbf{E}[X \mid A] = \int x f_{X|A}(x)\, dx$$

Def

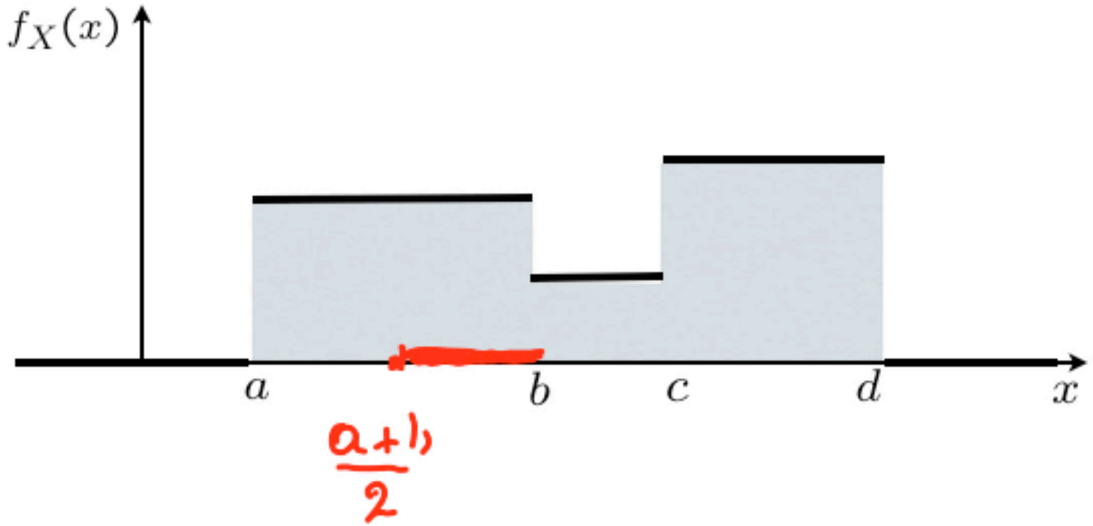
Expected value rule:

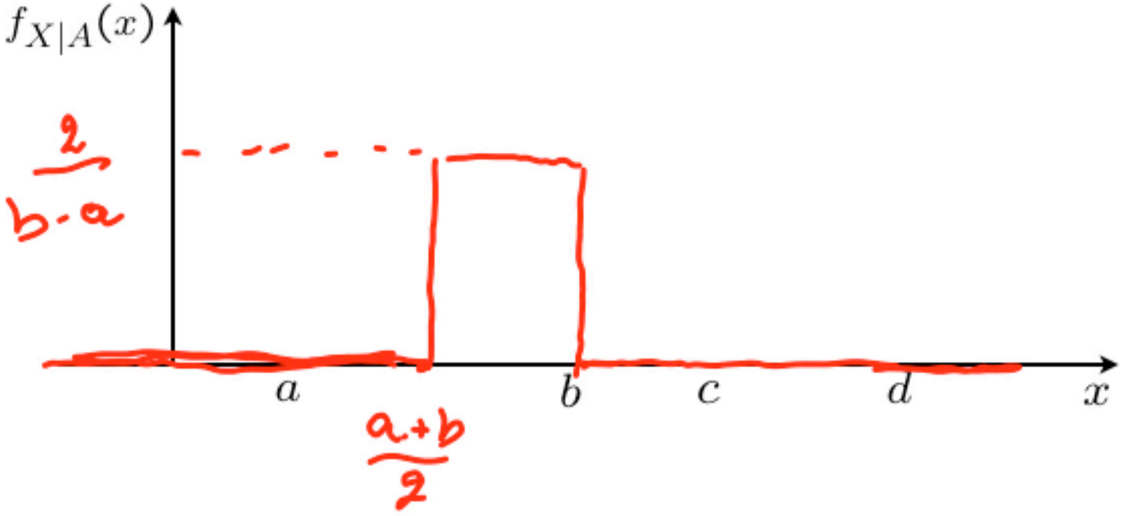
$$\mathbf{E}[g(X)] = \sum_x g(x) p_X(x) \qquad \mathbf{E}[g(X)] = \int g(x) f_X(x)\, dx$$

$$\mathbf{E}[g(X) \mid A] = \sum_x g(x) p_{X|A}(x) \qquad \mathbf{E}[g(X) \mid A] = \int g(x) f_{X|A}(x)\, dx$$

Example

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





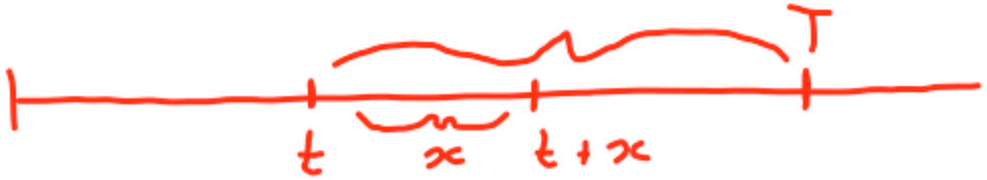
$$\mathbf{E}[X \mid A] = \frac{1}{2} \cdot \frac{a+b}{2} + \frac{1}{2} b$$

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

$$\mathbf{E}[X^2 \mid A] = \int_{\frac{a+b}{2}}^{b} \frac{2}{b-a} \cdot x^2 \, dx$$

Memorylessness of the exponential PDF

- Do you prefer a used or a new "exponential" light bulb? **Probabilistically identical!**
- Bulb lifetime T: exponential(λ)



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

  - we are told that $T > t$
  - r.v. X: remaining lifetime $= T - t$

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

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

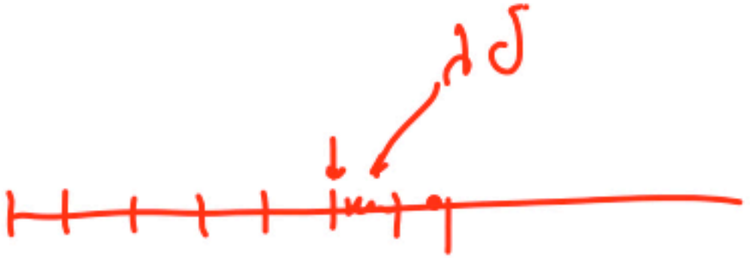
$$= \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x}$$

Memorylessness of the exponential PDF

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

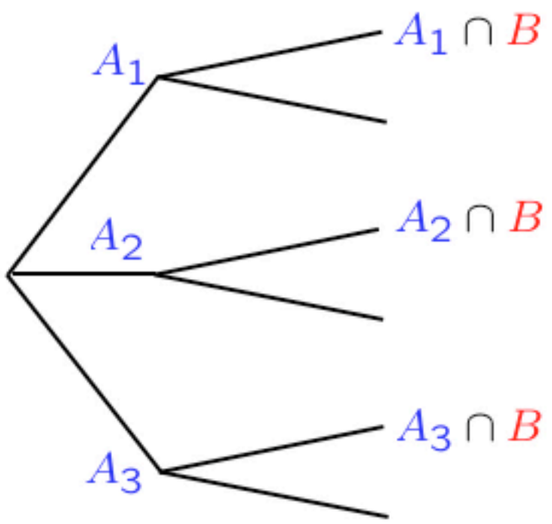
$\mathbf{P}(0 \leq T \leq \delta) \approx f_T(0) \cdot \delta = \lambda \delta$

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



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

Total probability and expectation theorems



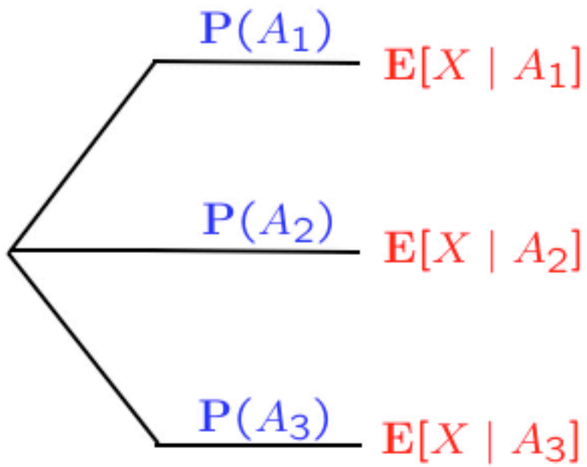
$$\mathbf{P}(B) = \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)$$

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

$$F_X(x) = P(X \le x) = P(A_1)\, P(X \le x \mid A_1) + \cdots$$

$$= P(A_1)\, F_{X|A_1}(x) + \cdots$$

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



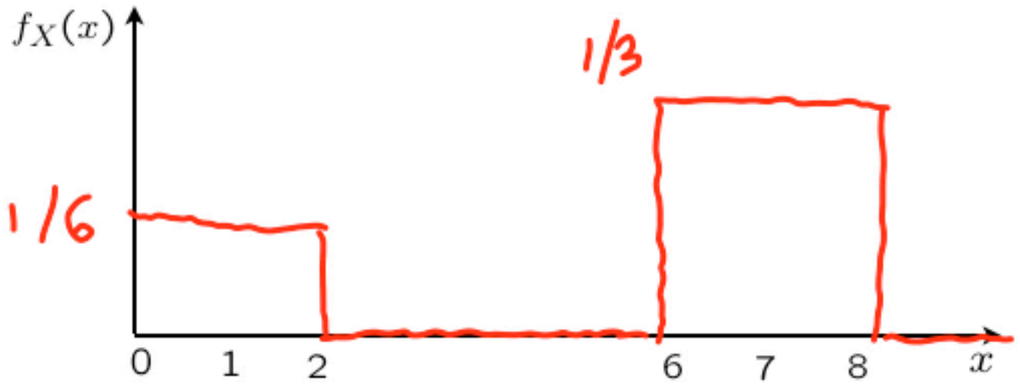
$$\int x f_X(x)\, dx = P(A_1) \int x f_{X|A_1}(x)\, dx + \cdots$$

$$\mathbf{E}[X] = \mathbf{P}(A_1)\mathbf{E}[X \mid A_1] + \cdots + \mathbf{P}(A_n)\mathbf{E}[X \mid A_n]$$

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

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



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

- $$\mathbf{E}[X] = \mathbf{P}(A_1)\mathbf{E}[X \mid A_1] + \cdots + \mathbf{P}(A_n)\mathbf{E}[X \mid A_n]$$

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

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

Is X continuous? No

$P(X=1) = 1/2$

X is mixed

Y discrete

Z continuous

$$X = \begin{cases} Y, & \text{with probability } p \\ Z, & \text{with probability } 1-p \end{cases}$$

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

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

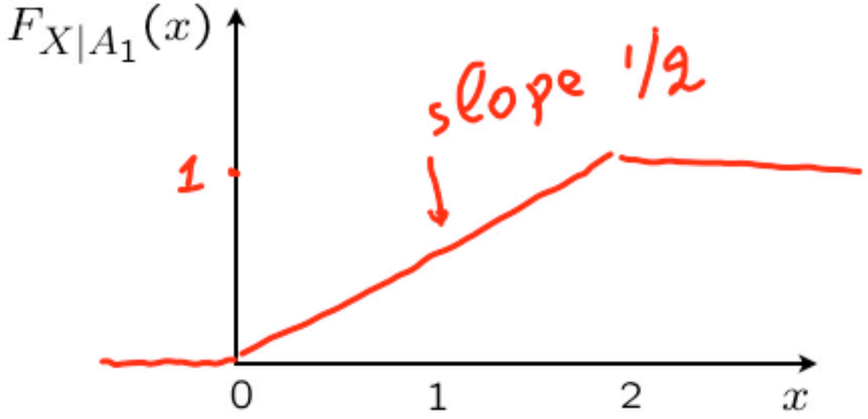
$$E[X] = p\, E[Y] + (1-p)\, E[Z]$$

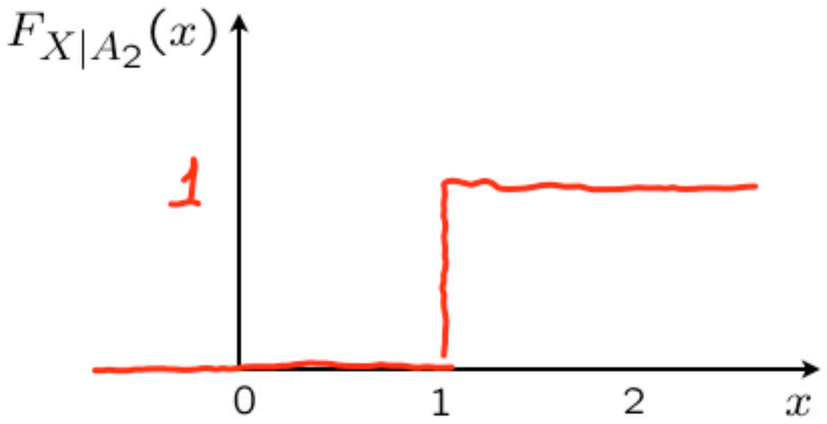
Mixed distributions

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

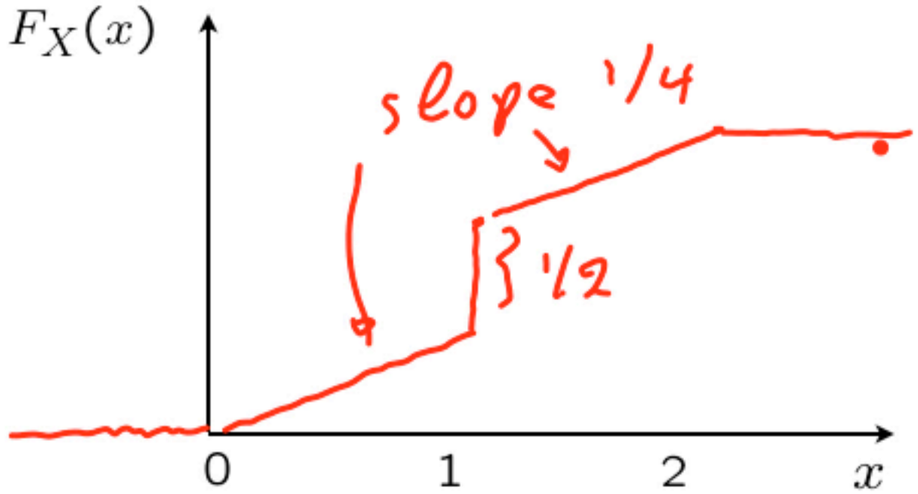
A_1

A_2





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



Jointly continuous r.v.'s and joint PDFs

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

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

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

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

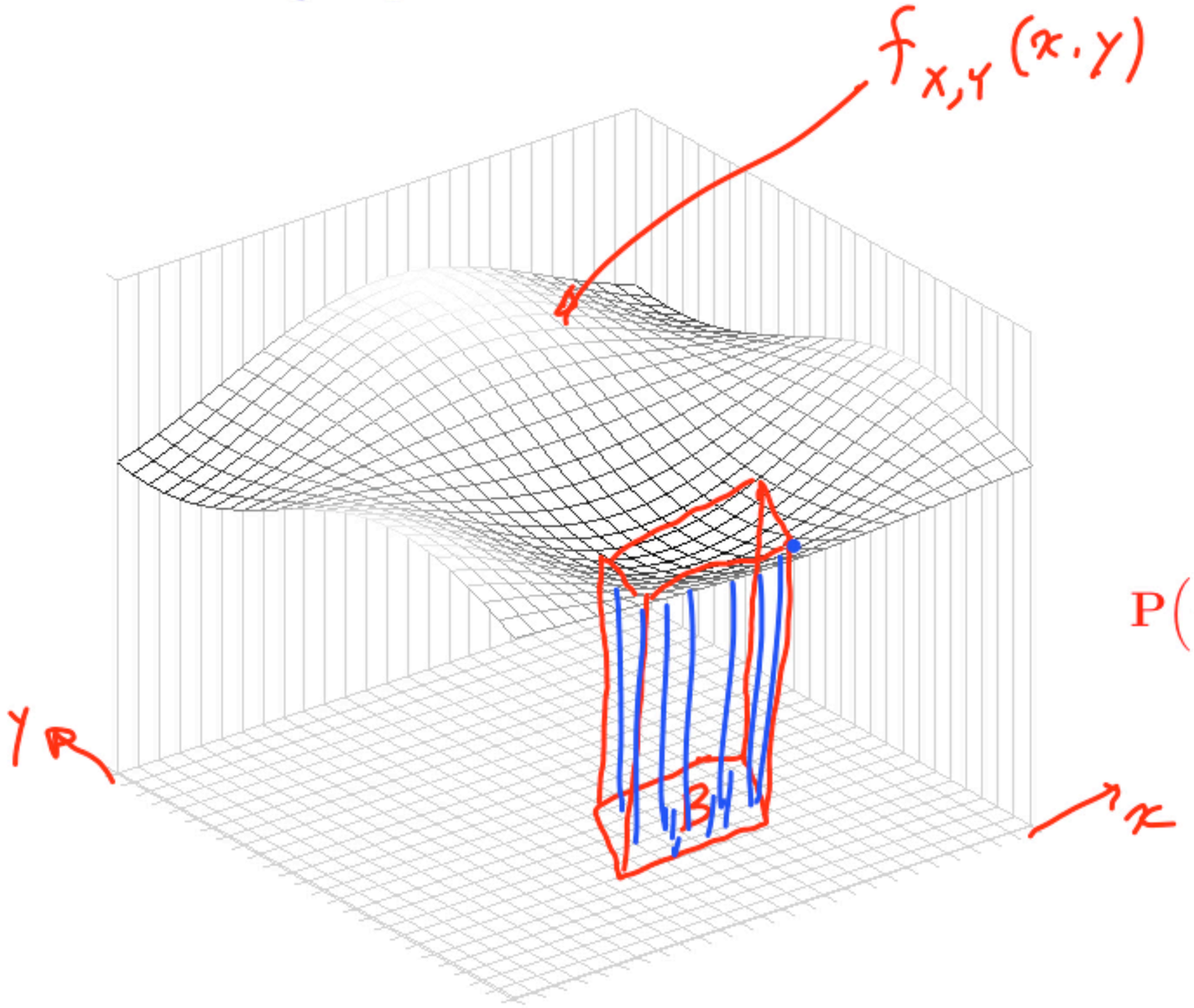
$$\mathbf{P}\Big((X,Y) \in B\Big) = \int\int_{(x,y)\in B} f_{X,Y}(x,y)\,dx\,dy \quad \bullet$$

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

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

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

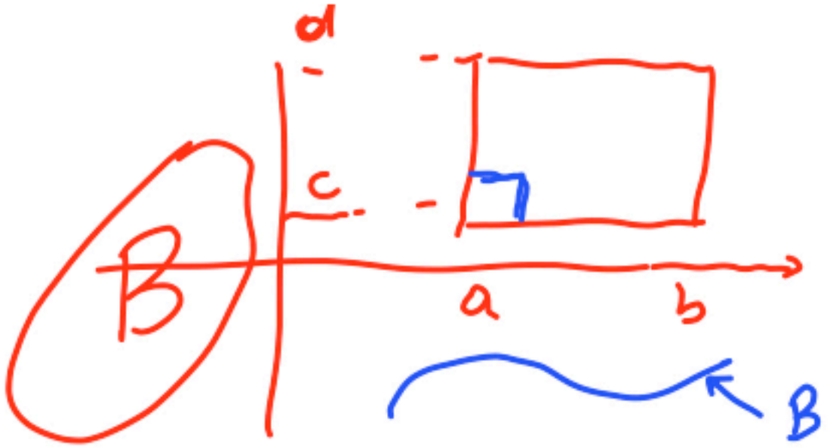
Visualizing a joint PDF



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

On joint PDFs

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

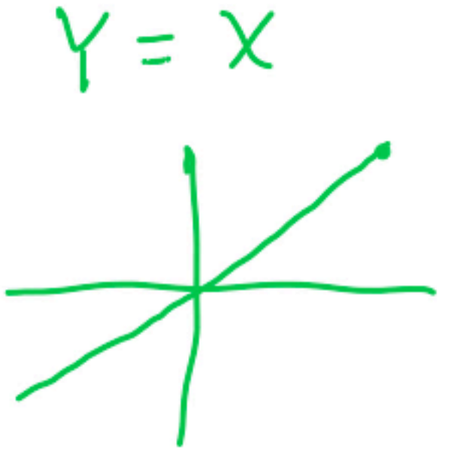


$$\mathbf{P}(a\le X\le b,\ c\le Y\le d) = \int_c^d\int_a^b f_{X,Y}(x,y)\,dx\,dy$$

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

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

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

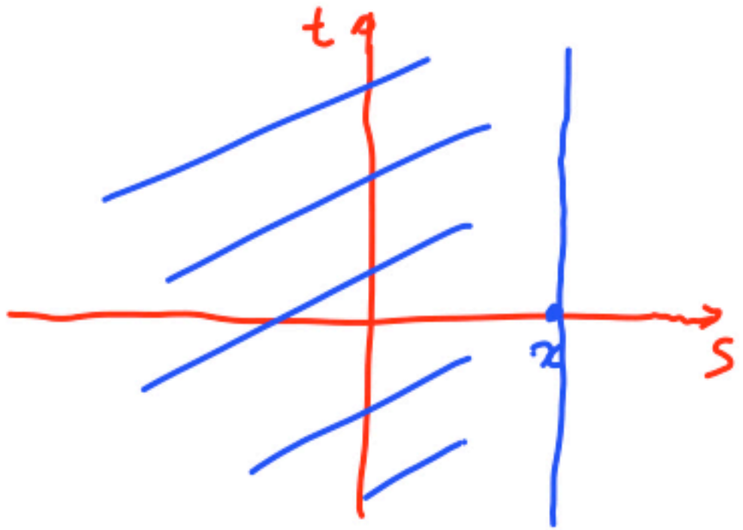


From the joint to the marginals

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

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

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

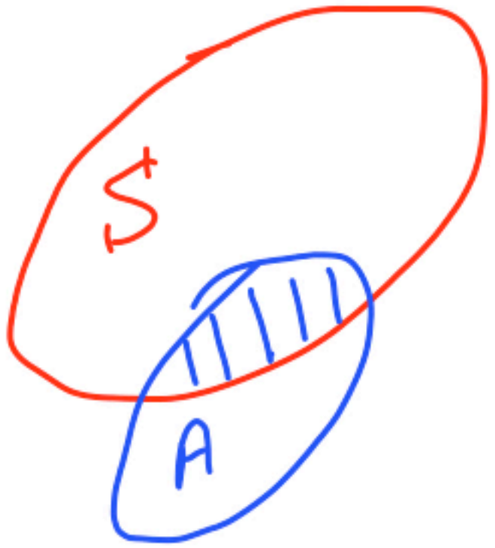


$f_x(x) = \dfrac{dF_x}{dx}(x) = [\ ]$

Uniform joint PDF on a set S

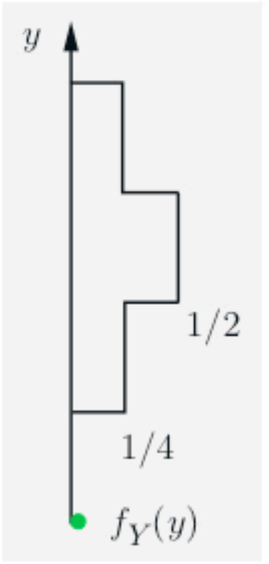
$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\text{area of } S}, & \text{if } (x,y) \in S, \\ 0, & \text{otherwise.} \end{cases}$$

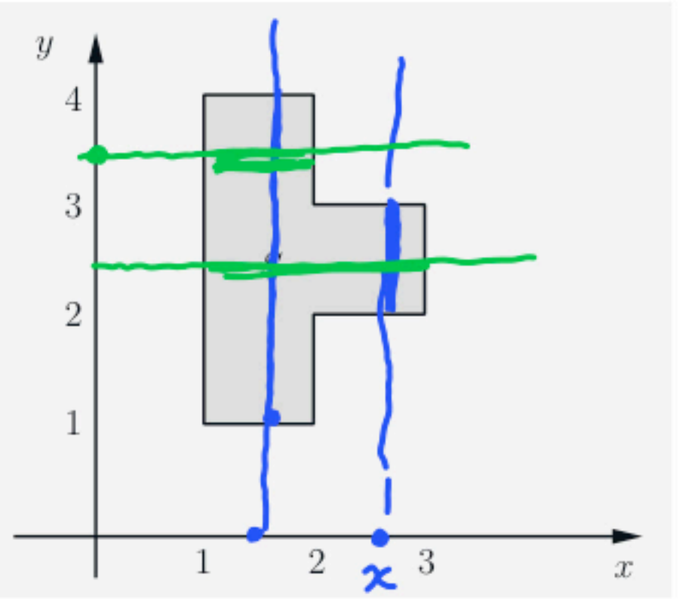
$f_{X,Y} = \frac{1}{4}$

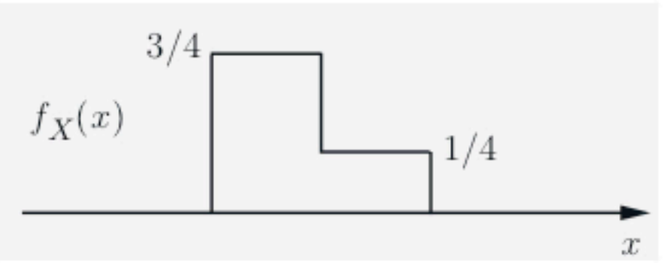


$= \dfrac{\text{area}(A \cap S)}{\text{area}(S)}$

$P(A)$







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:

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

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

Linearity of expectations

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

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

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

The joint CDF

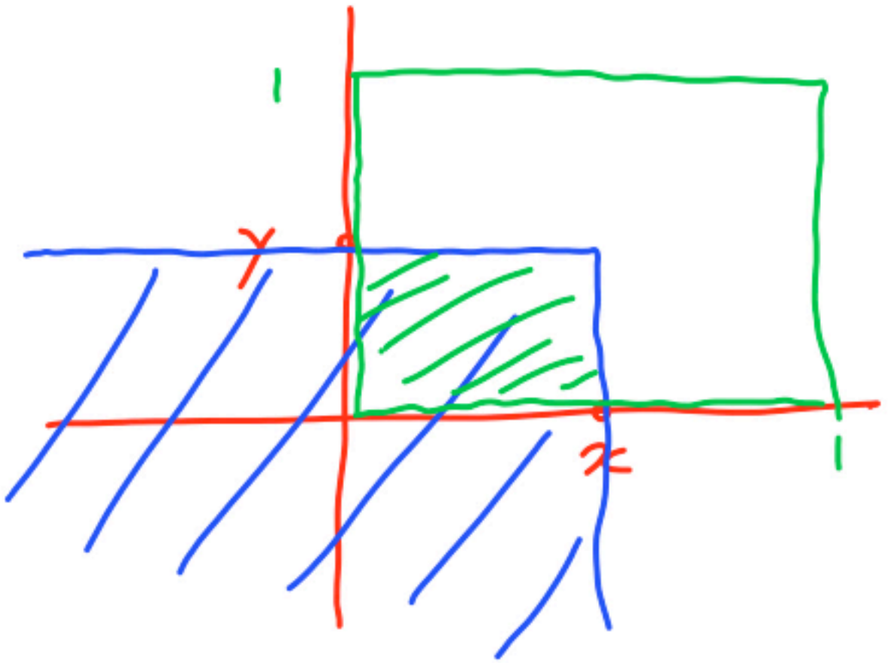
$$F_X(x) = \mathbf{P}(X \le x) = \int_{-\infty}^{x} f_X(t)\,dt \qquad f_X(x) = \frac{dF_X}{dx}(x)$$

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

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

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

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



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.

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

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

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

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

$Y = y$ $\quad$ $Y \approx y$

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

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

Comments on conditional PDFs

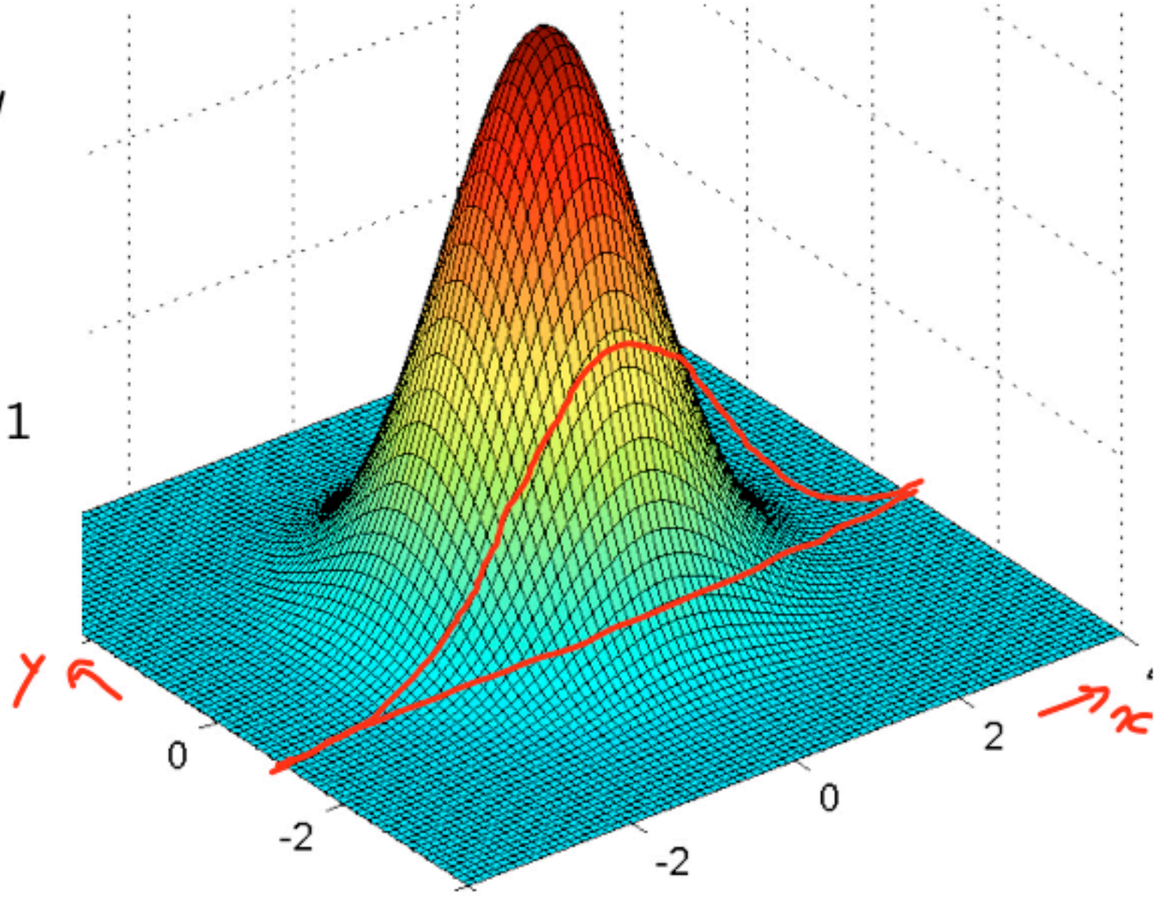
$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

- $f_{X|Y}(x \mid y) \geq 0$
- Think of value of Y as fixed at some y

  shape of $f_{X|Y}(\cdot \mid y)$: slice of the joint
- $$\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)\, dx = \frac{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx}{f_Y(y)} = 1$$
- Multiplication rule:

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

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



Total probability and expectation theorems

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

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

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

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

$$\mathbf{E}[X] = \sum_y p_Y(y) \mathbf{E}[X \mid Y = y]$$

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} f_Y(y) \mathbf{E}[X \mid Y = y]\, dy$$

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

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

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

- Expected value rule...

$$E[g(X) \mid Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\, dx$$

Independence

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

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

$f_{Y|X} = f_Y$

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

- 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:** $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$

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

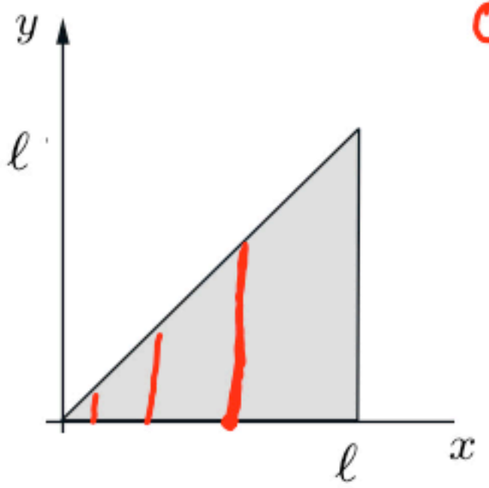
$g(X)$ and $h(Y)$ are also independent: $\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)] \cdot \mathbf{E}[h(Y)]$

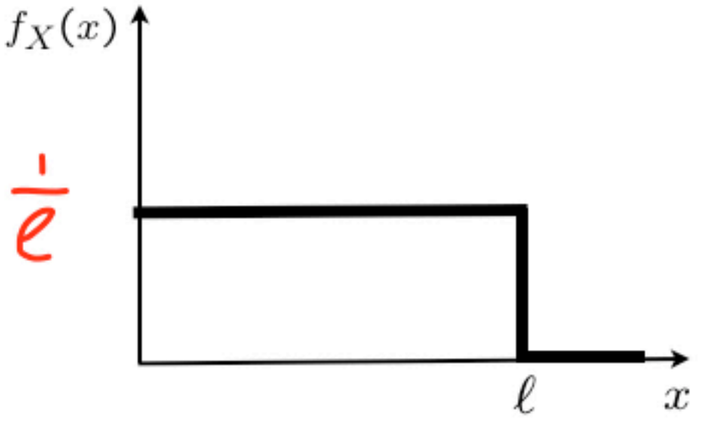
Stick-breaking example

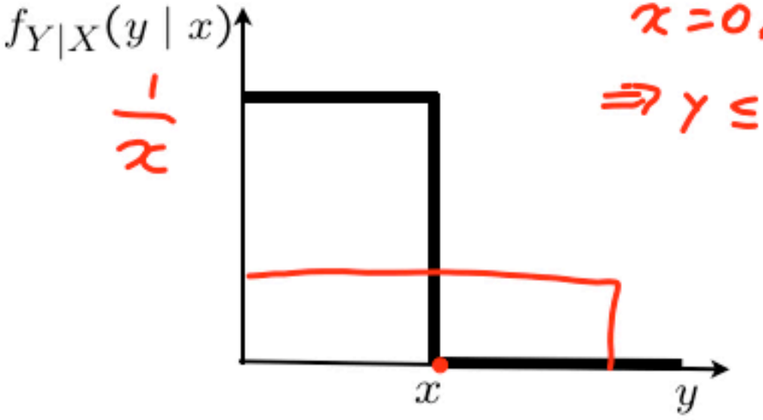
- Break a stick of length ℓ twice
  - first break at X: uniform in $[0, \ell]$
  - second break at Y: uniform in $[0, X]$

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

$0 \le y \le x \le \ell$

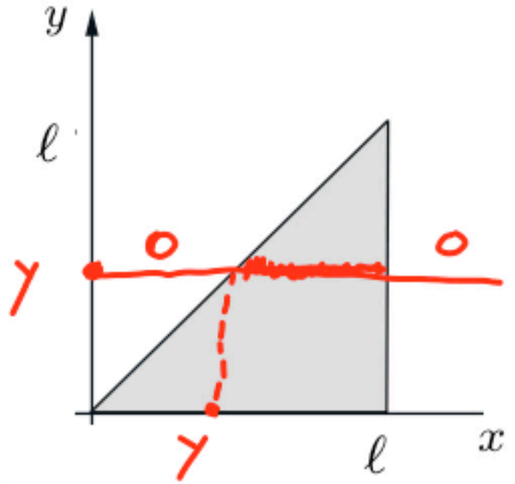






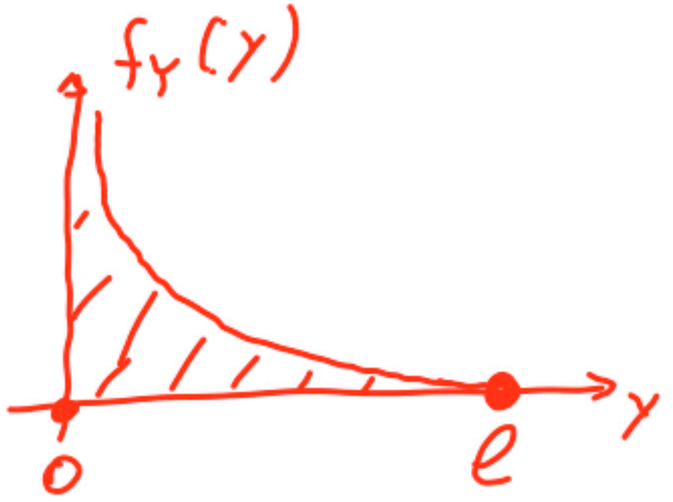
Stick-breaking example

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



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

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



- Using total expectation theorem:

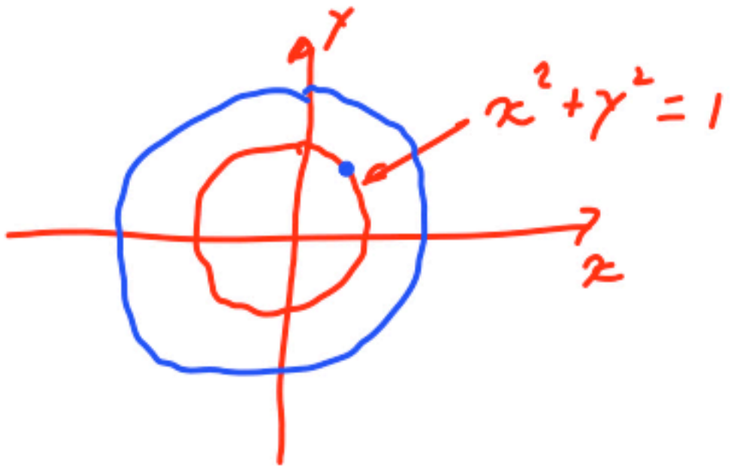
$$E[Y] = \int_0^\ell \frac{1}{\ell}\,E[Y|X=x]\,dx = \int_0^\ell \underbrace{\frac{1}{\ell}}_{f_X(x)}\,\frac{x}{2}\,dx = \frac{1}{2}E[X] = \frac{1}{2}\cdot\frac{\ell}{2} = \frac{\ell}{4}.$$

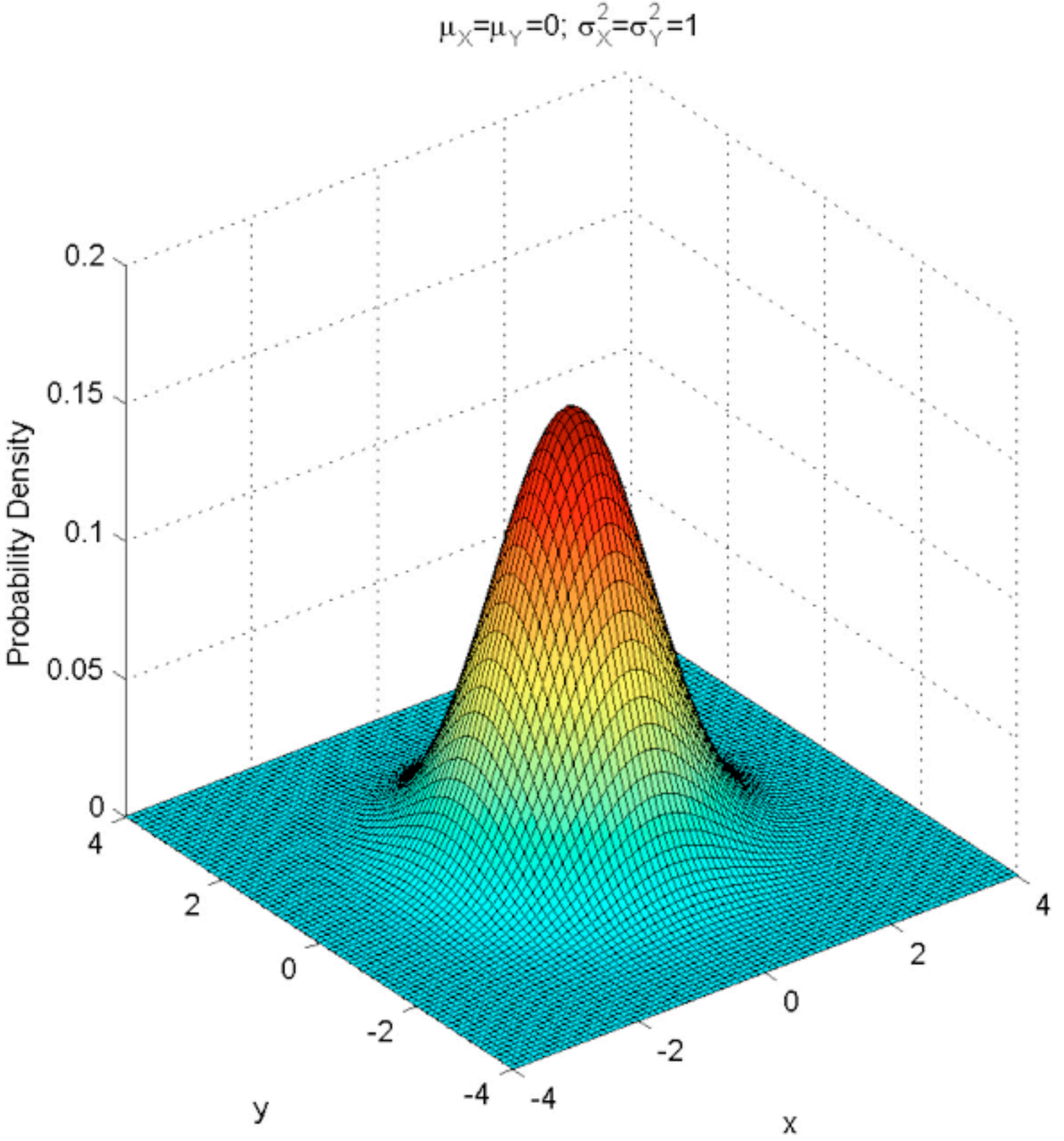
Independent standard normals

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

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

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

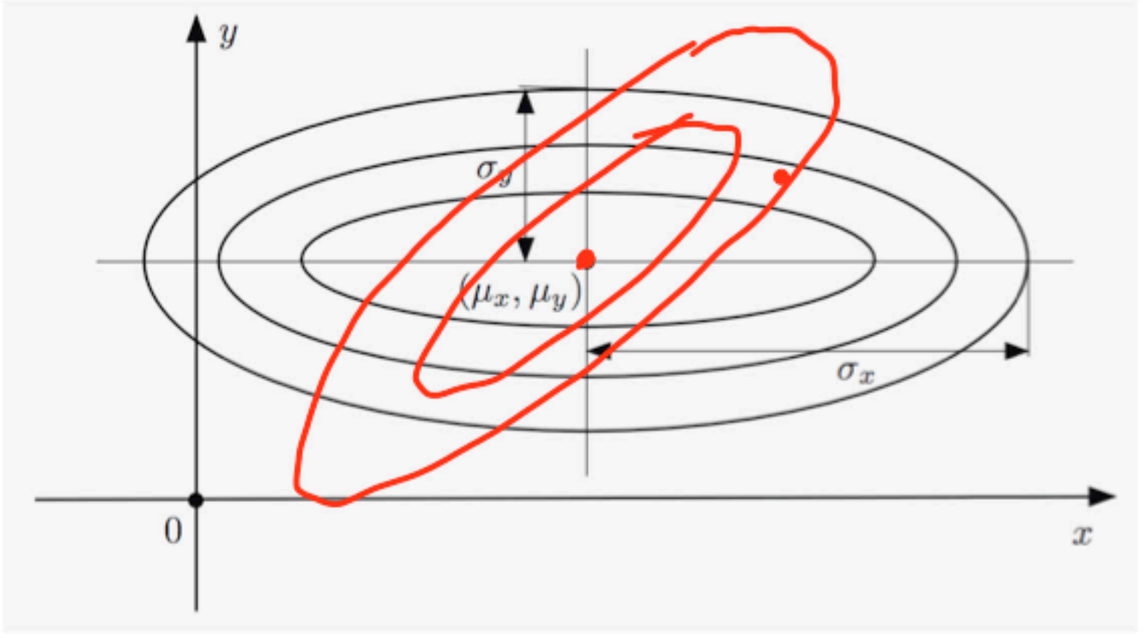


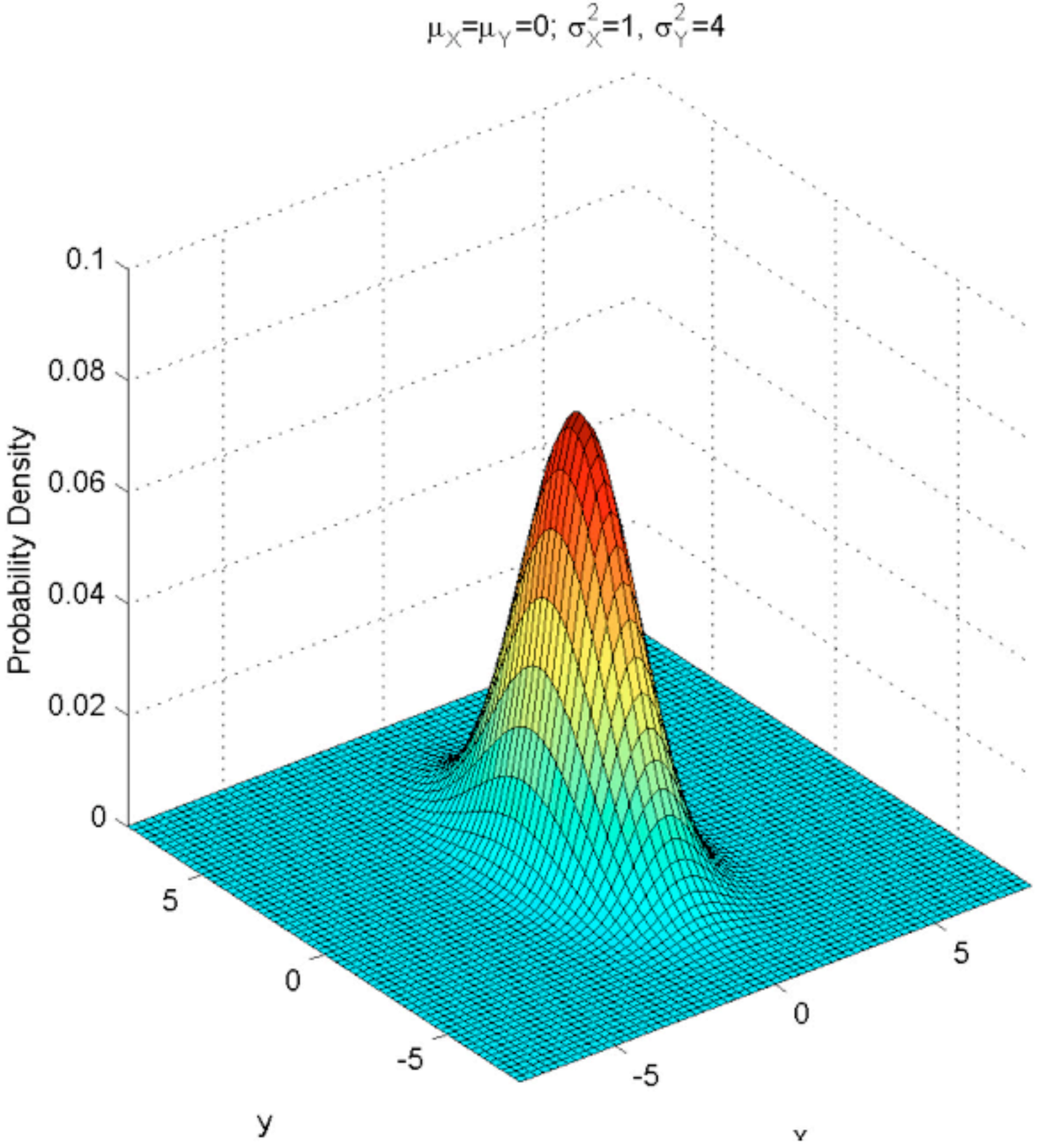


Independent normals

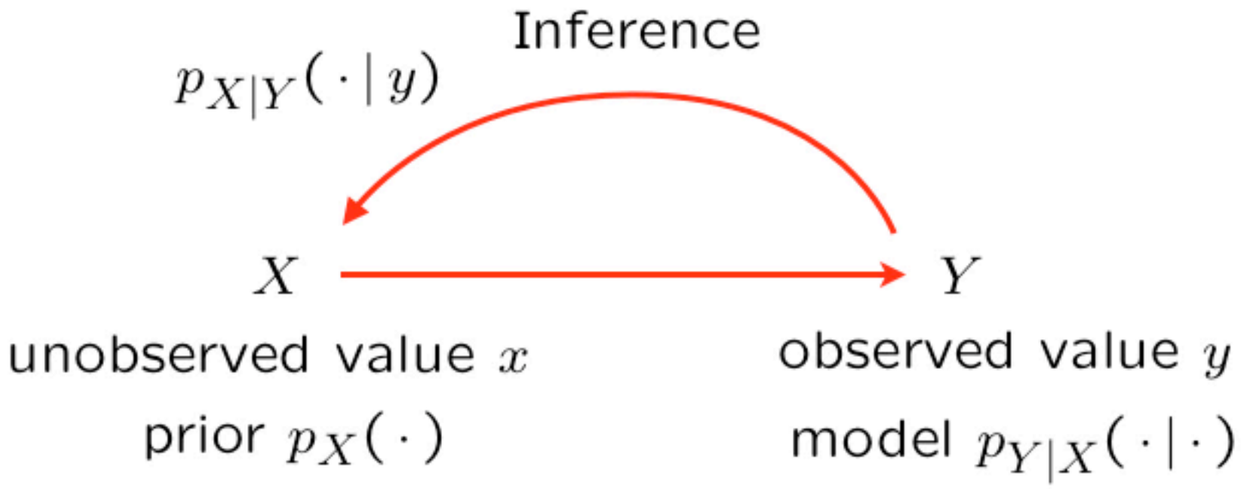
$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left\{ -\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2} \right\}$$





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

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

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

The Bayes rule — one discrete and one continuous random variable

K: discrete $\qquad$ Y: continuous

$$P(K=k,\ y\le Y\le y+\delta) \qquad \delta>0,\ \delta\approx 0$$

$$= P(K=k)\,P(y\le Y\le y+\delta \mid K=k) \approx p_K(k)\,f_{Y|K}(y|k)\,\delta$$

$$= P(y\le Y\le y+\delta)\,P(K=k \mid y\le Y\le y+\delta) \approx f_Y(y)\,\delta\,p_{K|Y}(k|y)$$

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

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

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

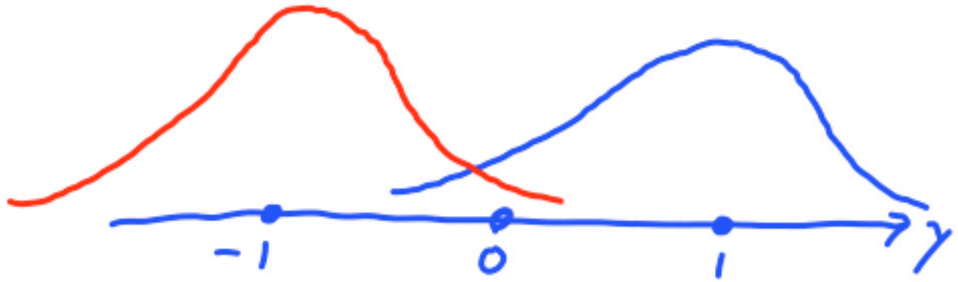
$$p_K(k) = \int f_Y(y')\,p_{K|Y}(k\,|\,y')\,dy'$$

The Bayes rule — discrete unknown, continuous measurement

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

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

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



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

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

$k = -1, +1$

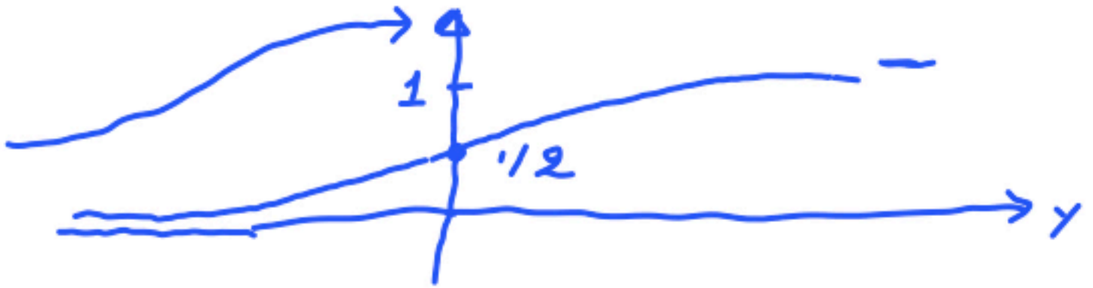
$$p_{K|Y}(k|y) = \frac{p_K(k)\, f_{Y|K}(y|k)}{f_Y(y)}$$

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

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

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

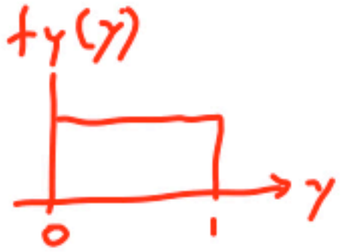
algebra



The Bayes rule — continuous unknown, discrete measurement

- measurement K: Bernoulli with parameter Y

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



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

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

- Distribution of Y given that $K = 1$? $f_{Y|K}(y|1)$

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

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

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

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

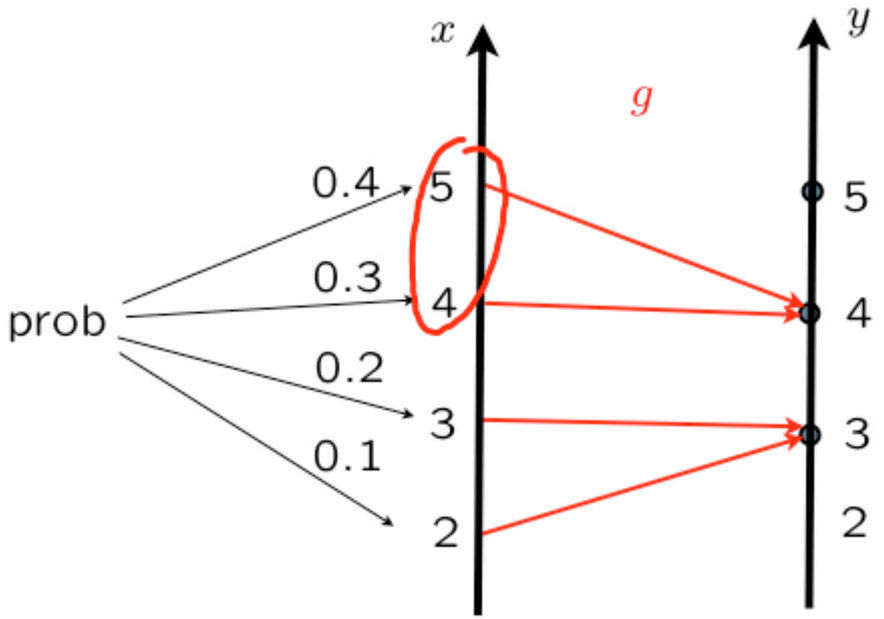
LECTURE 11: Derived distributions

- 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

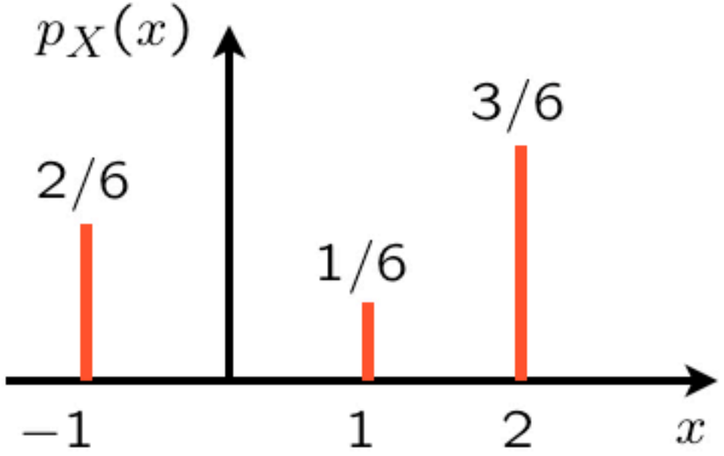
$Y = g(X)$



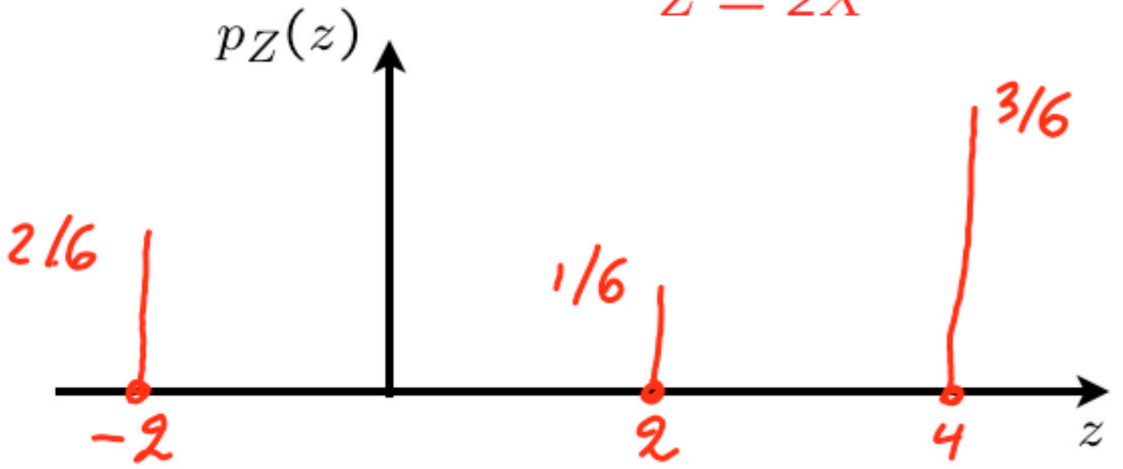
$$p_Y(4) = P(Y=4)$$
$$= P(X=4) + P(X=5)$$
$$= p_X(4) + p_X(5) = 0.3 + 0.4$$

$$p_Y(y) = \mathbf{P}(g(X)=y) = \sum_{x:\, g(x)=y} p_X(x)$$

A linear function of a discrete r.v.



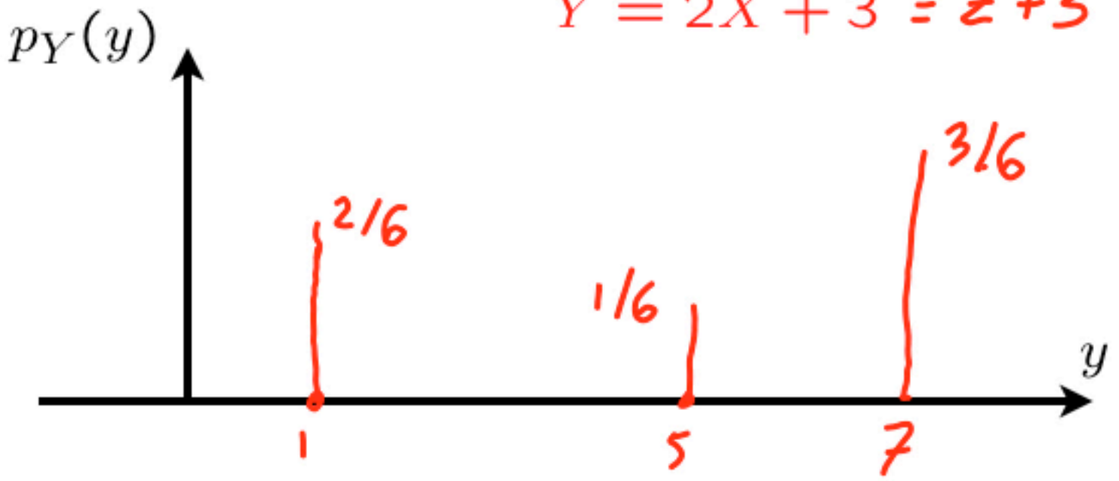
$Z = 2X$



$$P_Y(y) = P(Y = y) = P(2X + 3 = y)$$

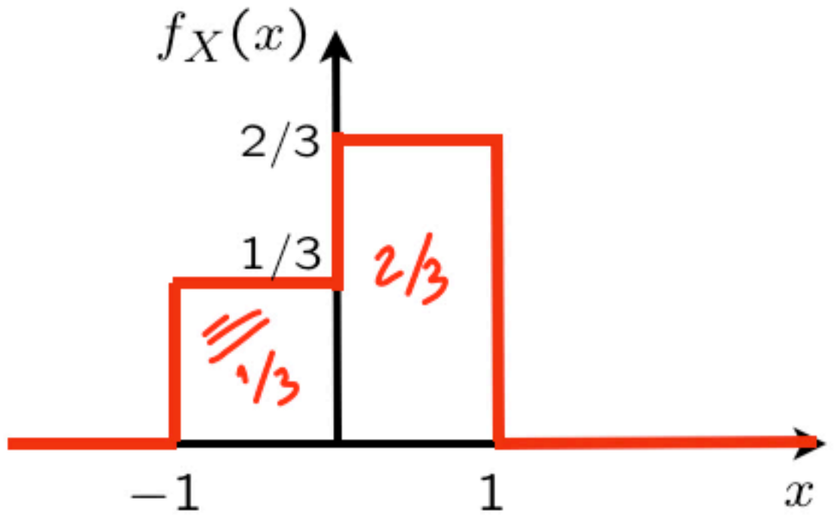
$$= P\left(X = \frac{y-3}{2}\right) = P_X\left(\frac{y-3}{2}\right)$$

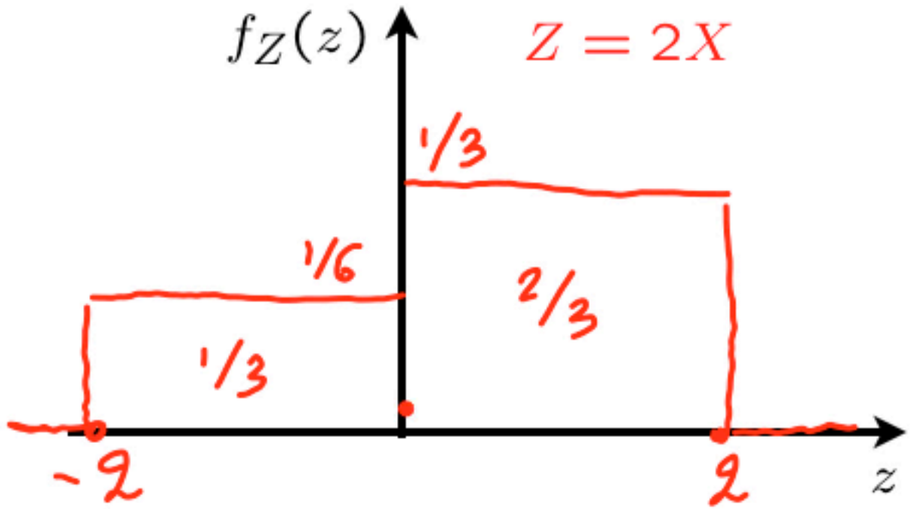
$Y = 2X + 3 = Z + 3$

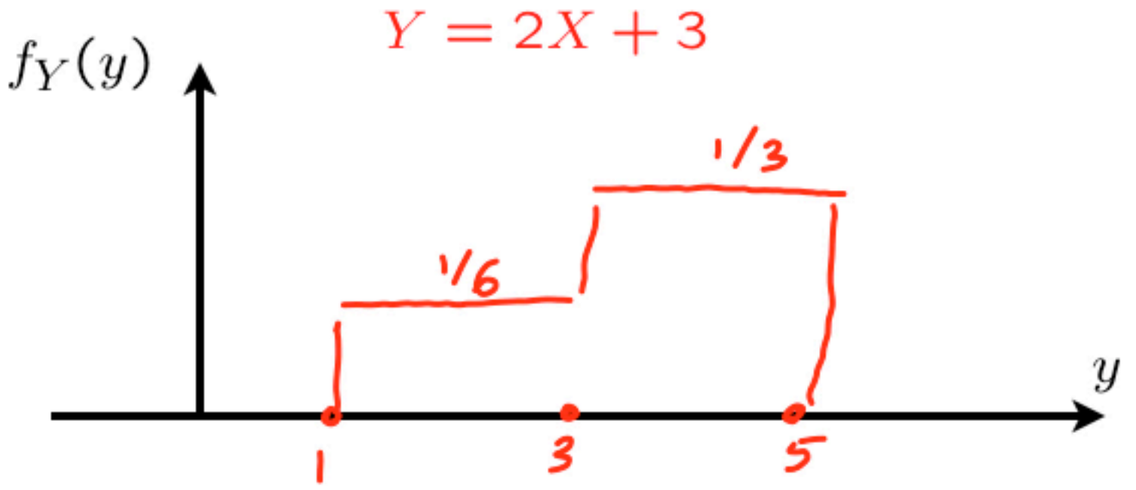


$$Y = aX + b: \quad p_Y(y) = p_X\left(\frac{y-b}{a}\right)$$

A linear function of a continuous r.v.







A linear function of a continuous r.v.

$Y = aX + b$

$a > 0$

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

$$F_Y(y) = P(Y \le y) = P(aX + b \le y)$$

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

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

$a < 0$

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

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

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

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

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

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

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

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

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

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

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

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

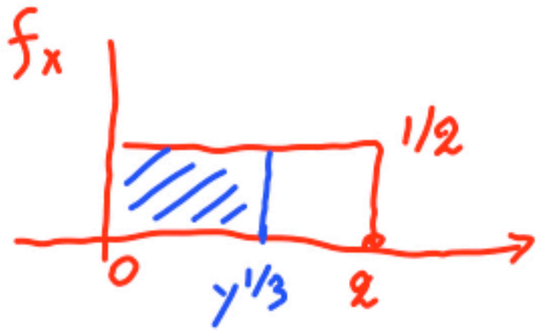
$$= \frac{1}{\sqrt{2\pi}\,\sigma|a|} e^{-(y-b-a\mu)^2/2\sigma^2a^2}$$

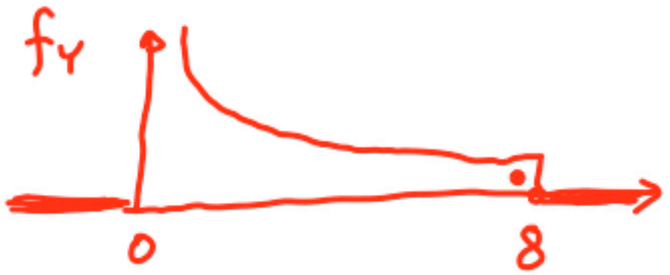
If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

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

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

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





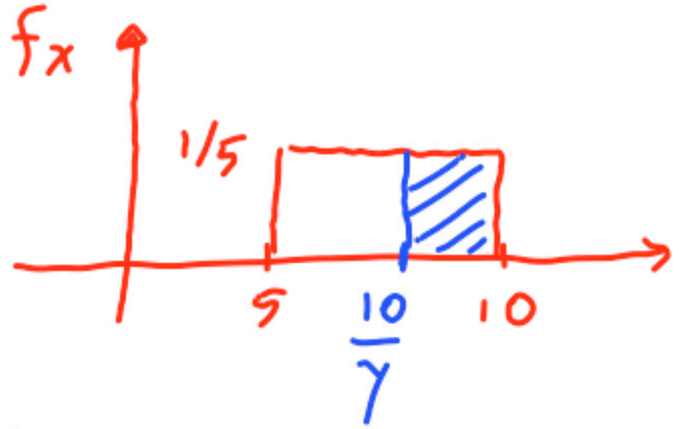
$0 \le y \le 8$

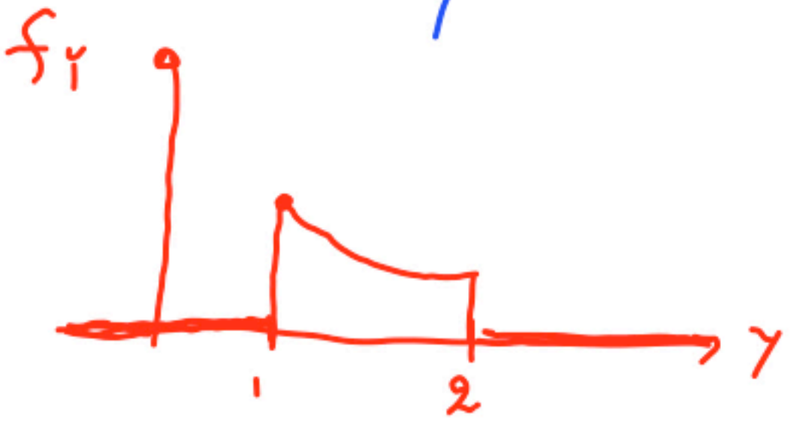
$$F_Y(y) = P(Y \le y) = P(X^3 \le y) = P(X \le y^{1/3}) = \frac{1}{2} y^{1/3}$$

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

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.





$$\text{time} = Y = \frac{10}{X} \qquad 1 \le y \le 2$$

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

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

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

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

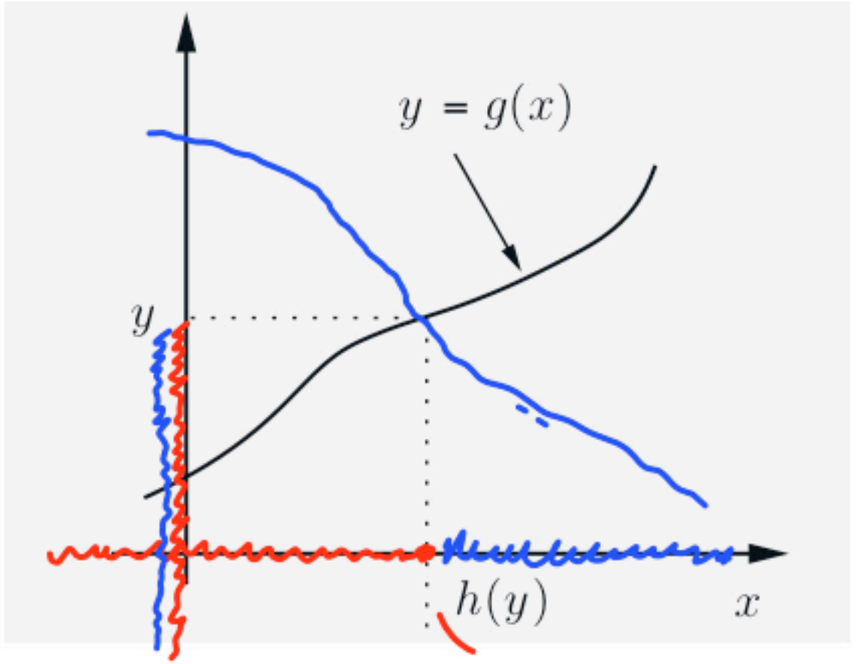
A general formula for the PDF of $Y = g(X)$ when g is monotonic

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

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

Assume g strictly increasing

and differentiable



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

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

$$F_Y(y) = P(Y \le y) = P(X \ge h(y))$$

$$= 1 - P(X \le h(y)) = 1 - F_X(h(y))$$

$$f_Y(y) = 4\, f_X(h(y)) \left| \frac{dh}{dy}(y) \right|$$

inverse function h → decreasing

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

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

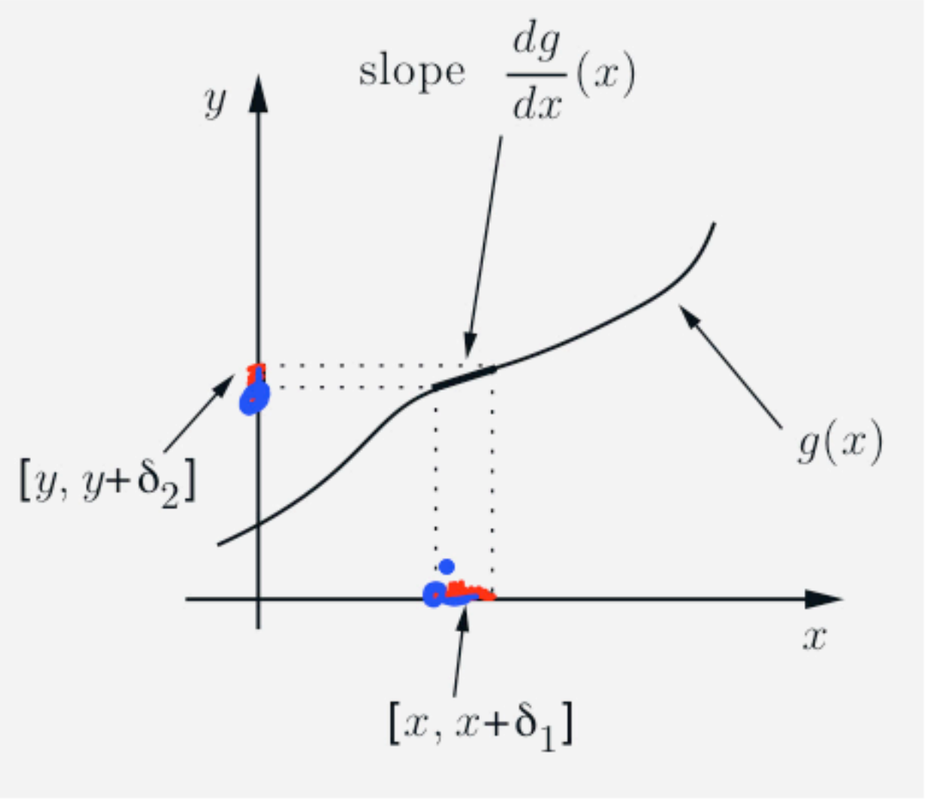
$$f_Y(y) = f_X\big(h(y)\big)\left|\frac{dh}{dy}(y)\right|$$

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

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

$0 \le y \le 1$

An intuitive explanation for the monotonic case



$y = g(x)$ $\quad \delta_2 \approx \delta_1 \frac{dg}{dx}(x)$

$x = h(y)$ $\quad \delta_1 \approx \delta_2 \cdot \frac{dh}{dy}(y)$ (*)

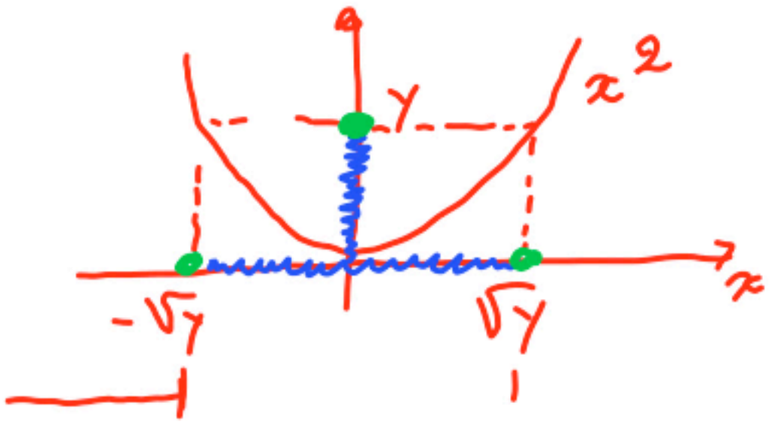
$f_Y(y)\delta_2 \approx P(y \le Y \le y+\delta_2) = P(x \le X \le x+\delta_1)$

$\approx f_X(x)\delta_1 \approx f_X(x)\delta_2 \frac{dh}{dy}(y)$

$f_Y(y) = f_X(x) \frac{dh}{dy}(y)$

$= f_X(h(y)) \frac{dh}{dy}(y)$

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: $y \geq 0$

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(|X| \leq \sqrt{y}) = P(-\sqrt{y} \leq X \leq \sqrt{y})$$

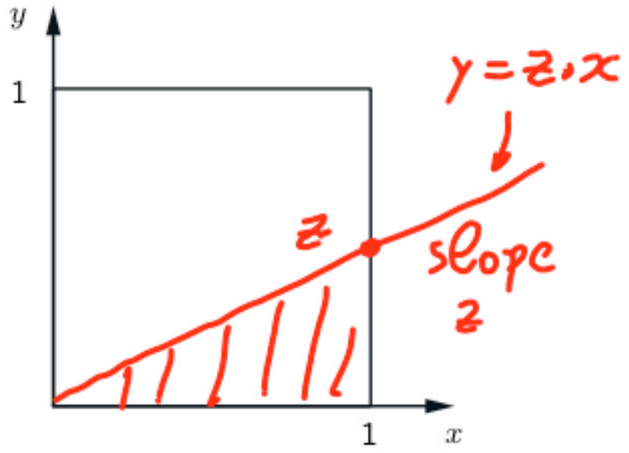
$$= F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

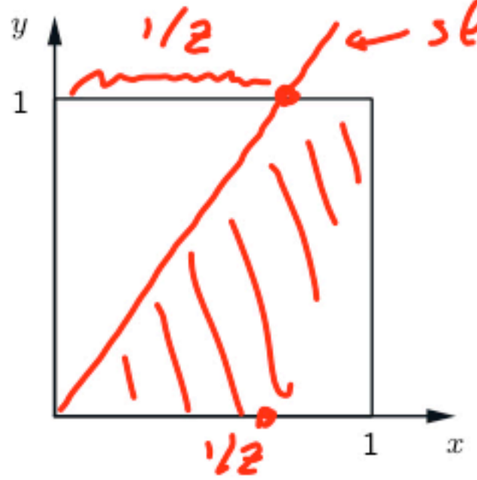
$$f_Y(y) = f_X(\sqrt{y})\frac{1}{2\sqrt{y}} + f_X(-\sqrt{y})\frac{(-1)}{2\sqrt{y}}$$

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

$Y = g(X)$

- Same methodology: find CDF of Z
- Let $Z = Y/X$; X, Y independent, uniform on $[0, 1]$

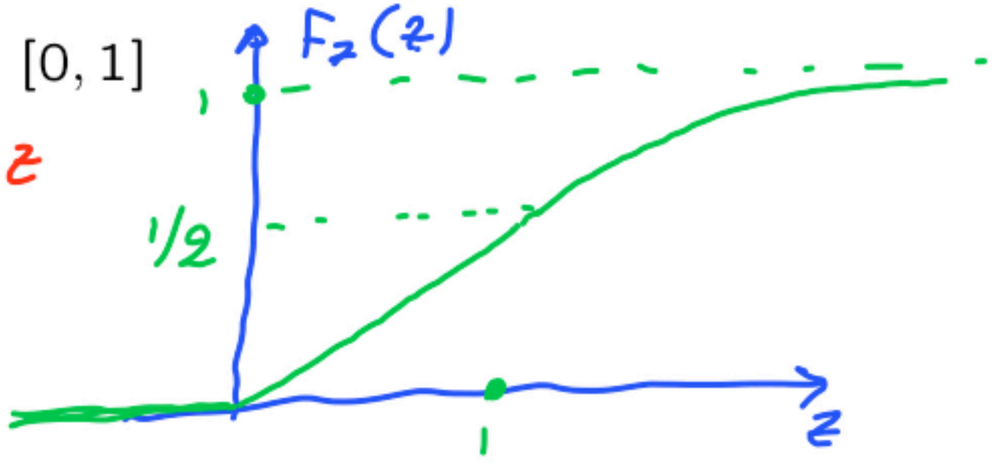


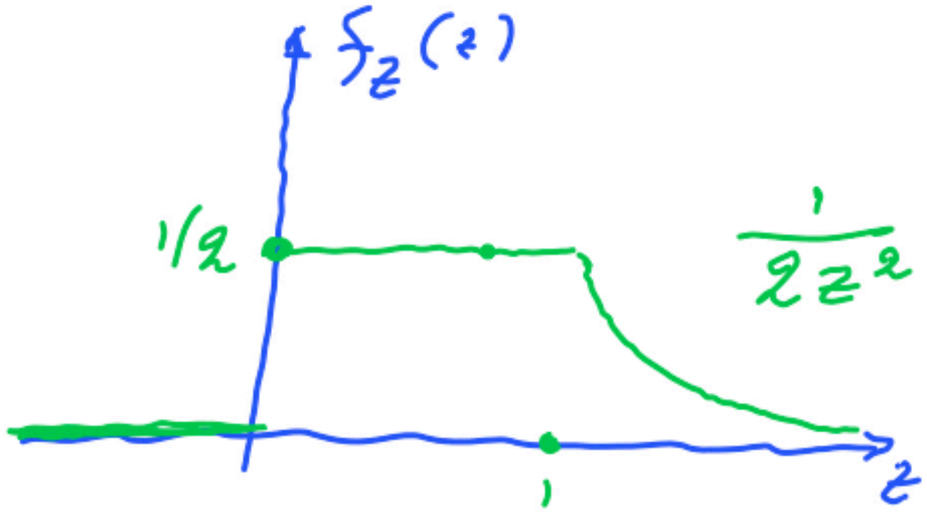


$$F_Z(z) = P\left(\frac{Y}{X} \le z\right) = 0, \quad z < 0$$

$$= \frac{1}{2} \cdot z, \quad 0 \le z \le 1$$

$$= 1 - \frac{1}{2z}, \quad z > 1$$





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

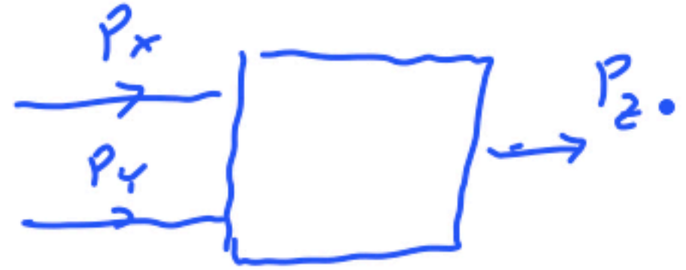
- Covariance and correlation
  - definitions
  - mathematical properties
  - interpretation

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

$$p_Z(z) = \sum_x p_X(x)\, p_Y(z - x)$$

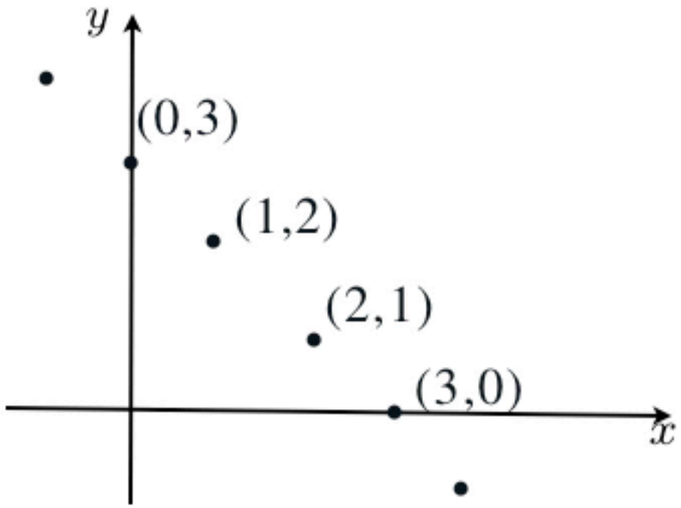
- $Z = X + Y$; X, Y independent, discrete

$g(x,y)$ known PMFs



$$p_Z(3) = \cdots + P(X=0, Y=3) + P(X=1, Y=2) + \cdots$$

$$= \cdots + p_X(0)\, p_Y(3) + p_X(1)\, p_Y(2) + \cdots$$

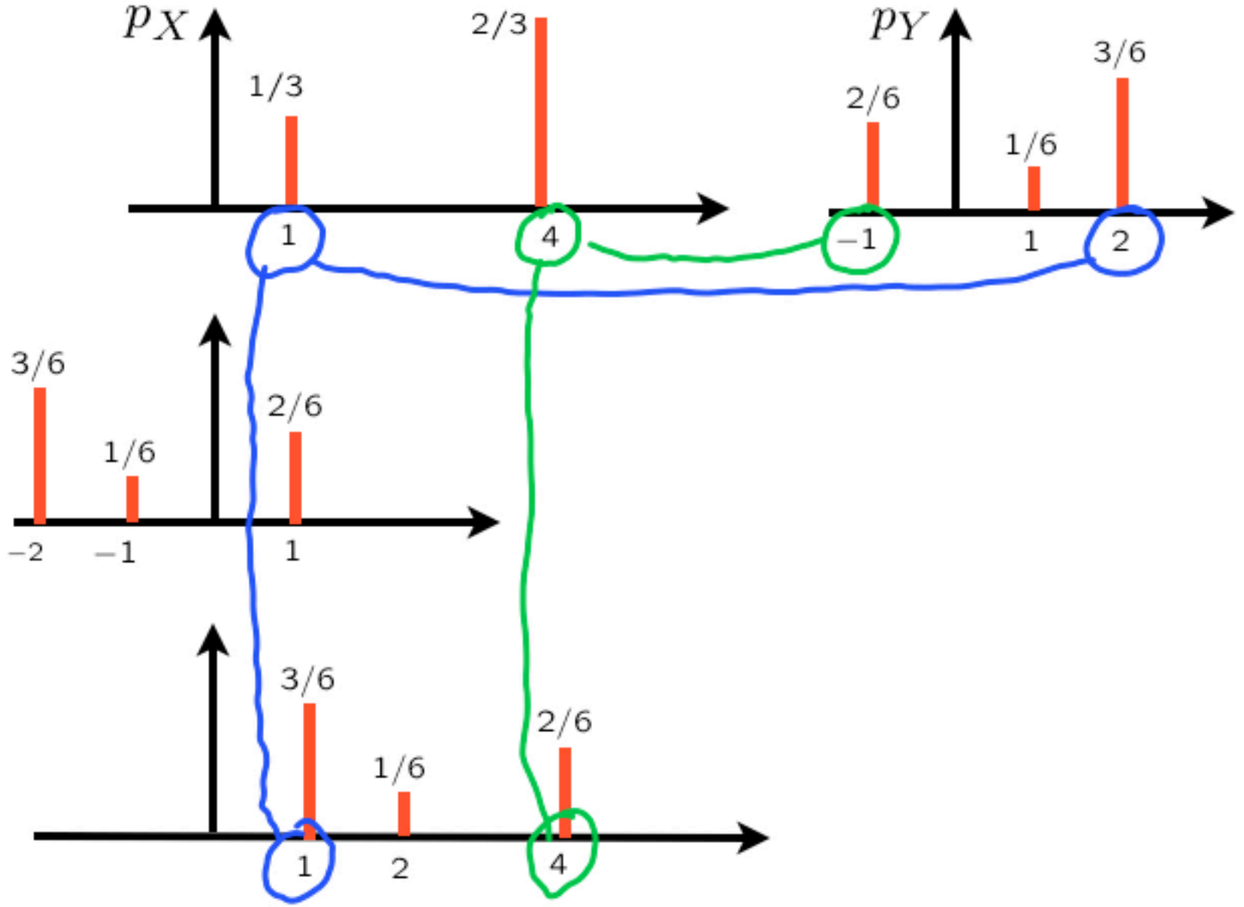


$$p_Z(z) = \sum_x P(X = x, Y = z - x)$$

$$= \sum_x p_X(x)\, p_Y(z - x)$$

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

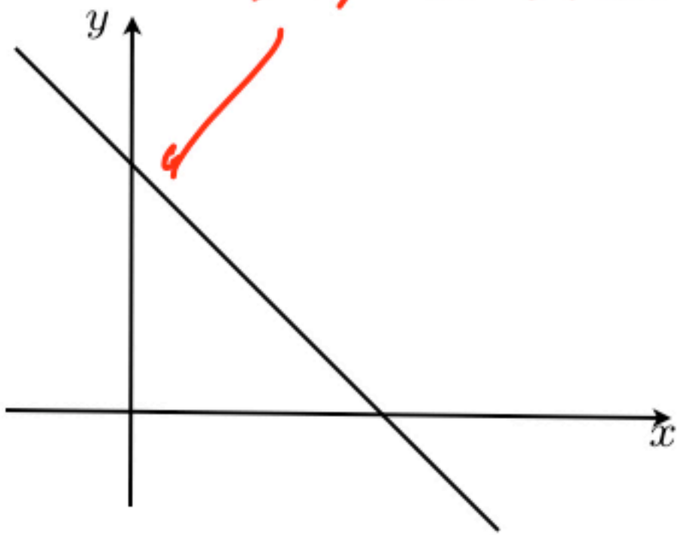
– Repeat for other values of z

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

$$p_Z(z) = \sum_x p_X(x)\, p_Y(z - x)$$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx$$

- $Z = X + Y$; X, Y independent, continuous known PDFs



Conditional on $X = x$: $Z = x + Y$ $x = 3$ $Z = Y + 3$

$f_{Z|X}(z|3) = f_{Y+3|X}(z|3) = f_{Y+3}(z) = f_Y(z-3)$

$f_{Z|X}(z|x) = f_Y(z - x)$ $f_{x+b}(x) = f_x(x-b)$

Joint PDF of Z and X:

$f_{X,Z}(x,z) = f_X(x)\, f_Y(z-x)$

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

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx$$

- $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$, independent $\qquad Z = X + Y$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma_x} e^{-(x-\mu_x)^2/2\sigma_x^2} \qquad f_Y(y) = \frac{1}{\sqrt{2\pi}\,\sigma_y} e^{-(y-\mu_y)^2/2\sigma_y^2}$$

$X + Y + W$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_x} \exp\left\{ -\frac{(x-\mu_x)^2}{2\sigma_x^2} \right\} \frac{1}{\sqrt{2\pi}\,\sigma_y} \exp\left\{ -\frac{(z-x-\mu_y)^2}{2\sigma_y^2} \right\} dx$$

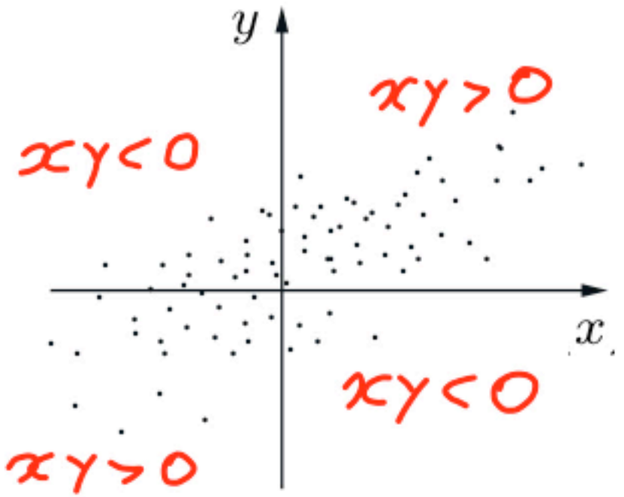
(algebra) $$= \frac{1}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2)}} \exp\left\{ -\frac{(z-\mu_x-\mu_y)^2}{2(\sigma_x^2+\sigma_y^2)} \right\}$$

$N(\mu_x+\mu_y, \sigma_x^2+\sigma_y^2)$

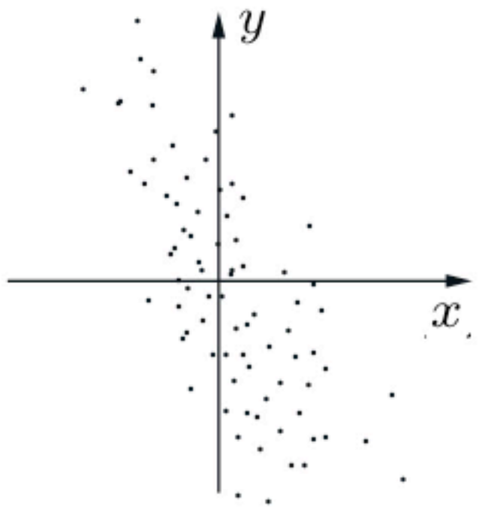
The sum of finitely many independent normals is normal

Covariance

- Zero-mean, discrete X and Y
  - if independent: $\mathbf{E}[XY] =$ $= E[X]\,E[Y] = 0$



$\mathbf{E}[XY] > 0$



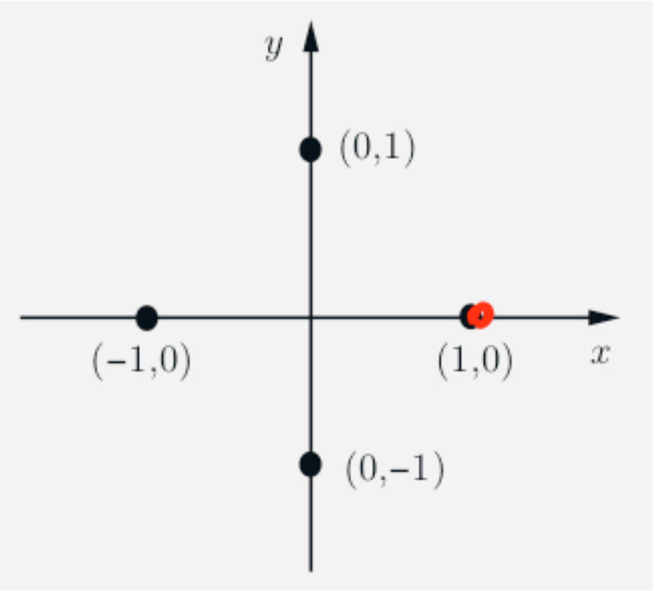
$\mathbf{E}[XY] < 0$

Definition for general case:

$$\operatorname{cov}(X,Y) = \mathbf{E}\Big[(X - \mathbf{E}[X]) \cdot (Y - \mathbf{E}[Y])\Big]$$

and $0 = E[(X - E[X])]\,E[Y - E[Y]]$

- independent $\Rightarrow$ $\operatorname{cov}(X,Y) = 0$
  (converse is not true)



$XY = 0$

$Cov = 0$

$X = 1 \Rightarrow Y = 0$

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)

$$= E[(aX+b)Y] = aE[XY] + bE[Y]$$

$$= a \cdot \text{cov}(X, Y)$$

$$\text{cov}(X, Y + Z) = E[X(Y+Z)]$$

$$= E[XY] + E[XZ] = \text{cov}(X, Y) + \text{cov}(X, Z)$$

$$\text{cov}(X, Y) = \mathbf{E}\Big[(X - \mathbf{E}[X]) \cdot (Y - \mathbf{E}[Y])\Big]$$

$$= E[XY] - E[XE[Y]]$$

$$- E[E[X]Y] + E[E[X]E[Y]]$$

$$= E[XY] - E[X]E[Y]$$

$$- E[X]E[Y] + E[X]E[Y]$$

$$\text{cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$$

The variance of a sum of random variables

$$\mathrm{var}(X_1 + X_2) = E\left[(X_1 + X_2 - E[X_1 + X_2])^2\right]$$

$$= E\left[\left((X_1 - E[X_1]) + (X_2 - E[X_2])\right)^2\right]$$

$$= E\Big[(X_1 - E[X_1])^2 + (X_2 - E[X_2])^2 + 2\,(X_1 - E[X_1])(X_2 - E[X_2])\Big]$$

$$= \mathrm{var}(X_1) + \mathrm{var}(X_2) + 2\,\mathrm{cov}(X_1, X_2).$$

The variance of a sum of random variables

$$\mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + 2\,\mathrm{cov}(X_1, X_2)$$

$$\mathrm{var}(X_1 + \cdots + X_n) =$$

(assume 0 means)

$$= E\left[(X_1 + \cdots + X_n)^2\right]$$

$$= E\left[\sum_{i=1}^{n} X_i^2 + \sum_{\substack{i=1,\ldots,n \\ j=1,\ldots,n \\ i \neq j}} X_i X_j\right] \quad \left.\begin{array}{l}\\\\\\\end{array}\right\} n^2 - n \text{ terms}$$

$$= \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$$

$$\mathrm{var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{var}(X_i) + \sum_{\{(i,j):\, i \neq j\}} \mathrm{cov}(X_i, X_j)$$

The Correlation coefficient

- Dimensionless version of covariance:

$$\rho(X,Y) = \mathbf{E}\left[\frac{(X-\mathbf{E}[X])}{\sigma_X}\cdot\frac{(Y-\mathbf{E}[Y])}{\sigma_Y}\right] = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}$$

$$-1 \le \rho \le 1$$

- Measure of the degree of "association" between X and Y
- Independent $\Rightarrow$ $\rho = 0$, "uncorrelated" (converse is not true)
- $\rho(X,X) = \dfrac{var(X)}{\sigma_x^2} = 1$
- $|\rho| = 1 \Leftrightarrow (X-\mathbf{E}[X]) = c(Y-\mathbf{E}[Y])$ (linearly related)
- $\mathrm{cov}(aX+b,Y) = a\cdot\mathrm{cov}(X,Y) \Rightarrow \rho(aX+b,Y) = \dfrac{a\,cov(X,Y)}{|a|\sigma_x\sigma_y} = sign(a)\cdot\rho(X,Y)$

Proof of key properties of the correlation coefficient

$$\rho(X,Y) = \mathbf{E}\left[\frac{(X-\mathbf{E}[X])}{\sigma_X}\cdot\frac{(Y-\mathbf{E}[Y])}{\sigma_Y}\right]$$

$$-1 \le \rho \le 1$$

- Assume, for simplicity, zero means and unit variances, so that $\rho(X,Y) = \mathbf{E}[XY]$

$$\mathbf{E}\left[(X-\rho Y)^2\right] = E[X^2] - 2\rho E[XY] + \rho^2 E[Y^2]$$

$$0 \le = 1 - 2\rho^2 + \rho^2 = 1-\rho^2 \qquad 1-\rho^2 \ge 0 \Rightarrow \rho^2 \le 1$$

If $|\rho| = 1$, then $X = \rho Y \Rightarrow X = Y$ or $X = -Y$

Interpreting the correlation coefficient

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- 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

    $X = Z + V \qquad Y = Z + W$

    Assume, for simplicity, that Z, V, W have zero means, unit variances

$$\rho(X,Y) = \frac{1}{\sqrt{2}\cdot\sqrt{2}} = \frac{1}{2}$$

$$\text{var}(X) = \text{var}(Z) + \text{var}(V) = 2 \Rightarrow \sigma_X = \sqrt{2} \qquad \sigma_Y = \sqrt{2}$$

$$\text{cov}(X,Y) = E[(Z+V)(Z+W)] = E[Z^2] + E[VZ] + E[ZW] + E[VW]$$

$$= 1 + 0 + 0 + 0$$

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).

$$\text{var}(X_1 + \cdots + X_{10}) = \sum_{i=1}^{10} \text{var}(X_i) + \sum_{\{(i,j):\, i \neq j\}} \text{cov}(X_i, X_j)$$

$E[X_1 + \cdots + X_{10}] = 10$

- If the X_i are uncorrelated, then:

$\text{var}(X_1 + \cdots + X_{10}) = 10 \cdot (1.3)^2 = 16.9$ $\sigma(X_1 + \cdots + X_{10}) = 4.1$

- If for $i \neq j$, $\rho(X_i, X_j) = 0.9$:

$\text{cov}(X_i, X_j) = \rho\, \sigma_{X_i} \sigma_{X_j} = 0.9 \times 1.3 \times 1.3 = 1.052$

$\text{var}(X_1 + \cdots + X_{10}) = 10 \cdot (1.3)^2 + 90 \cdot 1.052 = 154$

$\sigma(X_1 + \cdots + X_{10}) = 12.4$

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 $h(X)$?
- $h(X)$ is the r.v. that takes the value x^2, if X happens to take the value x

- $g(y) = \mathbf{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y)$
  (integral in continuous case)

- $g(Y)$: is the r.v. that takes the value $\mathbf{E}[X \mid Y = y]$, 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: $\mathbf{E}[X|Y] = g(Y)$

The mean of $\mathbf{E}[X \mid Y]$: Law of iterated expectations

- $g(y) = \mathbf{E}[X \mid Y = y]$

$$\mathbf{E}\Big[\mathbf{E}[X \mid Y]\Big] = \mathbf{E}[X]$$

$E[X|Y] \triangleq g(Y)$

$\mathbf{E}\Big[\mathbf{E}[X \mid Y]\Big] = E[g(Y)]$

$= \sum_y g(y)\, p_Y(y)$ exp. value rule

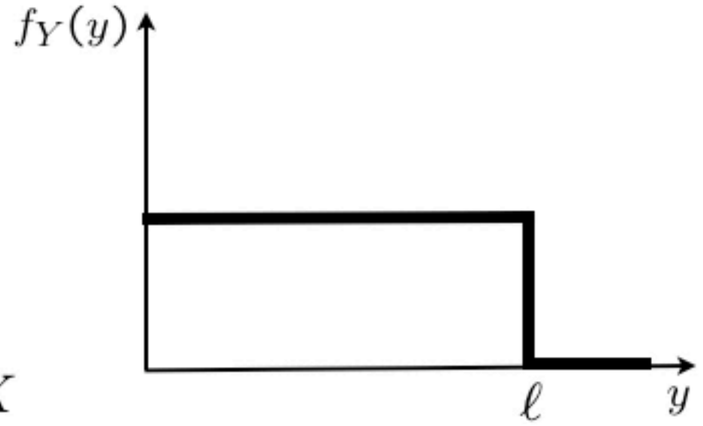
$= \sum_y E[X|Y=y]\, p_Y(y)$ • total exp thm

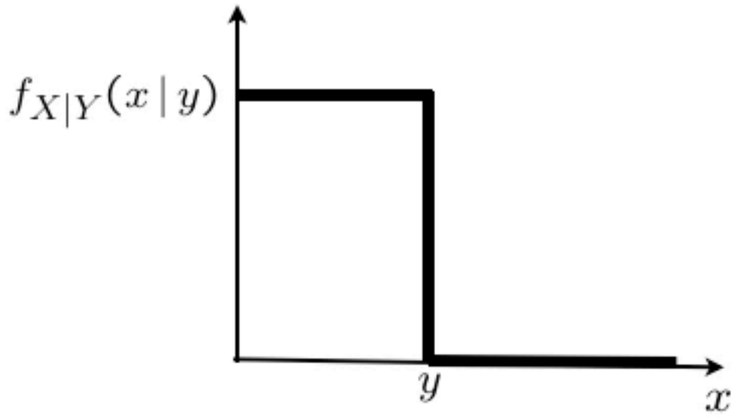
$= E[X]$

Stick-breaking example

- Stick example: stick of length ℓ
  break at uniformly chosen point Y

  break what is left at uniformly chosen point X





- $\mathbf{E}[X \mid Y = y] =$ $Y/2$
- $\mathbf{E}[X \mid Y] =$ $Y/2$

$\mathbf{E}[X] =$ $E[E[X|Y]] = E[Y/2] = \frac{1}{2}E[Y] = \frac{1}{2} \cdot \frac{\ell}{2} = \frac{\ell}{4}$

Forecast revisions

$$\mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \mathbf{E}[X]$$

- 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: $E[x \mid Y=y]$ $E[x|Y]$

- Law of iterated expectations:

$E[\text{revised forecast}] = E[x] = \text{original forecast}$

The conditional variance as a random variable

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

$$\text{var}(X \mid Y = y) = \mathbf{E}\Big[(X - \mathbf{E}[X \mid Y = y])^2 \mid Y = y\Big]$$

$\text{var}(X \mid Y)$ is the r.v. that takes the value $\text{var}(X \mid Y = y)$, when $Y = y$

- Example: X uniform on $[0, Y]$

$$\text{var}(X \mid Y = y) = y^2/12$$

$$\text{var}(X \mid Y) = Y^2/12$$

Law of total variance: $\text{var}(X) = \mathbf{E}\Big[\text{var}(X \mid Y)\Big] + \text{var}\Big(\mathbf{E}[X \mid Y]\Big)$

Derivation of the law of total variance

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \mid Y)\big] + \mathrm{var}\big(\mathbf{E}[X \mid Y]\big)$$

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

$\mathrm{var}(X \mid Y = y) = E[X^2 \mid Y=y] - (E[X \mid Y=y])^2$ for all y

$\mathrm{var}(X \mid Y) = E[X^2 \mid Y] - (E[X \mid Y])^2$

$\mathbf{E}\big[\mathrm{var}(X \mid Y)\big] = E[X^2] - E[(E[X \mid Y])^2]$

+

$\mathrm{var}\big(\mathbf{E}[X \mid Y]\big) = E[(E[X \mid Y])^2] - (E[E[X \mid Y]])^2$

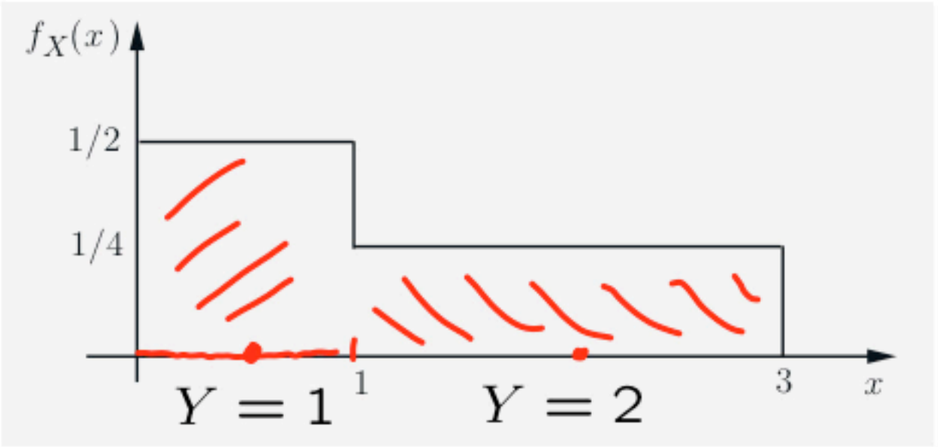
$(E[X])^2$

A simple example

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \mid Y)\big] + \mathrm{var}\big(\mathbf{E}[X \mid Y]\big)$$

$= 37/48$

$= 5/24 \quad + 9/16$



$$\mathrm{var}(X \mid Y) = \begin{cases} 1/2 & \mathrm{var}(X \mid Y = 1) = 1/12 \\ 1/2 & \mathrm{var}(X \mid Y = 2) = 2^2/12 = 4/12 \end{cases}$$

$$\mathbf{E}\big[\mathrm{var}(X \mid Y)\big] = \frac{1}{2} \cdot \frac{1}{12} + \frac{1}{2} \cdot \frac{4}{12} = \frac{5}{24}$$

$$\mathbf{E}[X \mid Y] = \begin{cases} 1/2 & \mathbf{E}[X \mid Y = 1] = \frac{1}{2} \\ 1/2 & \mathbf{E}[X \mid Y = 2] = 2 \end{cases}$$

$$\mathrm{var}\big(\mathbf{E}[X \mid Y]\big) = \frac{1}{2}\left(\frac{1}{2} - \frac{5}{4}\right)^2 + \frac{1}{2}\left(2 - \frac{5}{4}\right)^2 = \frac{9}{16}$$

$$\mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot 2 = \frac{5}{4} = E[X]$$

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: $y = 1: \ \frac{1}{10}\sum_{i=1}^{10} x_i = 90 \qquad y = 2: \ \frac{1}{20}\sum_{i=11}^{30} x_i = 60$

- $\mathbf{E}[X] = \frac{1}{30}\sum_{i=1}^{30} x_i = \frac{1}{30}(90 \cdot 10 + 60 \cdot 20) = 70$

$\mathbf{E}[X \mid Y = 1] = 90$

$\mathbf{E}[X \mid Y = 2] = 60$

$\mathbf{E}[X \mid Y] =$ 1/3 → 90; 2/3 → 60

- $\mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \frac{1}{3} \cdot 90 + \frac{2}{3} \cdot 60 = 70$

Section means and variances (ctd.)

$$\mathbf{E}[X \mid Y] = \begin{cases} 90, & \text{w.p. } 1/3 \\ 60, & \text{w.p. } 2/3 \end{cases}$$

$$\mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = 70 = \mathbf{E}[X]$$

$$\text{var}\big(\mathbf{E}[X \mid Y]\big) = \tfrac{1}{3}(90-70)^2 + \tfrac{2}{3}(60-70)^2 = 200$$

- More data: $\frac{1}{10}\sum_{i=1}^{10}(x_i - 90)^2 = 10$ $\quad$ $\frac{1}{20}\sum_{i=11}^{30}(x_i - 60)^2 = 20$

$\text{var}(X \mid Y = 1) =$ 10

$\text{var}(X \mid Y = 2) =$ 20

$\text{var}(X \mid Y) =$ 1/3 → 10; 2/3 → 20

$$\mathbf{E}\big[\text{var}(X \mid Y)\big] = \tfrac{1}{3}\cdot 10 + \tfrac{2}{3}\cdot 20 = \tfrac{50}{3}$$

$$\text{var}(X) = \mathbf{E}\big[\text{var}(X \mid Y)\big] + \text{var}\big(\mathbf{E}[X \mid Y]\big) = 50/3 + 200$$

var(X) = (average variability **within** sections) + (variability **between** sections)

Sum of a random number of independent r.v.'s

$$\mathbf{E}[Y] = \mathbf{E}[N] \cdot \mathbf{E}[X]$$

- N: number of stores visited
  (N is a nonnegative integer r.v.)
- Let $Y = X_1 + \cdots + X_N$
- X_i: money spent in store i
  - X_i independent, identically distributed
  - independent of N

$$\mathbf{E}[Y \mid N = n] = E[X_1 + \cdots + X_N \mid N = n] = E[X_1 + \cdots + X_n \mid N = n]$$
$$= E[X_1 + \cdots + X_n] = n\,E[X]$$

$E[Y|N] = NE[X]$

- Total expectation theorem:

$$\mathbf{E}[Y] = \sum_n p_N(n)\,\mathbf{E}[Y \mid N = n] = \sum_n p_N(n)\,n\,E[X] = E[N]E[X]$$

- Law of iterated expectations:

$$\mathbf{E}[Y] = \mathbf{E}\big[\mathbf{E}[Y \mid N]\big] = E[N\,E[X]] = E[N]\,E[X]$$

Variance of sum of a random number of independent r.v.'s

$Y = X_1 + \cdots + X_N$

- $\mathbf{E}[Y \mid N] = N\,\mathbf{E}[X]$
- $\mathrm{var}\big(\mathbf{E}[Y \mid N]\big) = \mathrm{var}(N\,E[X]) = (E[X])^2\,\mathrm{var}(N)$
- $\mathrm{var}(Y \mid N = n) = \mathrm{var}(X_1 + \cdots + X_n \mid N = n) = \mathrm{var}(X_1 + \cdots + X_n) = n\,\mathrm{var}(X)$

  $\mathrm{var}(Y \mid N) = N\,\mathrm{var}(X)$
- $\mathbf{E}\big[\mathrm{var}(Y \mid N)\big] = E[N\,\mathrm{var}(X)] = E[N]\,\mathrm{var}(X)$

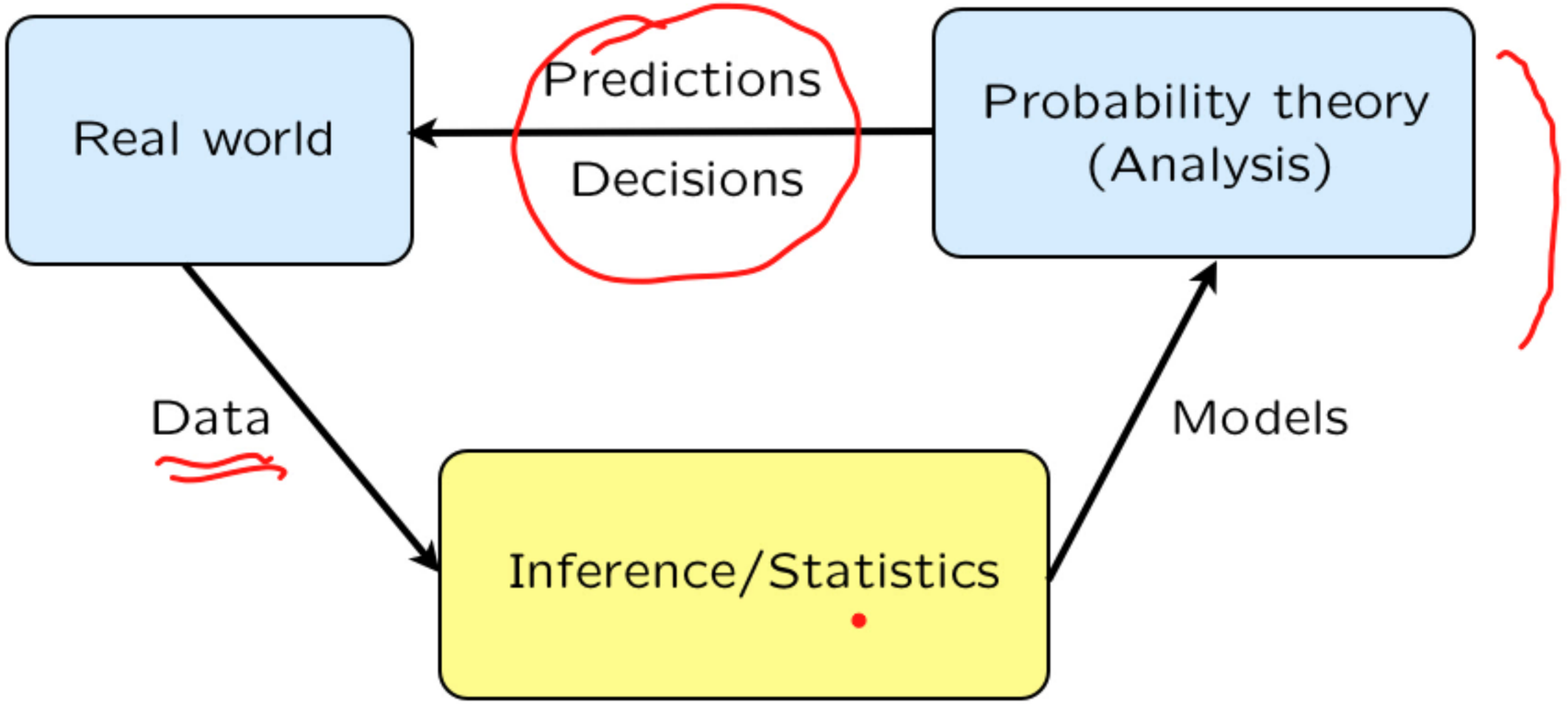
$$\mathrm{var}(Y) = \mathbf{E}\big[\mathrm{var}(Y \mid N)\big] + \mathrm{var}\big(\mathbf{E}[Y \mid N]\big)$$

$$\mathrm{var}(Y) = \mathbf{E}[N]\,\mathrm{var}(X) + (\mathbf{E}[X])^2\,\mathrm{var}(N)$$

LECTURE 14: Introduction to Bayesian inference

- The big picture
  - motivation, applications
  - problem types (hypothesis testing, estimation, etc.)
- The general framework
  - Bayes' rule $\longrightarrow$ posterior
    (4 versions)
  - point estimates (MAP, LMS)
  - performance measures)
    (prob. of error; mean squared error)
  - examples

Inference: the big picture



Inference then and now

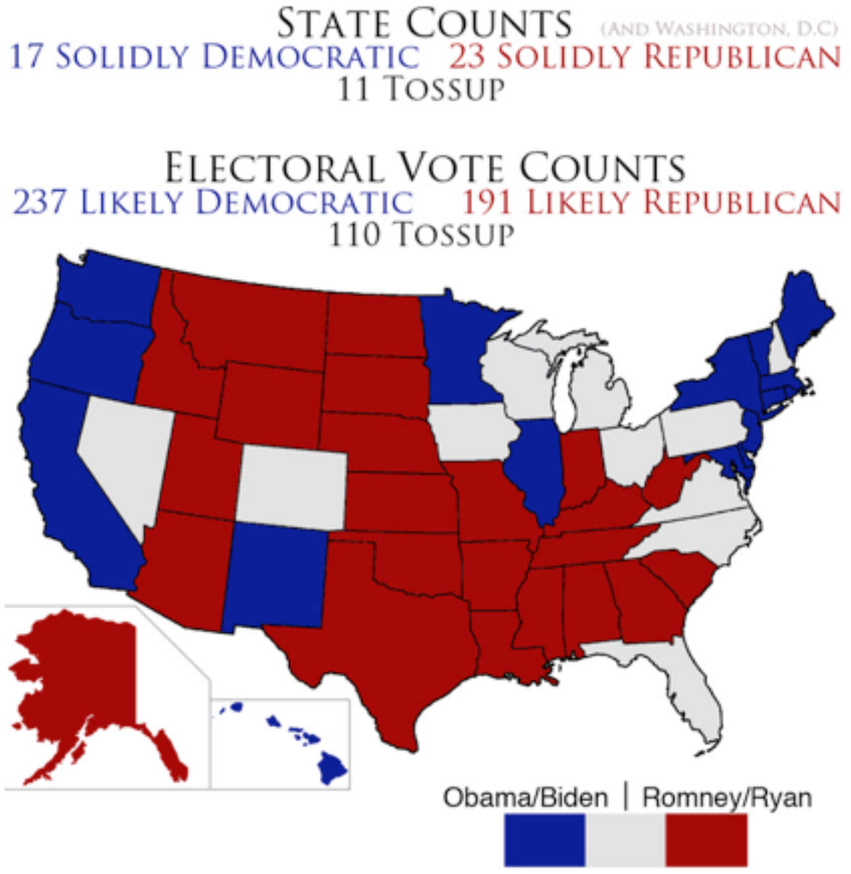
- Then:
  10 patients were treated: 3 died
  10 patients were not treated: 5 died
  Therefore ...

Now:

- Big data
- Big models
- Big computers

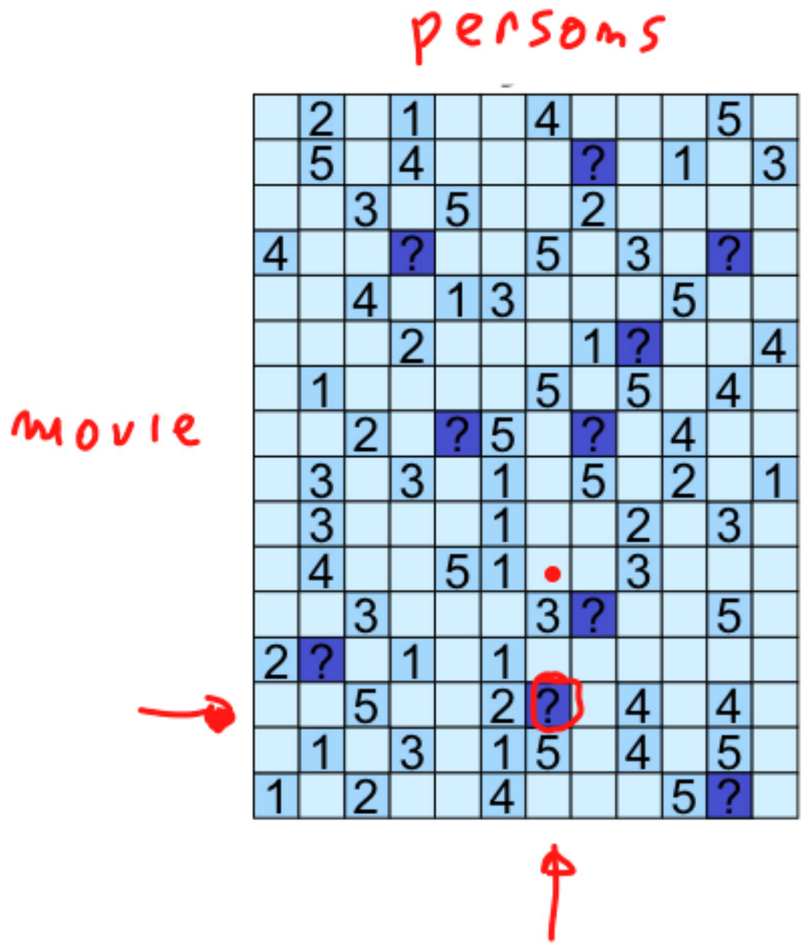
A sample of application domains

- Design and interpretation of experiments
  - polling



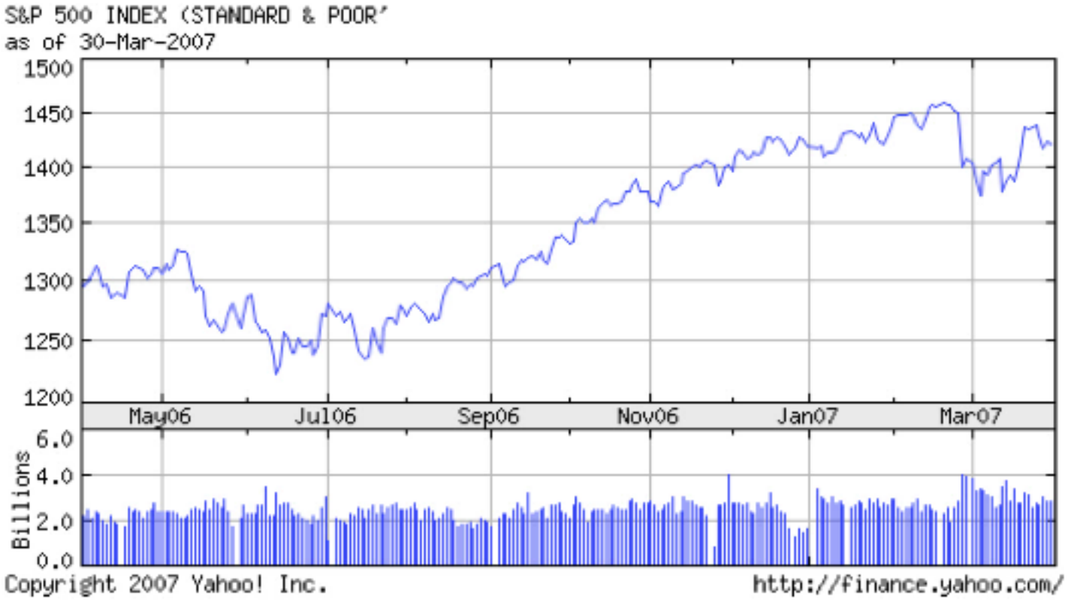
A sample of application domains

- marketing, advertising
- recommendation systems
  - Netflix competition



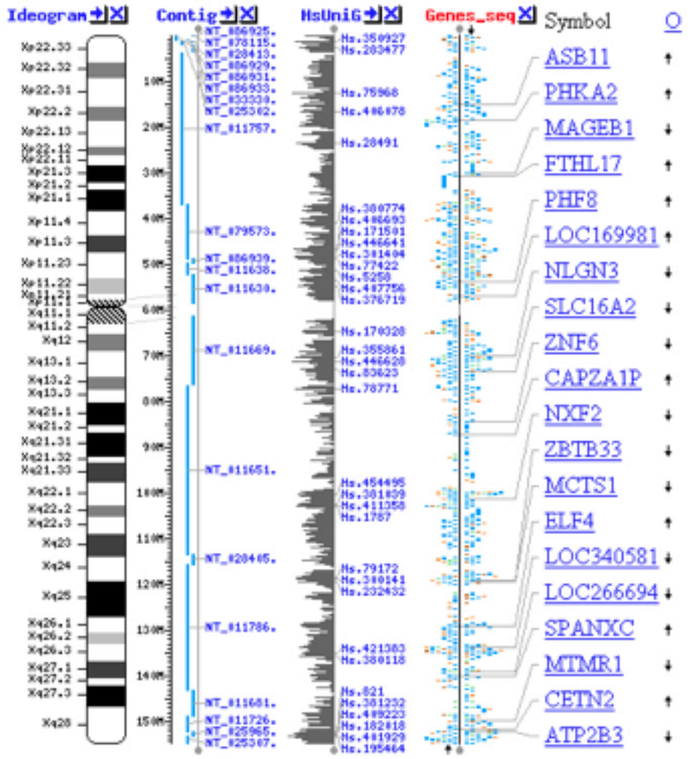
A sample of application domains

- Finance



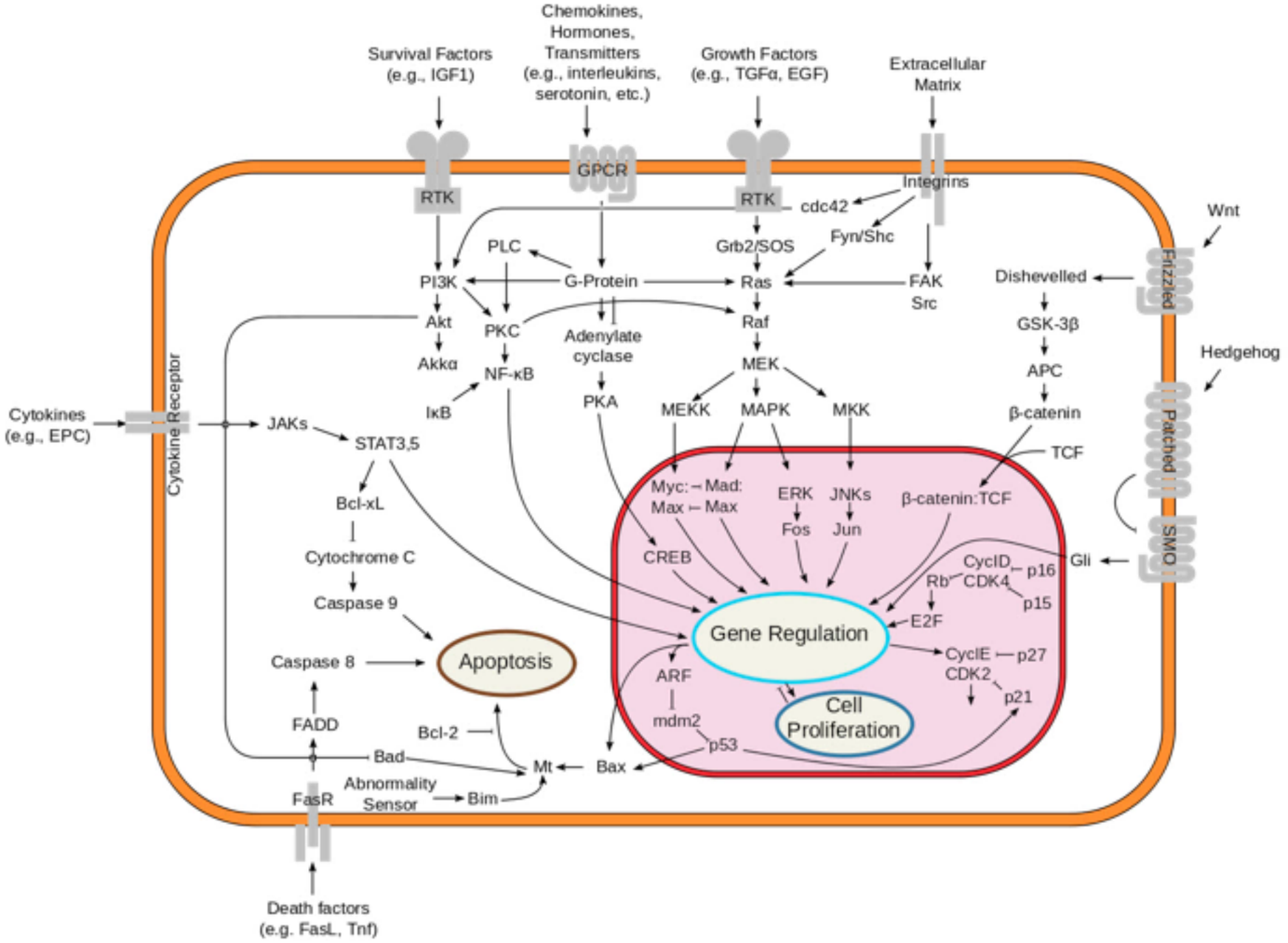
A sample of application domains

- Life sciences
  - genomics



  - neuroscience, etc., etc.

  - systems biology



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

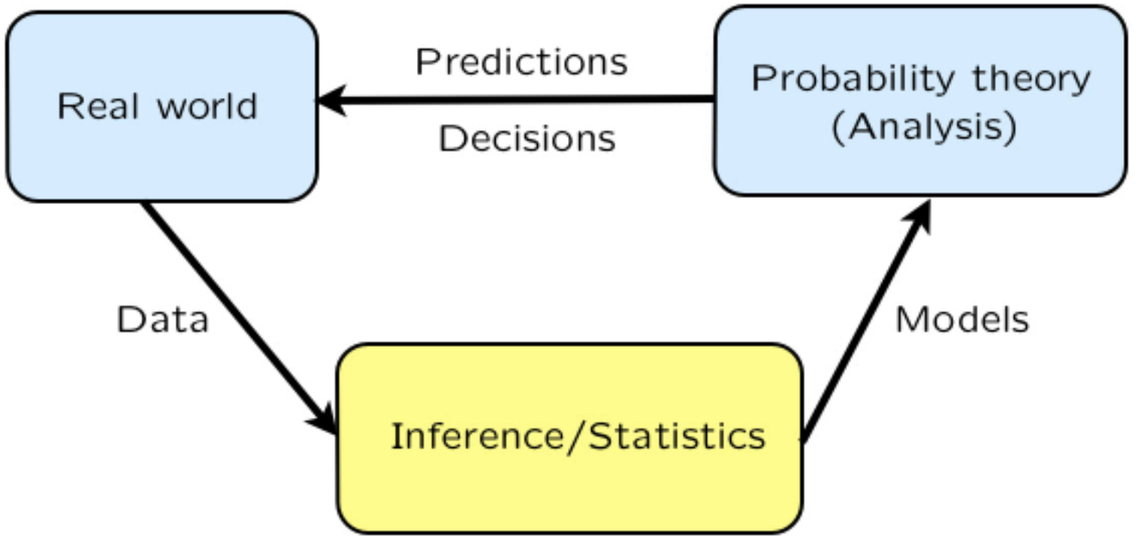
- Signal processing
  - communication systems (noisy ...)
  - speech processing and understanding
  - image processing and understanding
  - tracking of objects
  - positioning systems (e.g., GPS)
  - detection of abnormal events

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

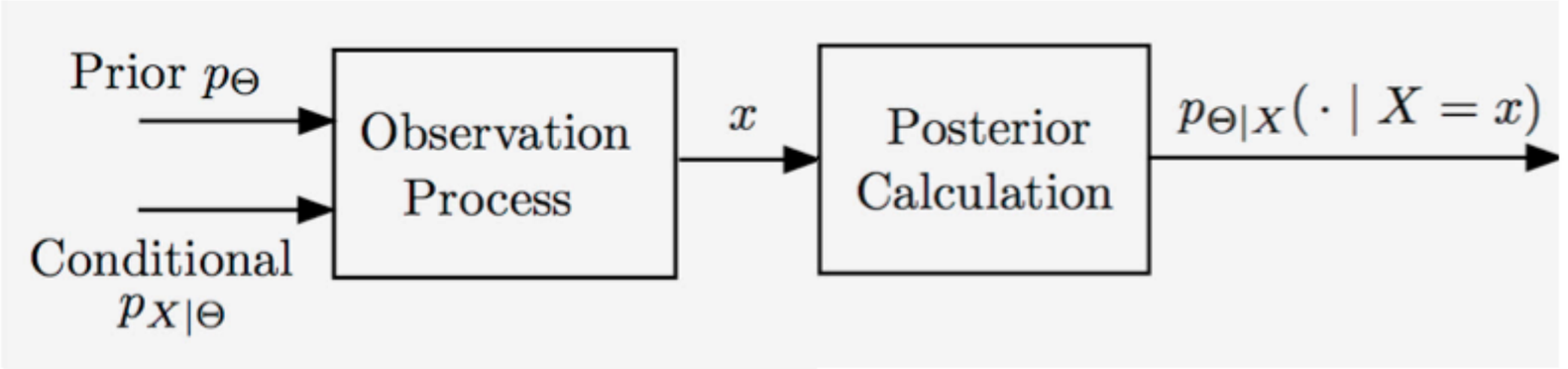
  Is it an airplane or a bird?

- 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

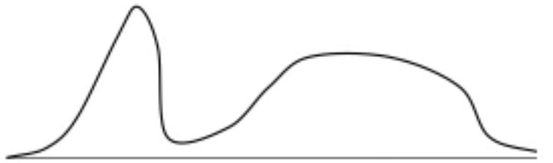


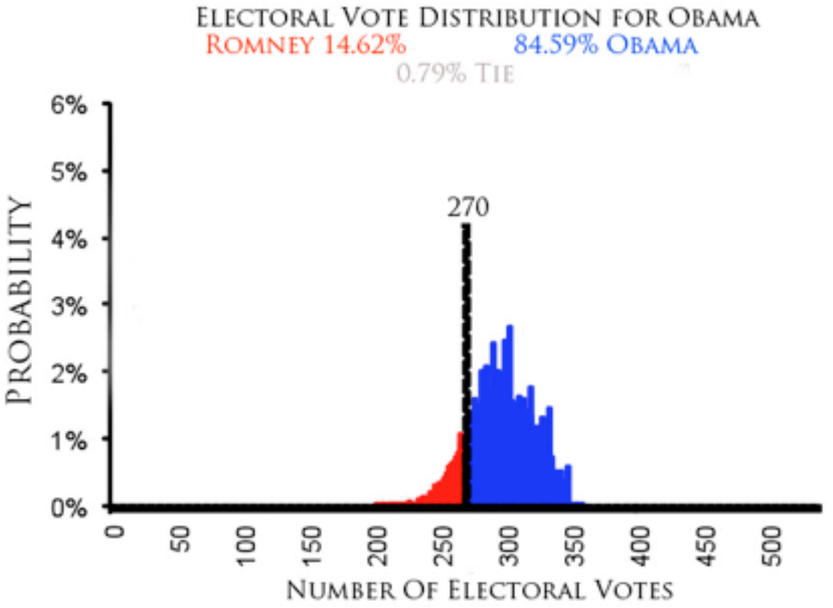
The output of Bayesian inference

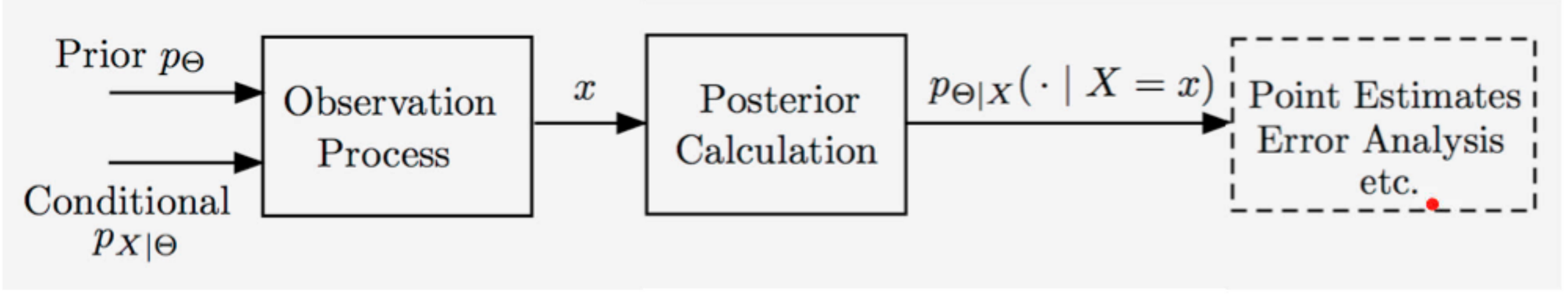
The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot \mid x)$ or PDF $f_{\Theta|X}(\cdot \mid x)$

$p_{\Theta|X}(\cdot \mid x)$

$f_{\Theta|X}(\cdot \mid x)$

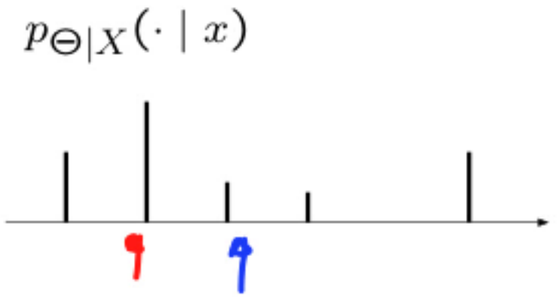


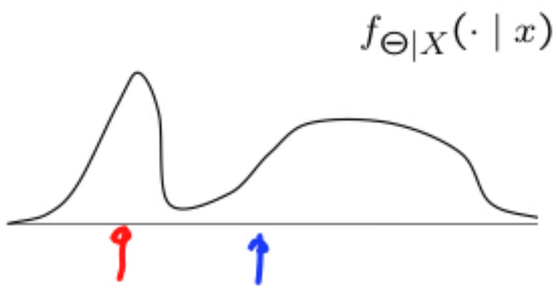




Point estimates in Bayesian inference

The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot \mid x)$ or PDF $f_{\Theta|X}(\cdot \mid x)$





estimate: $\widehat{\theta} = g(x)$
(number)

estimator: $\widehat{\Theta} = g(X)$
(random variable)

- Maximum a posteriori probability (MAP):

$$p_{\Theta|X}(\theta^* \mid x) = \max_{\theta} p_{\Theta|X}(\theta \mid x)$$

$$f_{\Theta|X}(\theta^* \mid x) = \max_{\theta} f_{\Theta|X}(\theta \mid x)$$

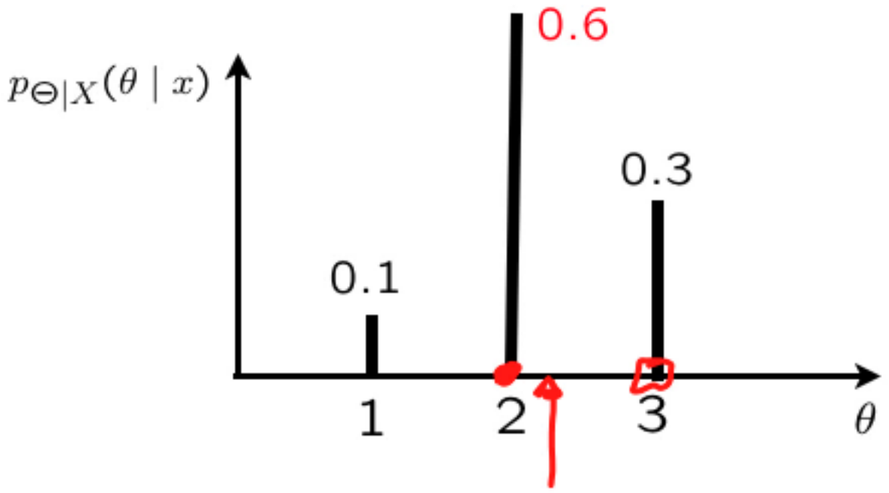
- Conditional expectation: $\mathbf{E}[\Theta \mid X = x]$ (LMS: Least Mean Squares)

Discrete Θ, discrete X

- values of Θ: alternative hypotheses

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, p_{X|\Theta}(x \mid \theta)}{p_X(x)}$$

$$p_X(x) = \sum_{\theta'} p_\Theta(\theta')\, p_{X|\Theta}(x \mid \theta')$$



- MAP rule: $\hat{\theta} = 2$

LMS: $\hat{\Theta} = E[\Theta \mid X = x] = 2.2$

- conditional prob of error:

$\mathbf{P}(\hat{\theta} \neq \Theta \mid X = x) = 0.4$

smallest under the MAP rule

- overall probability of error:

$$\mathbf{P}(\widehat{\Theta} \neq \Theta) = \sum_x \mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x)\, p_X(x)$$

$$= \sum_\theta \mathbf{P}(\widehat{\Theta} \neq \Theta \mid \Theta = \theta)\, p_\Theta(\theta)$$

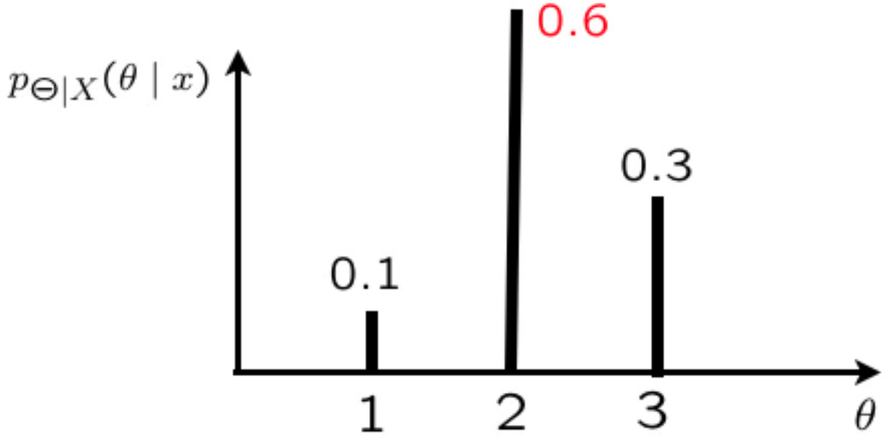
Discrete Θ, continuous X

- Standard example:
  - send signal $\Theta \in \{1, 2, 3\}$

    $X = \Theta + W$

    $W \sim N(0, \sigma^2)$, indep. of Θ

    $f_{X|\Theta}(x \mid \theta) = f_W(x - \theta)$



- MAP rule: $\hat{\theta} =$ 2

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \sum_{\theta'} p_\Theta(\theta') f_{X|\Theta}(x \mid \theta')$$

- conditional prob of error:

  $\mathbf{P}(\hat{\theta} \neq \Theta \mid X = x)$

  smallest under the MAP rule

- overall probability of error:

$$\mathbf{P}(\widehat{\Theta} \neq \Theta) = \int \mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x) f_X(x)\, dx$$

$$= \sum_{\theta} \mathbf{P}(\widehat{\Theta} \neq \theta \mid \Theta = \theta)\, p_\Theta(\theta)$$

Continuous Θ, continuous X

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta') f_{X|\Theta}(x \mid \theta')\, d\theta'$$

- 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$[0, \Theta]$

  Θ: uniform $[0, 1]$

- $\widehat{\Theta} = g(X)$ MAP LMS
- interested in:

$$\mathbf{E}\Big[(\widehat{\Theta} - \Theta)^2 \mid X = x\Big]$$

$$\mathbf{E}\Big[(\widehat{\Theta} - \Theta)^2\Big]$$

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 $[0,1]$

$$f_{\Theta|K}(\theta \mid k) = \frac{f_\Theta(\theta)\, p_{K|\Theta}(k \mid \theta)}{p_K(k)}$$

$$p_K(k) = \int f_\Theta(\theta') p_{K|\Theta}(k \mid \theta')\, d\theta'$$

$$f_{\Theta|K}(\theta \mid k) = \frac{1 \cdot \binom{n}{k} \theta^k (1-\theta)^{n-k}}{p_K(k)}$$

$\theta \in [0,1]$

$$= \frac{1}{d(n,k)} \theta^k (1-\theta)^{n-k}$$

"Beta distribution, with parameters $(k+1, n-k+1)$"

- If prior is Beta: $f_\Theta(\theta) = \frac{1}{c}\theta^\alpha (1-\theta)^\beta$ $\alpha, \beta \geq 0$

$$f_{\Theta|K}(\theta \mid k) = \frac{1}{c} \theta^\alpha (1-\theta)^\beta \binom{n}{k} \theta^k (1-\theta)^{n-k} / p_K(k) = d\, \theta^{\alpha+k} (1-\theta)^{\beta+n-k}$$

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 $[0,1]$

$$f_{\Theta|K}(\theta \mid k) = \frac{1}{d(n,k)}\,\theta^k(1-\theta)^{n-k}$$

- MAP estimate:

$$\widehat{\theta}_{\text{MAP}} = k/n$$

$$\max_\theta \left[k \log\theta + (n-k)\log(1-\theta)\right]$$

$$k/\theta - (n-k)/(1-\theta) = 0$$

$$\widehat{\Theta}_{\text{MAP}} = K/n$$

$$\int_0^1 \theta^\alpha (1-\theta)^\beta\, d\theta = \frac{\alpha!\,\beta!}{(\alpha+\beta+1)!} \qquad \alpha \geq 0,\ \beta \geq 0$$

$$\mathbf{E}[\Theta \mid K=k] = \int_0^1 \theta\, f_{\theta|k}(\theta|k)\, d\theta$$

$$= \frac{1}{d(n,k)} \int_0^1 \theta^{k+1}(1-\theta)^{n-k}\, d\theta$$

$$= \frac{1}{\dfrac{k!\,(n-k)!}{(n+1)!}} \cdot \frac{(k+1)!\,(n-k)!}{(n+2)!}$$

$$= \frac{k+1}{n+2} \approx \frac{k}{n} \quad (n \text{ large})$$

Summary

- Problem data: $p_\Theta(\cdot)$, $p_{X|\Theta}(\cdot \mid \cdot)$
- Given the value x of X: find, e.g., $p_{\Theta|X}(\cdot \mid x)$
  - using appropriate version of the Bayes rule (4 choices)

- Estimator $\widehat{\Theta} = g(X)$ Estimate $\hat{\theta} = g(x)$
  - MAP: $\hat{\theta}_{\text{MAP}} = g_{\text{MAP}}(x)$ maximizes $p_{\Theta|X}(\theta \mid x)$
  - LMS: $\hat{\theta}_{\text{LMS}} = g_{\text{LMS}}(x) = \mathbf{E}[\Theta \mid X = x]$

- Performance evaluation of an estimator $\widehat{\Theta}$
  - $\mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x)$
  - $\mathbf{P}(\widehat{\Theta} \neq \Theta)$
  - $\mathbf{E}\big[(\widehat{\Theta} - \Theta)^2 \mid X = x\big]$
  - $\mathbf{E}\big[(\widehat{\Theta} - \Theta)^2\big]$

total prob. thm.
exp

LECTURE 15: Linear models with normal noise

$$X_i = \sum_{j=1}^{m} a_{ij}\Theta_j + W_i \qquad W_i,\ \Theta_j\text{: independent, normal}$$

- 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

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

$2\alpha x + \beta = 0$

$c \cdot e^{-8(x-3)^2}$ $\qquad \mu = 3 \qquad \frac{1}{2\sigma^2} = 8 \Rightarrow \sigma^2 = \frac{1}{16} \qquad c = \frac{1}{\frac{1}{4}\sqrt{2\pi}}$

$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)}$ $\qquad \alpha > 0$ $\qquad$ Normal with mean $-\beta/2\alpha$ and variance $1/2\alpha$

$$\alpha x^2 + \beta x + \gamma = \alpha\left(x^2 + \frac{\beta}{\alpha}x + \frac{\gamma}{\alpha}\right) = \alpha\left(\left(x + \frac{\beta}{2\alpha}\right)^2 - \frac{\beta^2}{4\alpha^2} + \frac{\gamma}{\alpha}\right)$$

$$f_x(x) = \underbrace{c}_{\sqcup}\, e^{-\alpha\left(x + \frac{\beta}{2\alpha}\right)^2} \underbrace{e^{-\alpha\left(-\frac{\beta^2}{4\alpha^2} + \frac{\gamma}{\alpha}\right)}}_{\sqcup}$$

$\mu = -\frac{\beta}{2\alpha}$

$\frac{1}{2\sigma^2} = \alpha \Rightarrow \sigma^2 = 1/2\alpha$

Estimating a normal random variable in the presence of additive normal noise

$X = \Theta + W \qquad \Theta, W : N(0,1),$ independent

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

$f_{X|\Theta}(x \mid \theta)$: $X = \theta + W \qquad N(\theta, 1)$

$$f_{\Theta|X}(\theta \mid x) = \frac{1}{f_x(x)}\, c\, e^{-\frac{1}{2}\theta^2}\, c\, e^{-\frac{1}{2}(x-\theta)^2} = \underline{\underline{c(x)}}\, e^{-\text{quadratic}(\theta)}$$

Normal!

Fix x $\qquad \min_\theta \left[\frac{1}{2}\theta^2 + \frac{1}{2}(x-\theta)^2\right]$ $\qquad \theta + (\theta - x) = 0$

$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = x/2$

$\widehat{\Theta}_{\text{MAP}} = \mathbf{E}[\Theta \mid X] = X/2$

Estimating a normal parameter in the presence of additive normal noise

$$X = \Theta + W \qquad \Theta, W : N(0,1), \quad \text{independent}$$

$$\widehat{\Theta}_{\text{MAP}} = \widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X] = \frac{X}{2}$$

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

- Even with general means and variances:
  - posterior is normal
  - LMS and MAP estimators coincide
  - these estimators are "linear," of the form $\widehat{\Theta} = aX + b$

The case of multiple observations

$$X_1 = \Theta + W_1 \qquad \Theta \sim N(x_0, \sigma_0^2) \qquad W_i \sim N(0, \sigma_i^2)$$
$$\vdots$$
$$X_n = \Theta + W_n \qquad \Theta, W_1, \ldots, W_n \text{ independent}$$

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

$f_{X_i|\Theta}(x_i \mid \theta) =$ $c_i\, e^{-(x_i-\theta)^2/2\sigma_i^2}$

given $\Theta=\theta$: $X_i = \theta + W_i \sim N(\theta, \sigma_i^2)$

$f_{X|\Theta}(x \mid \theta) = f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n \mid \theta) = \prod_{i=1}^{n} f_{X_i|\Theta}(x_i \mid \theta)$

given $\Theta=\theta$: W_i independent $\Rightarrow$ X_i independent

$f_{\Theta|X}(\theta \mid x) = \frac{1}{f_X(x)} \cdot c_0\, e^{-(\theta-x_0)^2/2\sigma_0^2} \prod_{i=1}^{n} c_i\, e^{-(x_i-\theta)^2/2\sigma_i^2}$

Normal!

The case of multiple observations

$$f_{\Theta|X}(\theta\,|\,x) = c \cdot \exp\Big\{ -\,\text{quad}(\theta)\Big\} \qquad \text{quad}(\theta) = \frac{(\theta - x_0)^2}{2\sigma_0^2} + \frac{(\theta - x_1)^2}{2\sigma_1^2} + \cdots + \frac{(\theta - x_n)^2}{2\sigma_n^2}$$

find peak

$$\frac{d}{d\theta}\,\text{quad}(\theta) = 0: \quad \sum_{i=0}^{n} \frac{(\theta - x_i)}{\sigma_i^2} = 0 \quad \Rightarrow \quad \theta \sum_{i=0}^{n} \frac{1}{\sigma_i^2} = \sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}$$

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta\,|\,X = x] = \frac{\displaystyle\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\displaystyle\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

The case of multiple observations

- Key conclusions:
  - posterior is normal
  - LMS and MAP estimates coincide
  - these estimates are "linear," of the form $\hat{\theta} = a_0 + a_1 x_1 + \cdots + a_n x_n$
- Interpretations:
  - estimate $\hat{\theta}$: weighted average of x_0 (prior mean) and x_i (observations)
  - weights determined by variances

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = \frac{\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}.$$

σ_i^2 large

X_i very noisy

$\Rightarrow$ small weight

The mean squared error

$X_i = \Theta + W_i$

$$f_{\Theta|X}(\theta \,|\, x) = c \cdot \exp\Big\{ - \text{quad}(\theta)\Big\}$$

$$\text{quad}(\theta) = \frac{(\theta - x_0)^2}{2\sigma_0^2} + \frac{(\theta - x_1)^2}{2\sigma_1^2} + \cdots + \frac{(\theta - x_n)^2}{2\sigma_n^2}$$

$$\widehat{\theta} = \frac{\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

- Performance measures:

$$\mathbf{E}\Big[(\Theta - \widehat{\Theta})^2 \mid X = x\Big] = \mathbf{E}\Big[(\Theta - \hat{\theta})^2 \mid X = x\Big] = \text{var}(\Theta \mid X = x) = 1 \Big/ \sum_{i=0}^{n} \frac{1}{\sigma_i^2}$$

$$\mathbf{E}\Big[(\Theta - \widehat{\Theta})^2\Big] = \int E\Big[(\Theta - \hat{\Theta})^2 \,|\, X = x\Big] f_X(x)\, dx$$

$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)}$ $\quad \alpha > 0$ $\quad$ Normal with mean $-\beta/2\alpha$ and variance $1/2\alpha$

$$\alpha = \frac{1}{2\sigma_0^2} + \cdots + \frac{1}{2\sigma_n^2}$$

some σ_i^2 small $\rightarrow$ MSE small

all σ_i^2 large $\rightarrow$ MSE large

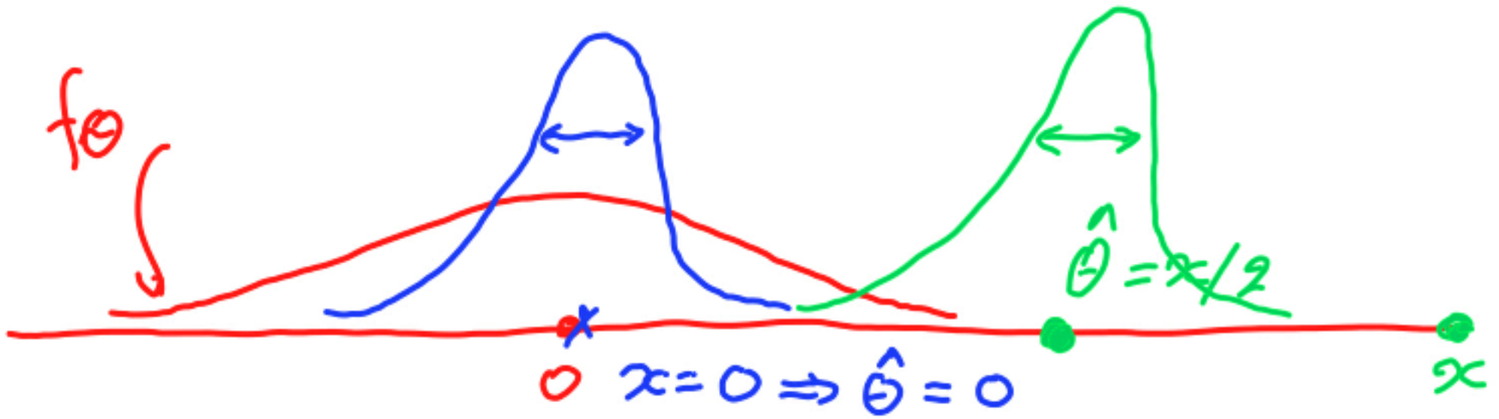
The mean squared error

$$\mathbf{E}\big[(\Theta - \widehat{\Theta})^2 \mid X = x\big] = \mathbf{E}\big[(\Theta - \widehat{\Theta})^2\big] = 1 \Big/ \sum_{i=0}^{n} \frac{1}{\sigma_i^2}$$

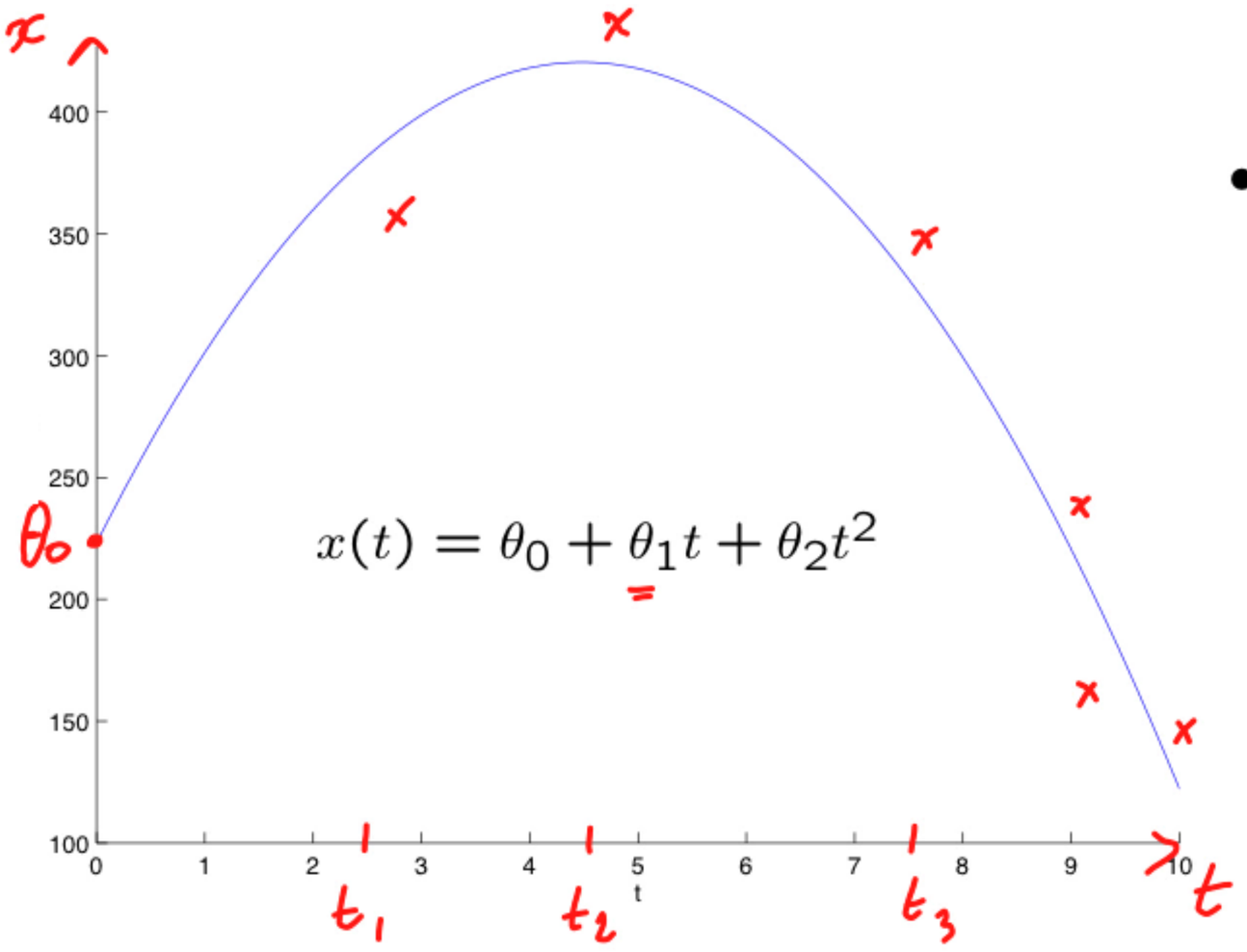
$$\hat{\theta} = \frac{\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

- Example: $\sigma_0^2 = \sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$ $\quad \frac{1}{(n+1)\frac{1}{\sigma^2}} = \frac{\sigma^2}{n+1}$
- conditional mean squared error same for all x
- Example: $X = \Theta + W$ $\quad \Theta \sim N(0,1)$, $\quad W \sim N(0,1)$

  independent Θ, W $\quad \widehat{\Theta} = X/2$ $\quad \mathbf{E}\big[(\Theta - \widehat{\Theta})^2 \mid X = x\big] = 1/2$



The case of multiple parameters: trajectory estimation



- Random variables $\Theta_0, \Theta_1, \Theta_2$

  independent; priors f_{Θ_j}

- Measurements at times $t_1, \ldots, t_n$

  $X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$

  noise model: f_{W_i}

  independent W_i; independent from Θ_j

A model with normality assumptions

$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i \qquad i = 1, \ldots, n$$

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

- 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: $N(\theta_0 + \theta_1 t_i + \theta_2 t_i^2, \sigma^2)$

$$f_{X_i|\Theta}(x_i \mid \theta) = c \cdot \exp\left\{ -(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2 / 2\sigma^2 \right\}$$

- posterior: $f_{\Theta|X}(\theta \mid x) = \frac{1}{f_X(x)} \prod_{j=0}^{2} f_{\Theta_j}(\theta_j) \prod_{i=1}^{n} f_{X_i|\Theta}(x_i \mid \theta)$

$$c(x) \exp\left\{ -\frac{1}{2}\left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2 \right\}$$

A model with normality assumptions

$$f_{\Theta|X}(\theta\,|\,x) = c(x)\exp\left\{-\frac{1}{2}\left(\frac{\theta_0^2}{\sigma_0^2}+\frac{\theta_1^2}{\sigma_1^2}+\frac{\theta_2^2}{\sigma_2^2}\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0-\theta_1 t_i-\theta_2 t_i^2)^2\right\}$$

- MAP estimate: maximize over $(\theta_0,\theta_1,\theta_2)$;
  (minimize quadratic function)

$\frac{\partial}{\partial\theta_j}(\text{quad}(\theta)) = 0$ 3 equations, 3 unknowns.
↑ linear

Linear normal models

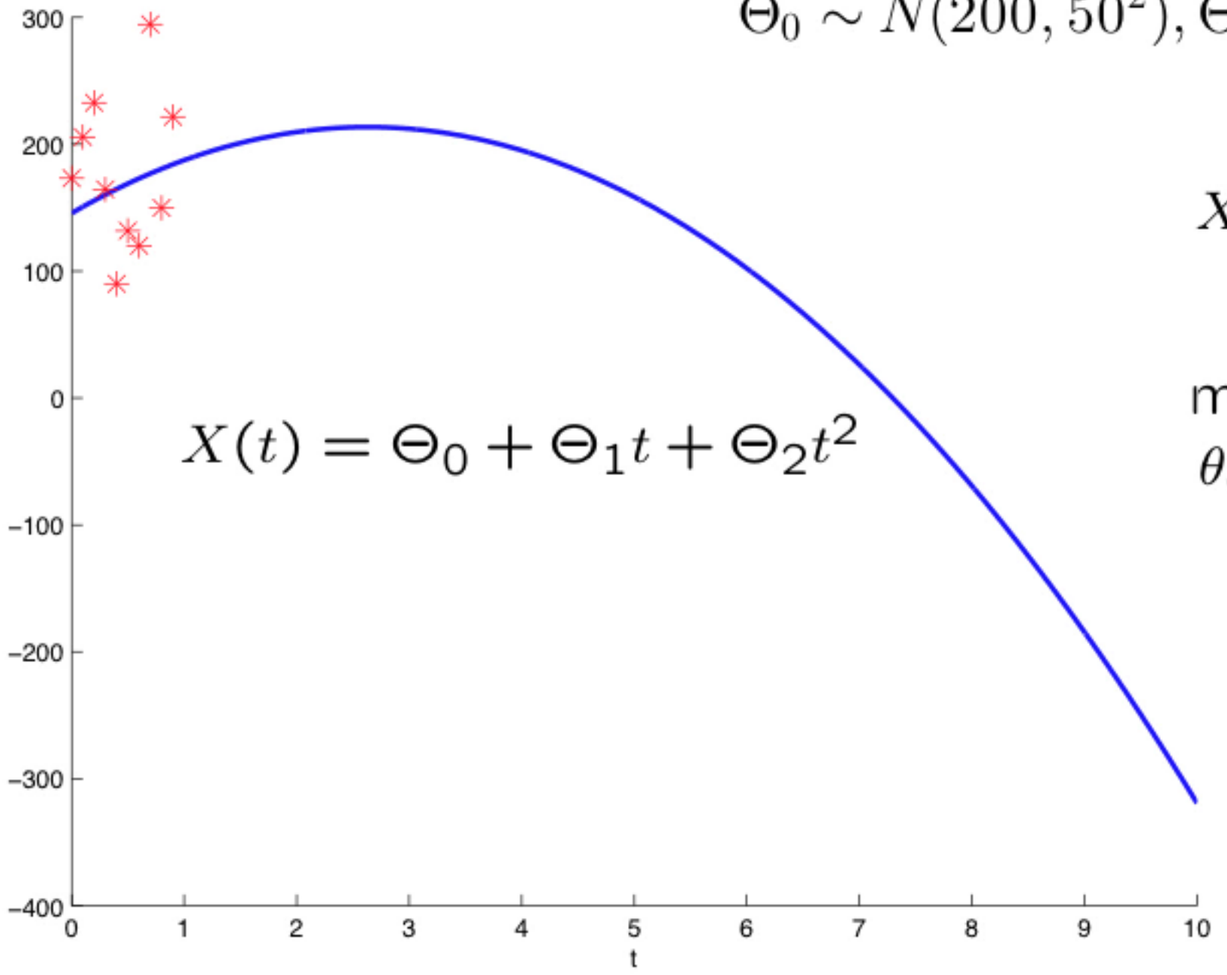
- Θ_j and X_i and are linear functions of independent normal random variables
- $f_{\Theta|X}(\theta \,|\, x) = c(x) \exp\Big\{ -\text{quadratic}(\theta_1, \ldots, \theta_m) \Big\}$ linear regression
- MAP estimate: maximize over $(\theta_1, \ldots \theta_m)$; (minimize quadratic function) linear equations

  $\widehat{\Theta}_{\text{MAP},j}$: linear function of $X = (X_1, \ldots, X_n)$
- 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
  - $\mathbf{E}\Big[(\widehat{\Theta}_{i,\text{MAP}} - \Theta_i)^2 \mid X = x\Big]$: same for all x

An illustration

Estimating the trajectory of a free-falling object

$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$



$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

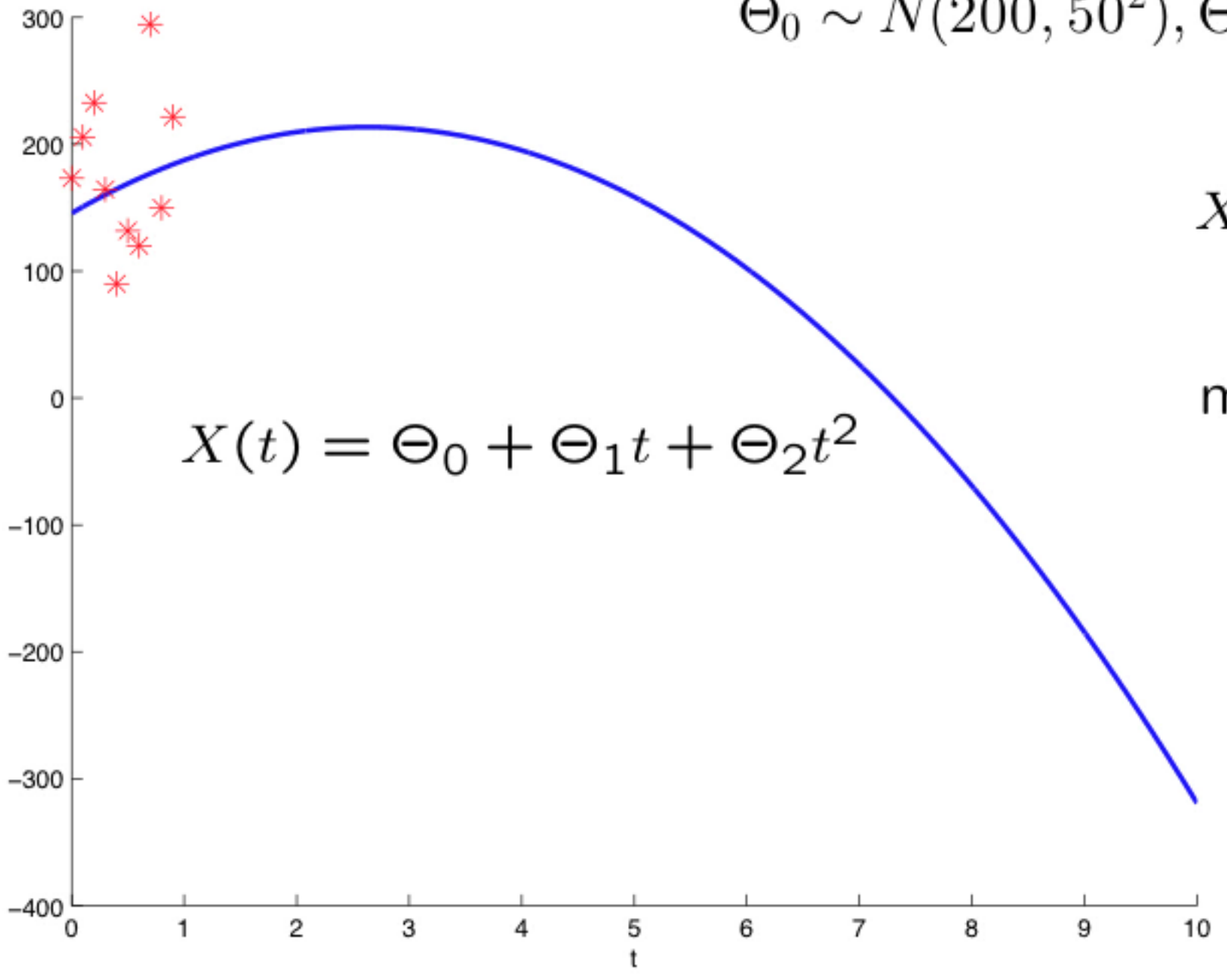
minimize $\theta_0, \theta_1, \theta_2$

$$\frac{1}{2}\left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2}\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2$$

An illustration

Estimating the trajectory of a free-falling object

$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$



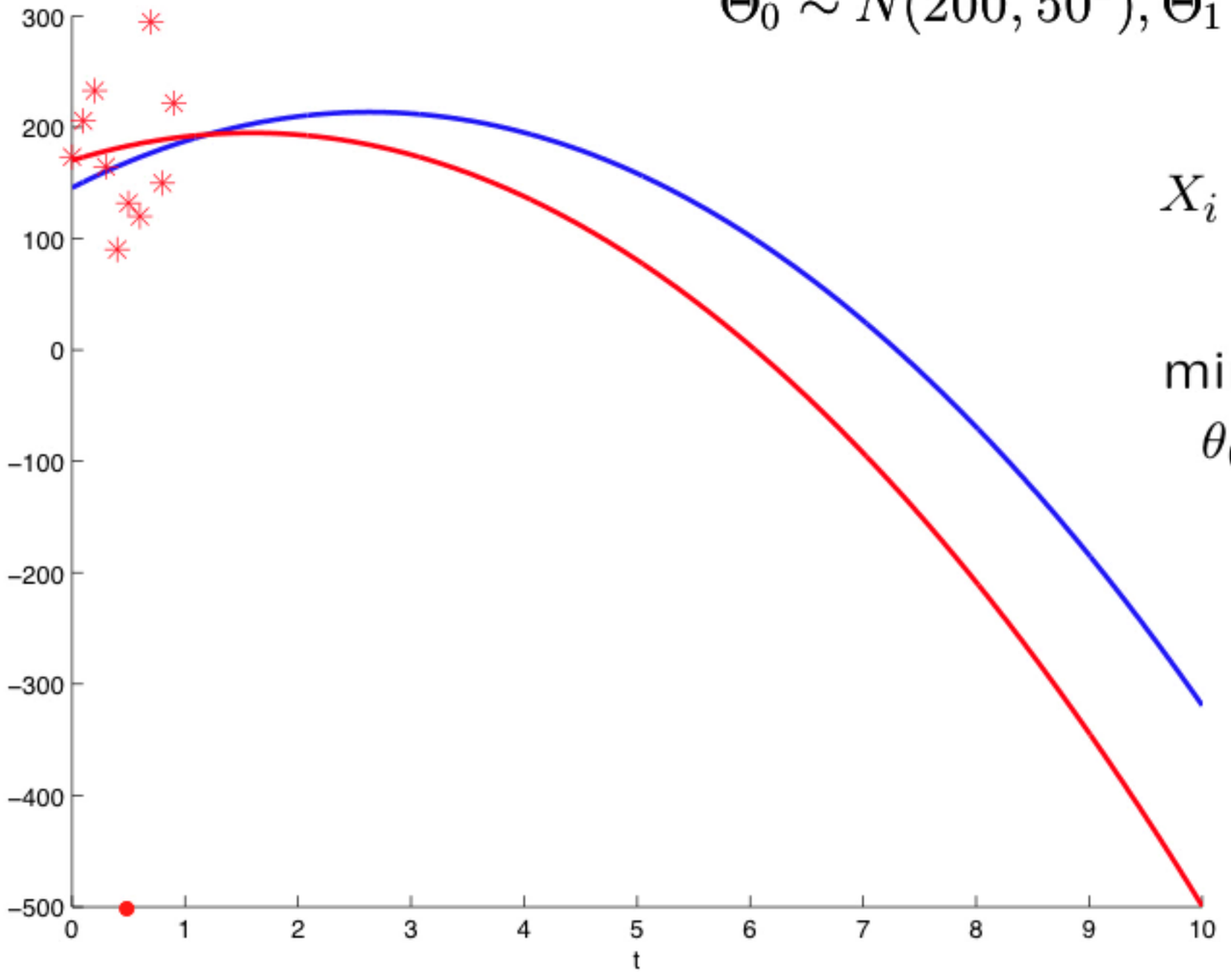
$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

minimize θ_0, θ_1

$$(\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

An illustration

Estimating the trajectory of a free-falling object



$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

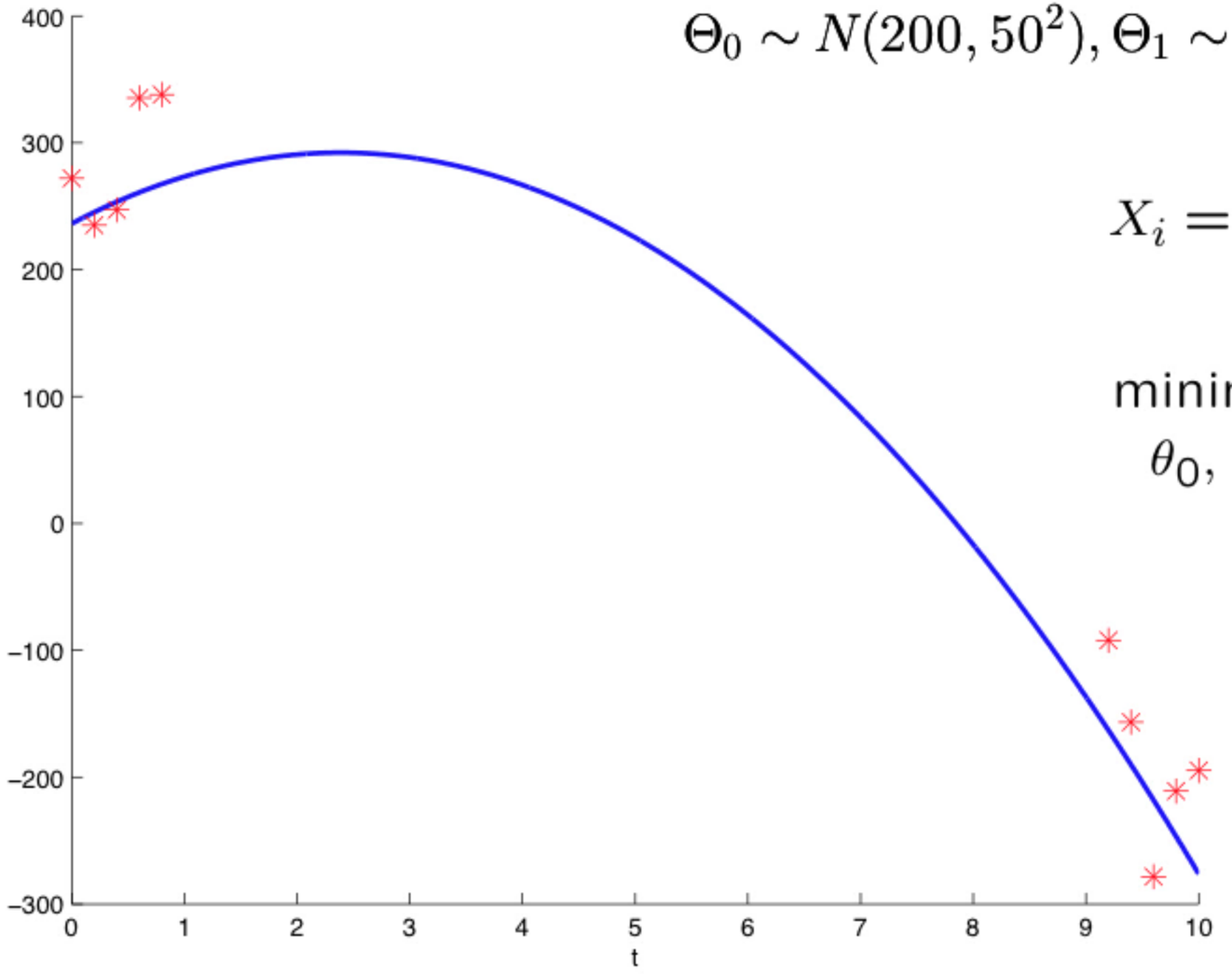
$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

minimize θ_0, θ_1

$$(\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

An illustration

Estimating the trajectory of a free-falling object



$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

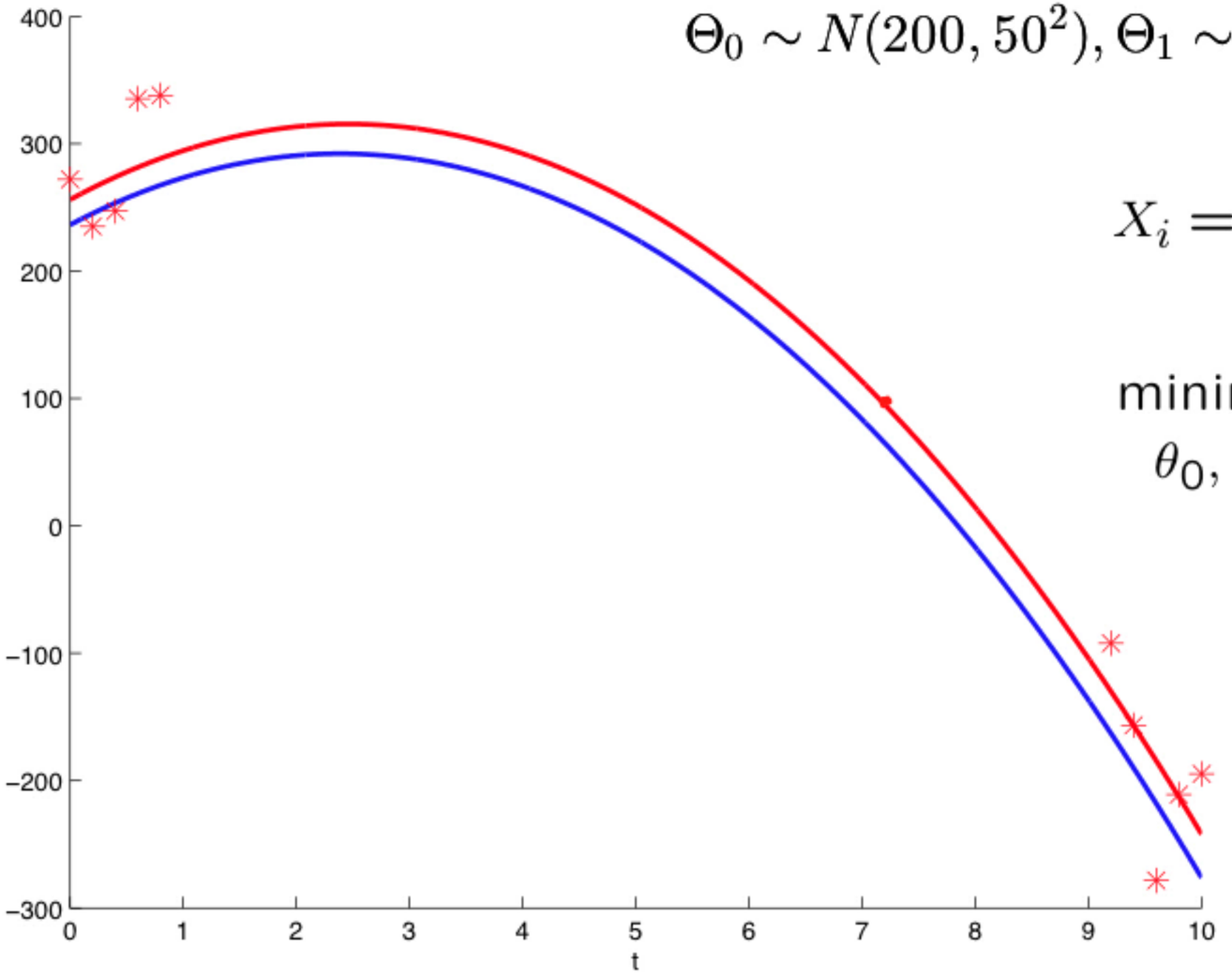
$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

minimize θ_0, θ_1

$$(\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

An illustration

Estimating the trajectory of a free-falling object



$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$X_i = \underbrace{\Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2}_{x(t)} + W_i$$

minimize θ_0, θ_1

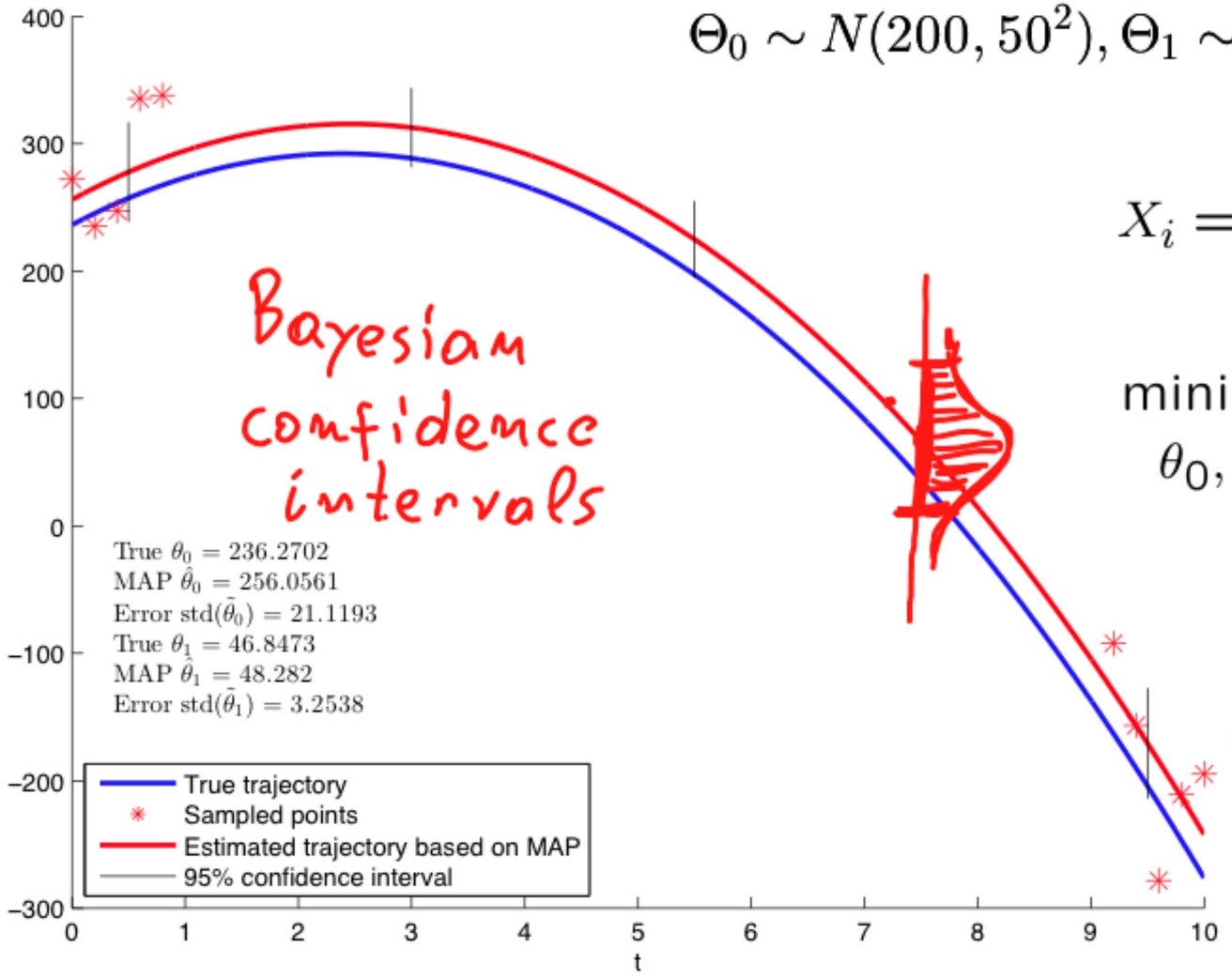
$$(\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

An illustration

Estimating the trajectory of a free-falling object

$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$X_i = \overbrace{\Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2}^{x(t)} + W_i$$



minimize θ_0, θ_1

$$(\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

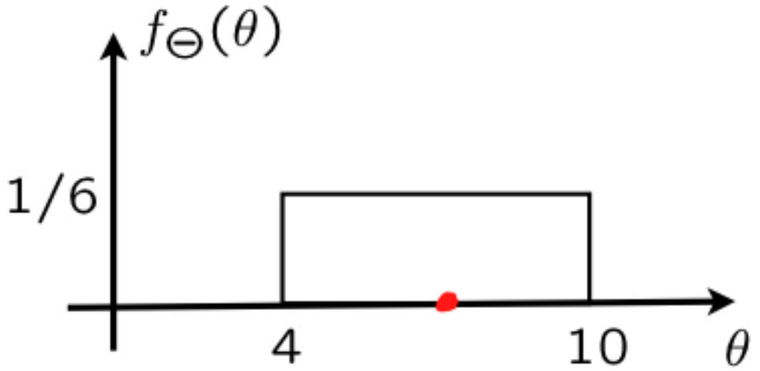
$$P(x(t) \in \text{interval} \mid \text{data}) = 0.95$$

LECTURE 16: Least mean squares (LMS) estimation

- minimize (conditional) mean squared error $\mathbf{E}\left[(\Theta - \widehat{\theta})^2 \,|\, X = x\right]$
  - solution: $\widehat{\theta} = \mathbf{E}[\Theta \,|\, X = x]$
  - general estimation method
- Mathematical properties
- Example

LMS estimation in the absence of observations

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\widehat{\theta}$
  - no observations available
  - MAP rule: any $\hat{\theta} \in [4, 10]$
  - (Conditional) expectation: $\hat{\theta} = 7$



- Criterion: Mean Squared Error (MSE): $\mathbf{E}\left[(\Theta - \widehat{\theta})^2\right]$

minimize mean squared error

LMS estimation in the absence of observations

- Least mean squares formulation:

minimize mean squared error (MSE), $\mathbf{E}\left[(\Theta - \hat{\theta})^2\right]$: $\hat{\theta} = \mathbf{E}[\Theta]$.

$E[\Theta^2] - 2E[\Theta]\hat{\theta} + \hat{\theta}^2$ $\quad \frac{d}{d\hat{\theta}} = 0 : \; -2E[\Theta] + 2\hat{\theta} = 0$

$\hat{\theta} = E[\Theta]$

$Var(\Theta - \hat{\theta}) + (E[\Theta - \hat{\theta}])^2$

$\|$

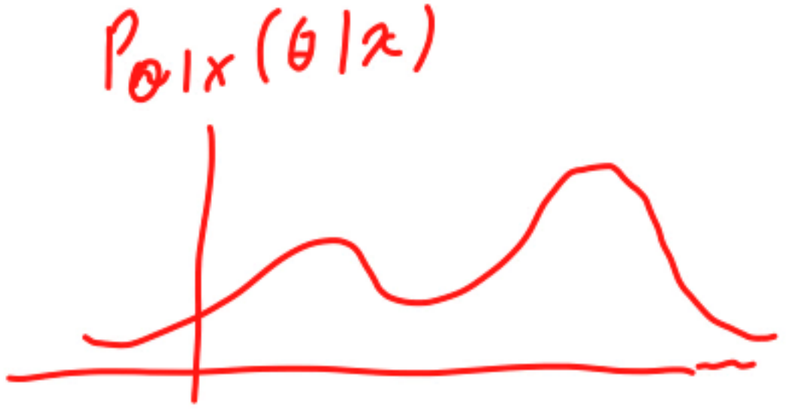
$Var(\Theta)$

minimized when $\hat{\theta} = E[\Theta]$

- Optimal mean squared error: $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta])^2\right] = \text{var}(\Theta)$

LMS estimation of Θ based on X

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\hat{\theta}$
- observation X; model $p_{X|\Theta}(x \mid \theta)$
  - observe that $X = x$



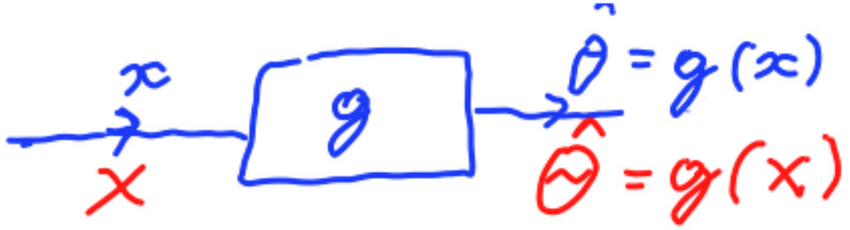
minimize mean squared error (MSE), $\mathbf{E}\left[(\Theta - \hat{\theta})^2\right]$: $\hat{\theta} = \mathbf{E}[\Theta]$

minimize conditional mean squared error, $\mathbf{E}\left[(\Theta - \hat{\theta})^2 \mid X = x\right]$: $\hat{\theta} = \mathbf{E}[\Theta \mid X = x]$

- LMS estimate: $\hat{\theta} = \mathbf{E}[\Theta \mid X = x]$

  estimator: $\widehat{\Theta} = \mathbf{E}[\Theta \mid X]$

LMS estimation of Θ based on X



- $\mathbf{E}[\Theta]$ minimizes $\mathbf{E}\big[(\Theta - \hat{\theta})^2\big]$

$$E\big[(\Theta - E[\Theta])^2\big] \le E\big[(\Theta - c)^2\big], \text{ for all } c$$

- $\mathbf{E}[\Theta \mid X = x]$ minimizes $\mathbf{E}\big[(\Theta - \hat{\theta})^2 \mid X = x\big]$

$$E\big[(\Theta - E[\Theta|X=x])^2 \mid X=x\big] \le E\big[(\Theta - g(x))^2 \mid X=x\big] \text{ for all } x$$

$$E\big[(\Theta - E[\Theta|X])^2 \mid X\big] \le E\big[(\Theta - g(X))^2 \mid X\big]$$

$$E\big[(\Theta - E[\Theta|X])^2\big] \le E\big[(\Theta - g(X))^2\big]$$

$\widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X]$ minimizes $\mathbf{E}\big[(\Theta - g(X))^2\big]$, over all estimators $\widehat{\Theta} = g(X)$

LMS performance evaluation

- LMS estimate: $\widehat{\theta} = \mathbf{E}[\Theta \mid X = x]$

  estimator: $\widehat{\Theta} = \mathbf{E}[\Theta \mid X]$

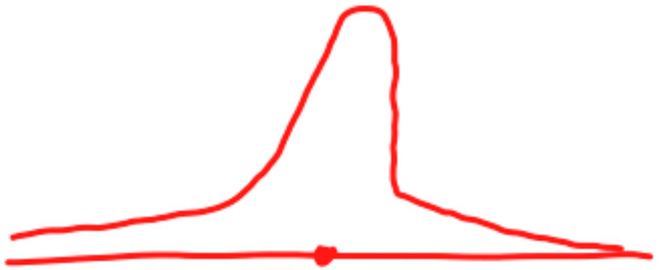
– Expected performance, once we have a measurement:

$$\text{MSE} = \mathbf{E}\Big[\big(\Theta - \mathbf{E}[\Theta \mid X = x]\big)^2 \mid X = x\Big] = \text{var}(\Theta \mid X = x)$$

– Expected performance of the design:

$$\text{MSE} = \mathbf{E}\Big[\big(\Theta - \mathbf{E}[\Theta \mid X]\big)^2\Big] = \mathbf{E}\big[\text{var}(\Theta \mid X)\big]$$

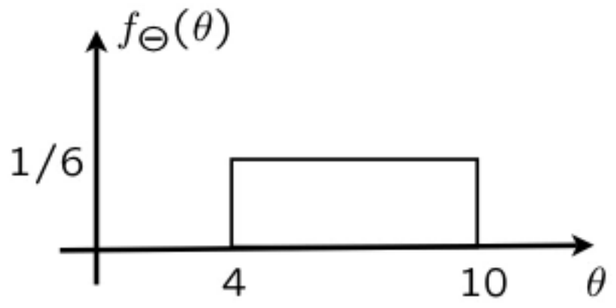
LMS estimation of Θ based on X

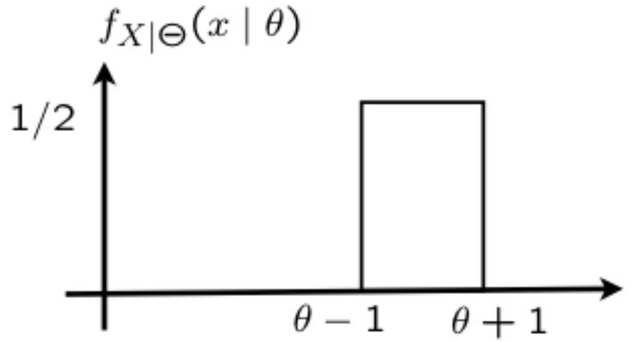


- LMS relevant to estimation (not hypothesis testing)

- 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



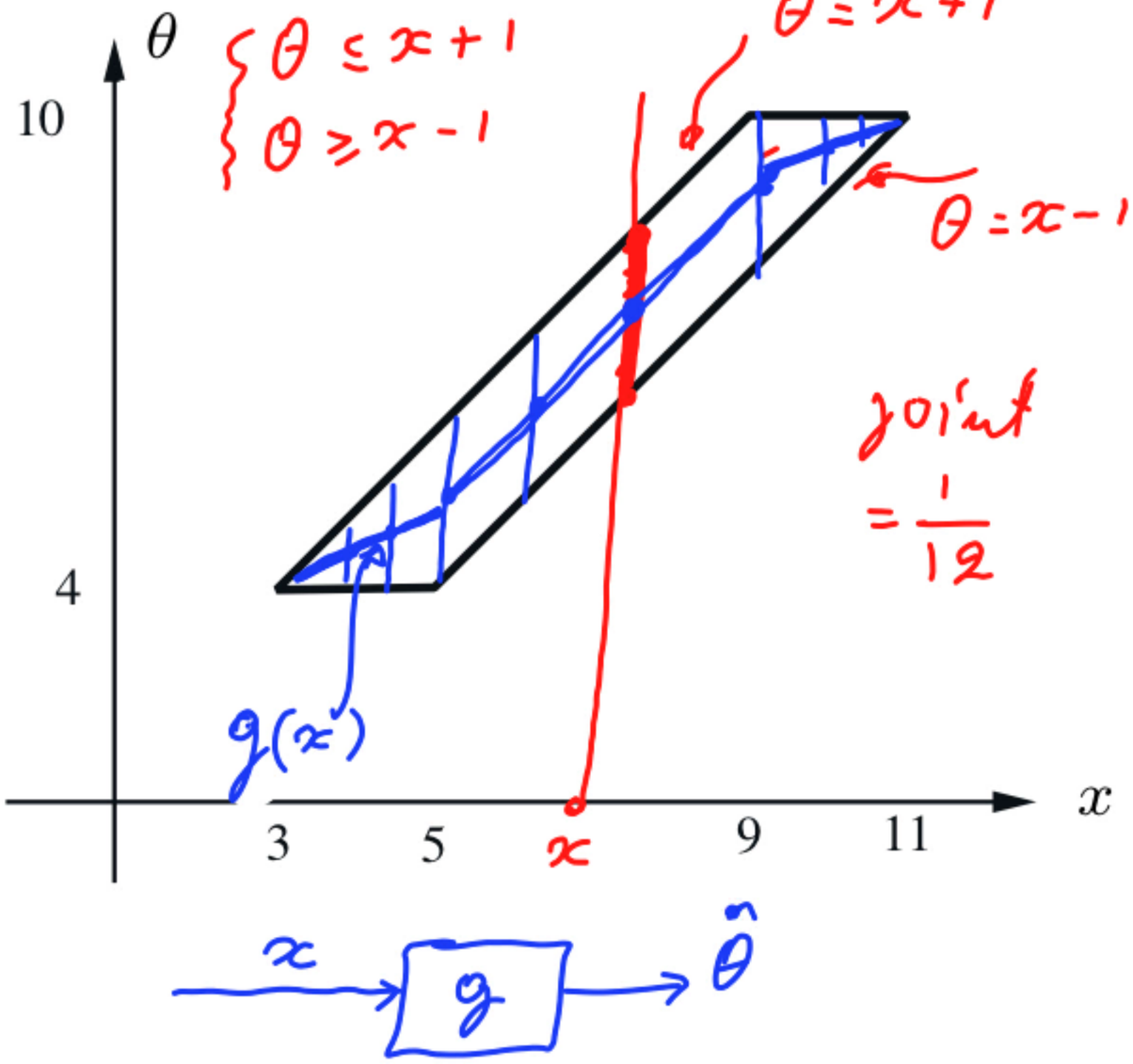


$X = \Theta + U \quad U \sim \text{unif}(-1, 1)$

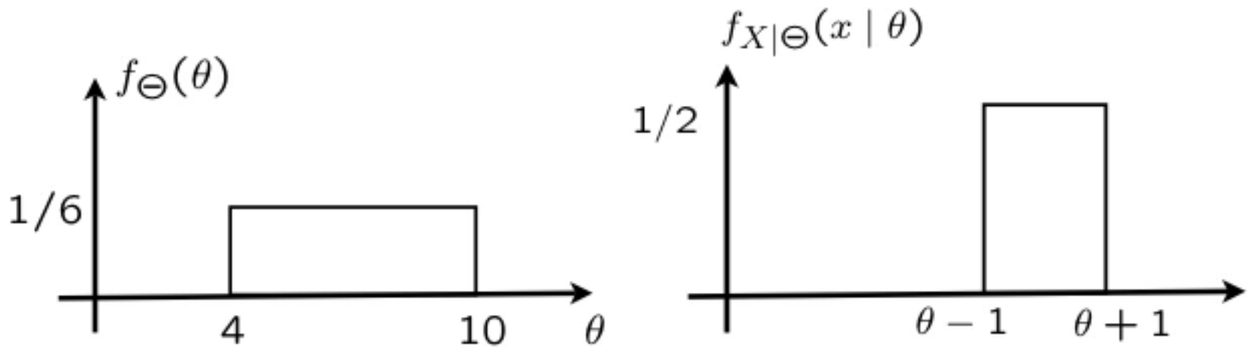
$\theta - 1 \le x \le \theta + 1$

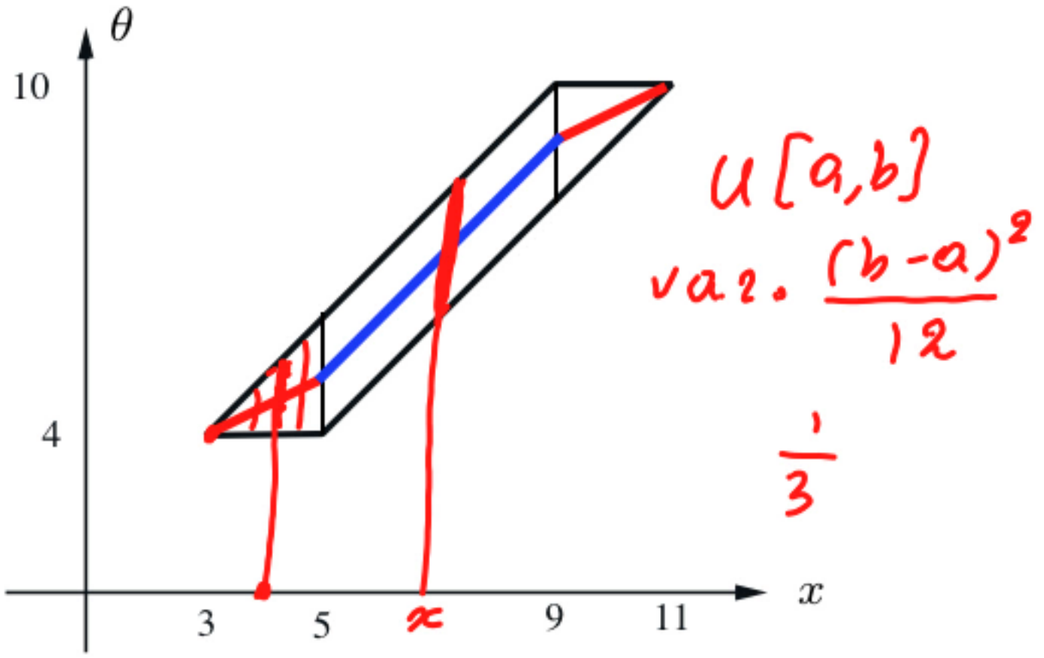
$\theta \le x + 1$

$\theta \ge x - 1$



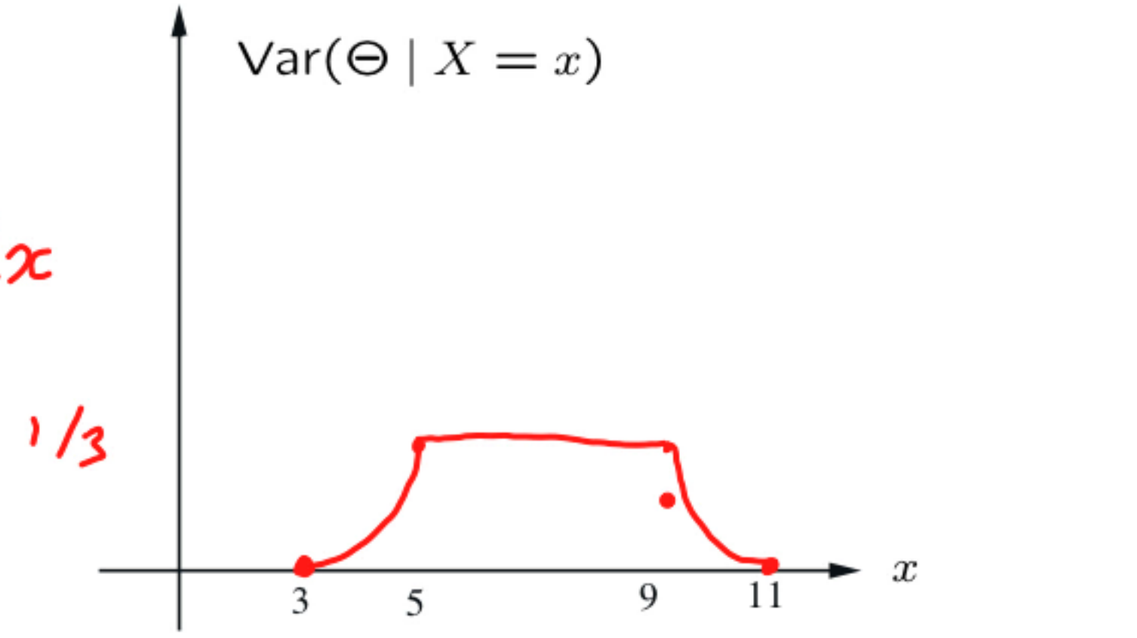
Conditional mean squared error





- $\mathbf{E}[(\Theta - \mathbf{E}[\Theta \mid X = x])^2 \mid X = x]$
  - same as $\mathrm{Var}(\Theta \mid X = x)$: variance of conditional distribution of Θ

$E[\mathrm{Var}(\Theta|X)] = \int f_X(x)\, \mathrm{Var}(\Theta|X=x)\, dx$



LMS estimation with multiple observations or unknowns

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\widehat{\theta}$

- observations $X = (X_1, X_2, \ldots, X_n)$; model $p_{X|\Theta}(x \,|\, \theta)$
  - observe that $X = x$
  - new universe: condition on $X = x$
- LMS estimate: $\mathbf{E}[\Theta \mid X_1 = x_1, \ldots, X_n = x_n]$

- If Θ is a vector, apply to each component separately

$\Theta = (\Theta_1, \ldots, \Theta_m)$ $\quad$ $\hat{\Theta}_j = E[\Theta_j \mid X_1 = x_1, \ldots, X_n = x_n]$

Some challenges in LMS estimation

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta') f_{X|\Theta}(x \mid \theta')\, d\theta'$$

- Full correct model, $f_{X|\Theta}(x \mid \theta)$, may not be available
- Can be hard to compute/implement/analyze

$$E[\Theta_j \mid X = x] = \int\cdots\int \theta_j\, f_{\Theta|X}(\theta \mid x)\, d\theta_1 \cdots d\theta_m$$

Properties of the estimation error in LMS estimation

- Estimator: $\widehat{\Theta} = \mathbf{E}[\Theta \mid X]$
- Error: $\widetilde{\Theta} = \widehat{\Theta} - \Theta$

$E[\hat{\Theta}] = E[\Theta]$

$E[\tilde{\Theta}] = 0$

$\mathbf{E}[\widetilde{\Theta} \mid X = x] = 0$

$E[\hat{\Theta} - \Theta \mid X = x] = \hat{\Theta} - E[\Theta \mid X = x] = 0$

$\mathrm{cov}(\widetilde{\Theta}, \widehat{\Theta}) = 0$

$E[\tilde{\Theta}\hat{\Theta}] - E[\tilde{\Theta}]E[\hat{\Theta}] \quad 0$

$E[\tilde{\Theta}\hat{\Theta} \mid X = x] = \hat{\Theta}E[\tilde{\Theta} \mid X = x] = 0$

$\mathrm{var}(\Theta) = \mathrm{var}(\widehat{\Theta}) + \mathrm{var}(\widetilde{\Theta})$ •

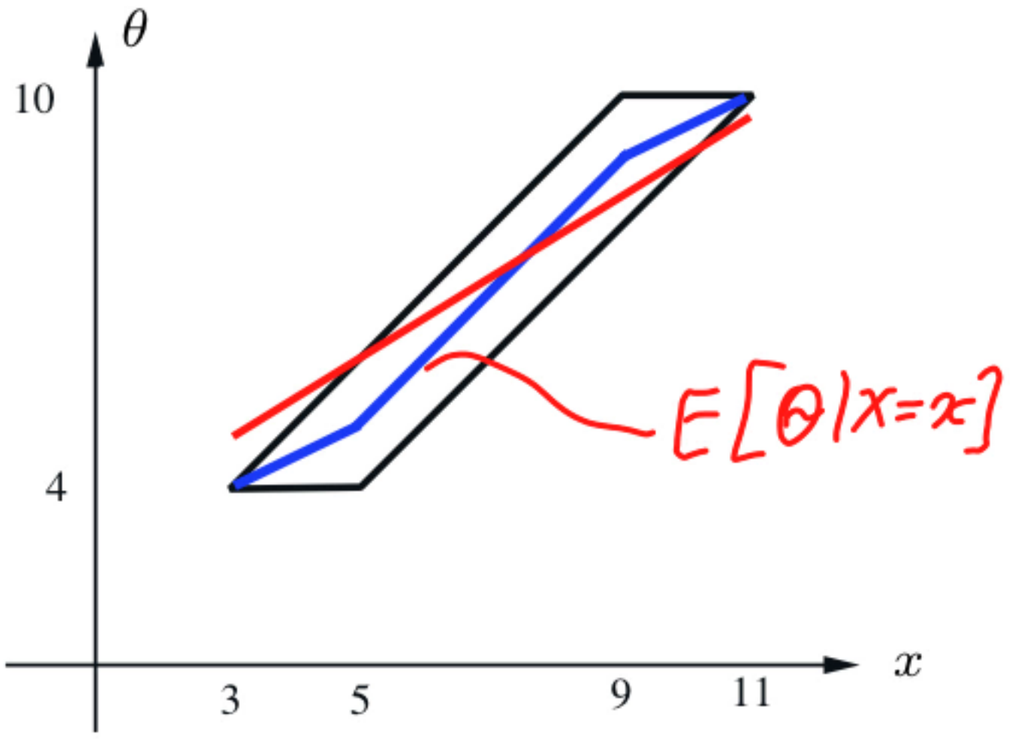
$\Theta = \hat{\Theta} - \tilde{\Theta}$

LECTURE 17: Linear least mean squares (LLMS) estimation

- Conditional expectation $\mathbf{E}[\Theta \,|\, X]$ 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 $\mathbf{E}\left[(\widehat{\Theta} - \Theta)^2\right]$
- Estimators $\widehat{\Theta} = g(X) \longrightarrow \widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \,|\, X]$
- Consider estimators of Θ, of the form $\widehat{\Theta} = aX + b$
- Minimize $\mathbf{E}\left[(\Theta - aX - b)^2\right]$, w.r.t. a, b
- If $\mathbf{E}[\Theta \,|\, X]$ is linear in X, then $\widehat{\Theta}_{\text{LMS}} = \widehat{\Theta}_{\text{LLMS}}$

Solution to the LLMS problem

- Minimize $\mathbf{E}\left[(\Theta - aX - b)^2\right]$, w.r.t. a, b
  - suppose a has already been found: $b = E[\Theta] - a E[X]$

$$\min \; E\left[\left(\Theta - aX - E[\Theta - aX]\right)^2\right] = var(\Theta - aX)$$

$$= var(\Theta) + a^2 var(X) - 2a\, cov(\Theta, X)$$

$$\frac{d}{da} = 0 : \quad 2a\, var(X) - 2\, cov(\Theta, X) = 0$$

$$a = cov(\Theta, X)/var(X)$$

$$\rho = \frac{cov(\Theta, X)}{\sigma_\Theta \sigma_X}$$

$$a = \frac{\rho \sigma_\Theta \sigma_X}{\sigma_X^2}$$

$$\widehat{\Theta}_L = \mathbf{E}[\Theta] + \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{var}(X)}\big(X - \mathbf{E}[X]\big) = \mathbf{E}[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}\big(X - \mathbf{E}[X]\big)$$

Remarks on the solution and on the error variance

$$\widehat{\Theta}_L = \mathbf{E}[\Theta] + \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{var}(X)}\big(X - \mathbf{E}[X]\big) = \mathbf{E}[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}\big(X - \mathbf{E}[X]\big)$$

- Only means, variances, covariances matter
- $\rho > 0$: $X > E[X] \Rightarrow \hat{\Theta}_L > E[\Theta]$
- $\rho = 0$: $\hat{\Theta}_L = E[\Theta]$

$|\rho| = 1$

$\hat{\Theta}_L = \Theta$

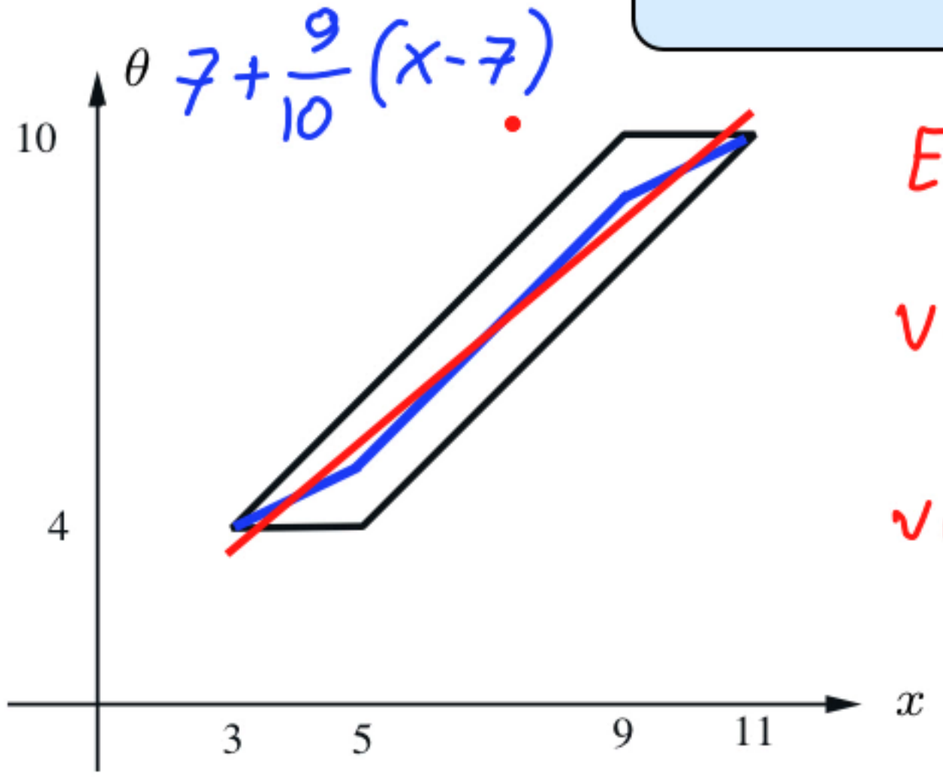
$$\mathbf{E}\big[(\widehat{\Theta}_L - \Theta)^2\big] = (1 - \rho^2)\,\mathrm{var}(\Theta)$$

assume $E[\Theta] = E[X] = 0$

$$E\left[\left(\Theta - \rho\frac{\sigma_\Theta}{\sigma_X}X\right)^2\right] = \sigma_\Theta^2 - 2\rho\frac{\sigma_\Theta}{\sigma_X}\rho\sigma_\Theta\sigma_X + \rho^2\frac{\sigma_\Theta^2}{\sigma_X^2}\sigma_X^2$$

Example

$$\widehat{\Theta}_L = \mathbf{E}[\Theta] + \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{var}(X)}\big(X - \mathbf{E}[X]\big) = \mathbf{E}[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}\big(X - \mathbf{E}[X]\big)$$



$E[\Theta] = 7 \quad E[U] = 0 \quad E[X] = 7$

$\mathrm{var}(\Theta) = \frac{6^2}{12} = 3 \quad \mathrm{var}(U) = \frac{2^2}{12} = \frac{1}{3}$

$\mathrm{var}(X) = 3 + \frac{1}{3} = \frac{10}{3}$

$\mathrm{cov}(\Theta, \Theta + U) =$

$= \mathrm{cov}(\Theta, \Theta) + \mathrm{cov}(\Theta, U) = 3$

Θ : uniform $[4, 10]$

$X = \Theta + U$ uniform $[-1, 1]$

Θ, U independent

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 $[0,1]$

$$\widehat{\Theta}_{\text{LMS}} = \frac{X+1}{n+2} = \widehat{\Theta}_{\text{LLMS}}$$

$$\widehat{\Theta}_{\text{LLMS}} = \mathbf{E}[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}\big(X - \mathbf{E}[X]\big)$$

LLMS for inferring the parameter of a coin

- Θ: uniform on $[0,1]$ $\quad \mathbf{E}[\Theta] = \frac{1}{2}$ $\quad \mathrm{var}(\Theta) = \frac{1}{12}$ $\quad \mathbf{E}[\Theta^2] = \frac{1}{12} + \frac{1}{2^2} = \frac{1}{3}$
- $p_{X|\Theta}$: $\mathrm{Bin}(n,\Theta)$ $\quad \mathbf{E}[X \mid \Theta] = n\Theta$ $\quad \mathrm{var}(X \mid \Theta) = n\Theta(1-\Theta)$

$\mathbf{E}[X] = E[n\Theta] = n/2$ $\quad \mathbf{E}[X^2 \mid \Theta] = n\Theta(1-\Theta) + n^2\Theta^2$

- $\mathbf{E}[X^2] = E[E[X^2 \mid \Theta]] = E[n\Theta + (n^2-n)\Theta^2] = \frac{n}{2} + \frac{n^2-n}{3} = \frac{n}{6} + \frac{n^2}{3}$

$\mathrm{var}(X) = E[X^2] - (E[X])^2 = \frac{n}{6} + \frac{n^2}{3} - \frac{n^2}{4} = \frac{n}{6} + \frac{n^2}{12} = \frac{n(n+2)}{12}$

$\mathbf{E}[\Theta X \mid \Theta] = \Theta E[X \mid \Theta] = n\Theta^2$

$\mathbf{E}[\Theta X] = E[E[\Theta X \mid \Theta]] = E[n\Theta^2] = n/3$

$\mathrm{cov}(\Theta, X) = E[\Theta X] - E[\Theta]E[X] = \frac{n}{3} - \frac{n}{4} = \frac{n}{12}$

LLMS for inferring the parameter of a coin

$$\widehat{\Theta}_{\text{LLMS}} = \mathbf{E}[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}\big(X - \mathbf{E}[X]\big)$$

$$\text{cov}(\Theta, X) = \frac{n}{12} \qquad \text{var}(X) = \frac{n(n+2)}{12} \qquad \mathbf{E}[X] = \frac{n}{2}$$

$$\widehat{\Theta}_{\text{LLMS}} = \frac{X+1}{n+2} = \hat{\Theta}_{LMS}$$

LLMS with multiple observations

- Unknown Θ; observations $X = (X_1, \ldots, X_n)$
- Consider estimators of the form: $\widehat{\Theta} = a_1 X_1 + \cdots + a_n X_n + b$
- Find best choices of $a_1, \ldots, a_n, b$

  minimize: $\mathbf{E}\big[(a_1 X_1 + \cdots + a_n X_n + b - \Theta)^2\big]$ $= a_1^2 E[X_1^2] + 2a_1 a_2 E[X_1 X_2] + \cdots + a_1 E[X_1 \Theta] + \cdots$
- If $\mathbf{E}[\Theta \,|\, X]$ 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

$$X_1 = \Theta + W_1 \qquad \Theta \sim x_0,\ \sigma_0^2 \qquad W_i \sim 0,\ \sigma_i^2$$
$$\vdots$$
$$X_n = \Theta + W_n \qquad \Theta, W_1, \ldots, W_n \text{ uncorrelated}$$

- Suppose $\Theta, W_1, \ldots, W_n$ are independent normal

$$\widehat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = \frac{\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}} \qquad \widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X] = \frac{\frac{x_0}{\sigma_0^2} + \sum_{i=1}^{n} \frac{X_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}} = \widehat{\Theta}_{\text{LLMS}}$$

- 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: $\mathbf{E}[\Theta \mid X]$ is the same as $\mathbf{E}[\Theta \mid X^3]$
  - LLMS is different: estimator $\widehat{\Theta} = aX + b$ versus $\widehat{\Theta} = aX^3 + b$

$cov(\Theta, X^3)$ $var(X^3)$

  - can also consider $\widehat{\Theta} = a_1 X + a_2 X^2 + a_3 X^3 + b$
  - can also consider $\widehat{\Theta} = a_1 X + a_2 e^X + a_3 \log X + b$

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: $X, X_1, \ldots, X_n$ i.i.d.

$$\frac{X_1 + \cdots + X_n}{n} \longrightarrow \mathbf{E}[X]$$

  - 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 $\mathbf{E}[X]$ is small, then X is unlikely to be very large"

Markov inequality: If $X \geq 0$ and $a > 0$, then $\mathbf{P}(X \geq a) \leq \dfrac{\mathbf{E}[X]}{a}$.

$$Y = \begin{cases} 0, & \text{if } X < a \\ a, & \text{if } X \geq a \end{cases}$$

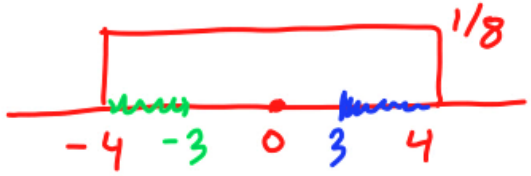
$$a\,P(X \geq a) = E[Y] \leq E[X]$$

The Markov inequality

Markov inequality: If $X \geq 0$ and $a > 0$, then $\mathbf{P}(X \geq a) \leq \dfrac{\mathbf{E}[X]}{a}$

- **Example:** X is Exponential($\lambda = 1$): $\mathbf{P}(X \geq a) \leq \dfrac{1}{a}$

- **Example:** X is Uniform$[-4, 4]$: $\mathbf{P}(X \geq 3) \leq$ $P(|X| \geq 3) \leq \dfrac{E[|X|]}{3} = \dfrac{2}{3}$



$= \frac{1}{2} P(|X| \geq 3) \leq \frac{1}{3}$

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: $\mathbf{P}\big(|X-\mu| \geq c\big) \leq \dfrac{\sigma^2}{c^2}$

Markov inequality: If $X \geq 0$ and $a > 0$, then $\mathbf{P}(X \geq a) \leq \dfrac{\mathbf{E}[X]}{a}$

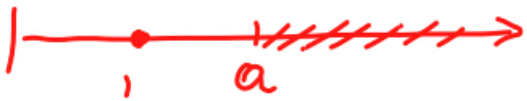
$$P(|X-\mu| \geq c) = P\big((X-\mu)^2 \geq c^2\big) \leq \frac{E[(X-\mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$$

The Chebyshev inequality

Chebyshev inequality: $\mathbf{P}\big(|X-\mu| \geq c\big) \leq \dfrac{\sigma^2}{c^2}$ •

$$\mathbf{P}\big(|X-\mu| \geq k\sigma\big) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2} \qquad k=3 \qquad \leq \frac{1}{9}$$

- **Example:** X is Exponential($\lambda = 1$): $\mathbf{P}(X \geq a) \leq \dfrac{1}{a}$ (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)

- $X_1, X_2, \ldots$ i.i.d.; finite mean μ and variance σ^2

  Sample mean: $M_n = \dfrac{X_1 + \cdots + X_n}{n}$ $\qquad \mu = E[X_i]$

- $\mathbf{E}[M_n] = \dfrac{E[X_1 + \cdots + X_n]}{n} = \dfrac{n\mu}{n} = \mu$

- $\mathrm{Var}(M_n) = \dfrac{\mathrm{Var}(X_1 + \cdots + X_n)}{n^2} = \dfrac{n\sigma^2}{n^2} = \dfrac{\sigma^2}{n}$

$$\mathbf{P}(|M_n - \mu| \geq \epsilon) \leq \frac{\mathrm{var}(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \xrightarrow{n \to \infty} 0 \quad (\text{fixed } \epsilon > 0)$$

WLLN: For $\epsilon > 0$, $\mathbf{P}\big(|M_n - \mu| \geq \epsilon\big) = \mathbf{P}\left(\left|\dfrac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0$, as $n \to \infty$

Interpreting the WLLN

$$M_n = (X_1 + \cdots + X_n)/n$$

WLLN: For $\epsilon > 0$, $\mathbf{P}\big(|M_n - \mu| \geq \epsilon\big) = \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0$, as $n \to \infty$

- One experiment
  - many measurements $X_i = \mu + W_i$
  - W_i: measurement noise; $\mathbf{E}[W_i] = 0$; 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 $p = \mathbf{P}(A)$
  - X_i: indicator of event A
  - the sample mean M_n is the **empirical frequency** of event A

$X_i = 1$, if A occurs
0, o.w.

$E[X_i] = p$

The pollster's problem

- p: fraction of population that will vote "yes" in a referendum
- ith (randomly selected) person polled: $X_i = \begin{cases} 1, & \text{if yes,} \\ 0, & \text{if no.} \end{cases}$

uniformly, independently

$E[X_i] = p$

- $M_n = (X_1 + \cdots + X_n)/n$: fraction of "yes" in our sample
- Would like "small error," e.g.: $|M_n - p| < 0.01$
- Try $n = 10,000$

- $\mathbf{P}\big(|M_{10,000} - p| \geq 0.01\big) \leq \dfrac{\sigma^2}{n\varepsilon^2} = \dfrac{p(1-p)}{10^4 \cdot 10^{-4}} \leq \dfrac{1}{4}$ ← want ≤ 5%

$\dfrac{1/4}{n\,10^{-4}} \leq \dfrac{5}{10^2} \iff n \geq \dfrac{10^6}{20} = 50,000$ ← will suffice

Convergence "in probability"

WLLN: For any $\epsilon > 0$, $\mathbf{P}\Big(|M_n - \mu| \geq \epsilon\Big) \to 0$, as $n \to \infty$

$M_n \xrightarrow[n \to \infty]{i.p} \mu$

- 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:

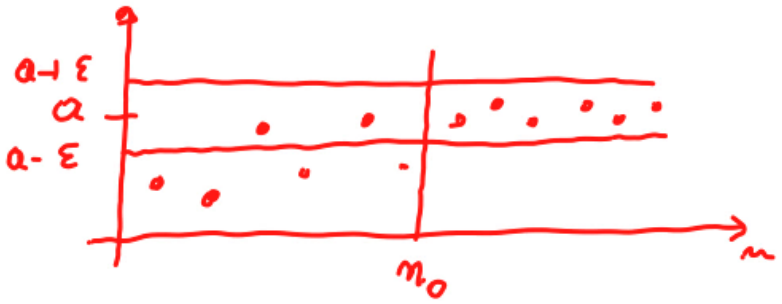
for any $\epsilon > 0$, $\lim_{n \to \infty} \mathbf{P}\Big(|Y_n - a| \geq \epsilon\Big) = 0$

Understanding convergence "in probability"

- Ordinary convergence
  - Sequence a_n; number a

    $a_n \to a$

    "a_n eventually gets and stays (arbitrarily) close to a"

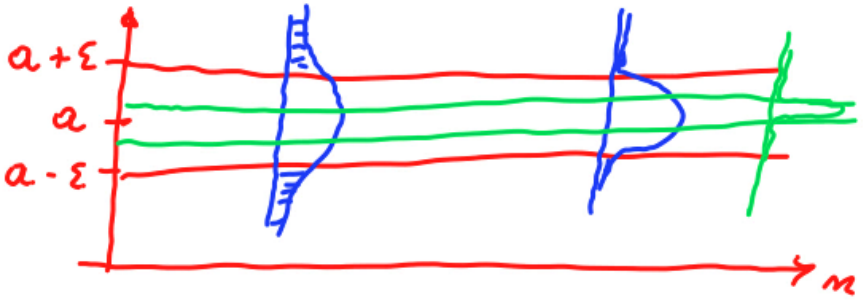


- For every $\epsilon > 0$, there exists n_0, such that for every $n \geq n_0$, we have $|a_n - a| \leq \epsilon$

- Convergence in probability
  - Sequence Y_n; number a

    $Y_n \to a$

- for any $\epsilon > 0$, $\mathbf{P}\Big(|Y_n - a| \geq \epsilon\Big) \to 0$



"(almost all) of the PMF/PDF of Y_n eventually gets concentrated (arbitrarily) close to a"

Some properties

- Suppose that $X_n \to a$, $Y_n \to b$, in probability
- If g is continuous, then $g(X_n) \to g(a)$ $\qquad X_n^2 \to a^2$
- $X_n + Y_n \to a + b$
- **But:** $\mathbf{E}[X_n]$ need not converge to a

Convergence in probability examples

$Y_n \xrightarrow[n\to\infty]{i.p.} 0.$

$\varepsilon > 0 \qquad P(|Y_n - 0| \geq \varepsilon) = 1/n \xrightarrow[n\to\infty]{} 0$

$\mathbf{E}[Y_n] = n^2 \cdot \frac{1}{n} = n \xrightarrow[n\to\infty]{} \infty$

- convergence in probability does **not** imply convergence of expectations

Convergence in probability examples

- X_i: i.i.d., uniform on $[0,1]$
- $Y_n = \min\{X_1, \ldots, X_n\}$

$Y_{n+1} \le Y_n$

$\mathbf{P}\big(|Y_n - 0| \ge \epsilon\big) = P(Y_n \ge \varepsilon)$.

$\varepsilon > 0$

$= P(X_1 \ge \varepsilon, \ldots, X_n \ge \varepsilon)$

$\varepsilon > 1$

$= P(X_1 \ge \varepsilon) \cdots P(X_n \ge \varepsilon)$

$\varepsilon \le 1$

$= (1-\varepsilon)^n \xrightarrow[n\to\infty]{} 0$

$Y_n \xrightarrow[n\to\infty]{i.p.} 0$

Related topics

- Better bounds/approximations on tail probabilities $\leq \sim \frac{1}{n}$
  - Markov and Chebyshev inequalities
  - Chernoff bound $P(|M_n - \mu| \geq a) \leq e^{-n \underbrace{h(a)}_{>0}}$
  - Central limit theorem "$M_n \sim N(\mu, \sigma^2/n)$"
- Different types of convergence
  - Convergence in probability
  - Convergence "with probability 1" $P\left(\{\omega : Y_n(\omega) \xrightarrow[n\to\infty]{} Y(\omega)\}\right) = 1$
  - Strong law of large numbers $M_n \xrightarrow[n\to\infty]{wp1} \mu$
  - Convergence of a sequence of distributions (CDFs) to a limiting CDF

LECTURE 19: The Central Limit Theorem (CLT)

- WLLN: $\dfrac{X_1 + \cdots + X_n}{n} \longrightarrow \mathbf{E}[X]$

- CLT: $X_1 + \cdots + X_n \approx$ 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

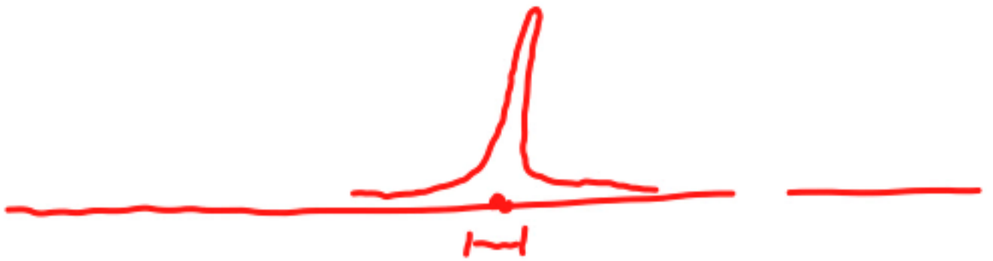
- $X_1, \ldots, X_n$ i.i.d., finite mean μ and variance σ^2

- $S_n = X_1 + \cdots + X_n$

  variance: $n\sigma^2$

- $M_n = \dfrac{S_n}{n} = \dfrac{X_1 + \cdots + X_n}{n}$

  variance: $\dfrac{\sigma^2}{n} \rightsquigarrow 0$



- $\dfrac{S_n}{\sqrt{n}} = \dfrac{X_1 + \cdots + X_n}{\sqrt{n}}$

  variance: $\sigma^2 = \dfrac{n\sigma^2}{n}$

The Central Limit Theorem (CLT)

- $X_1, \ldots, X_n$ i.i.d., finite mean μ and variance σ^2
- $S_n = X_1 + \cdots + X_n$ variance: $n\sigma^2$
- $\dfrac{S_n}{\sqrt{n}} = \dfrac{X_1 + \cdots + X_n}{\sqrt{n}}$ variance: σ^2

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma} \qquad \mathbf{E}[Z_n] = 0 \qquad \text{var}(Z_n) = 1$$

- Let Z be a standard normal r.v. (zero mean, unit variance)

Central Limit Theorem: For every z: $\lim_{n\to\infty} \mathbf{P}(Z_n \leq z) = \mathbf{P}(Z \leq z)$

- $\mathbf{P}(Z \leq z)$ is the standard normal CDF, $\Phi(z)$, available from the normal tables

Usefulness of the CLT

$$S_n = X_1 + \cdots + X_n \qquad Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma} \qquad Z \sim N(0,1)$$

Central Limit Theorem: For every z: $\lim_{n\to\infty} \mathbf{P}(Z_n \le z) = \mathbf{P}(Z \le z)$

- universal and easy to apply; only means, variances matter
- fairly accurate computational shortcut
- justification of normal models

What exactly does the CLT say? — Theory

$$S_n = X_1 + \cdots + X_n \qquad Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma} \qquad Z \sim N(0,1)$$

Central Limit Theorem: For every z: $\lim_{n\to\infty} \mathbf{P}(Z_n \le z) = \mathbf{P}(Z \le z)$

- 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": $\mathbf{E}[e^{sZ_n}] \to \mathbf{E}[e^{sZ}]$, for all s

What exactly does the CLT say? — Practice

$$S_n = X_1 + \cdots + X_n \qquad Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma} \qquad Z \sim N(0,1)$$

Central Limit Theorem: For every z: $\lim_{n\to\infty} \mathbf{P}(Z_n \le z) = \mathbf{P}(Z \le z)$

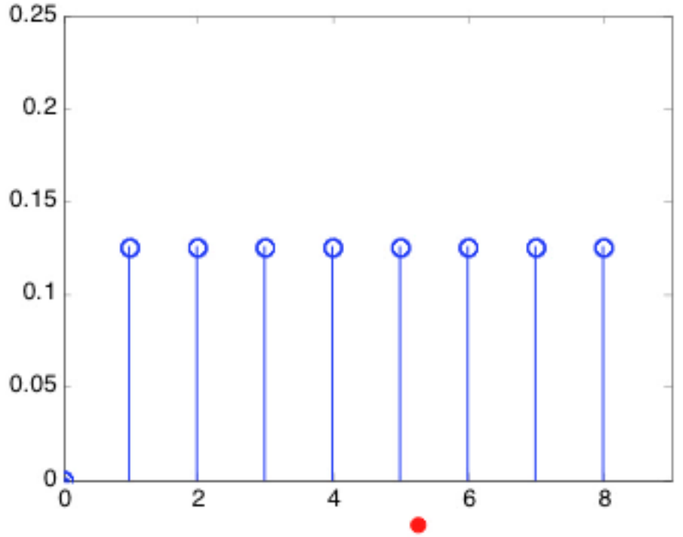
- The **practice** of normal approximations:
  - treat Z_n as if it were normal
  - hence treat S_n as if normal: $\mathcal{N}(n\mu, n\sigma^2)$

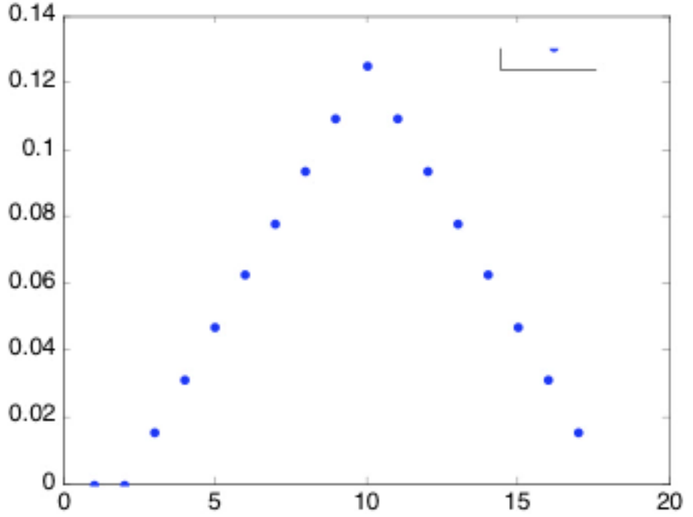
$$S_n = \sqrt{n}\,\sigma Z_n + n\mu$$

- Can we use the CLT when n is "moderate"?

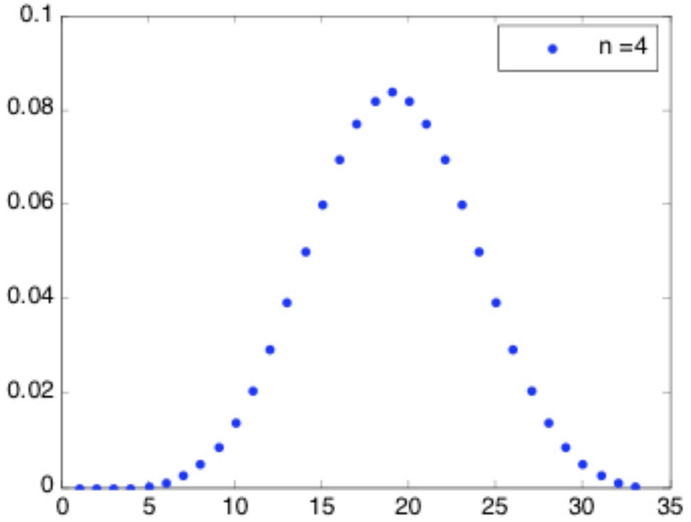
$$n = 30\ ?$$

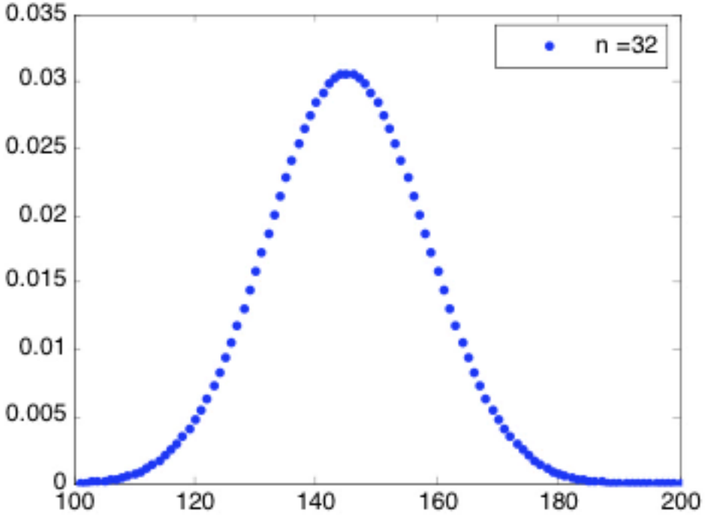
  - usually, yes
  - symmetry and unimodality help

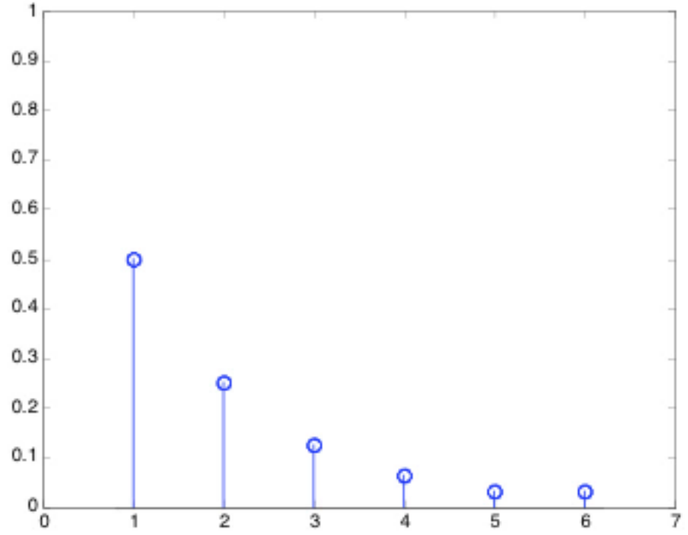

0.25
0.2
0.15
0.1
0.05
0
0
2
4
6
8

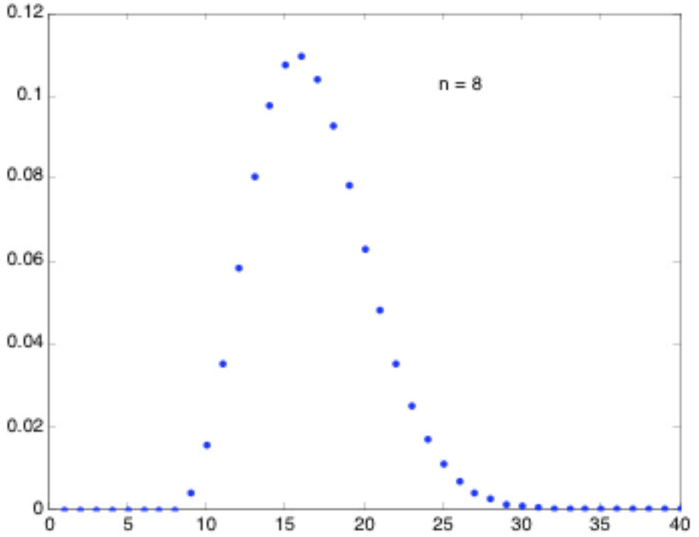

0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
20

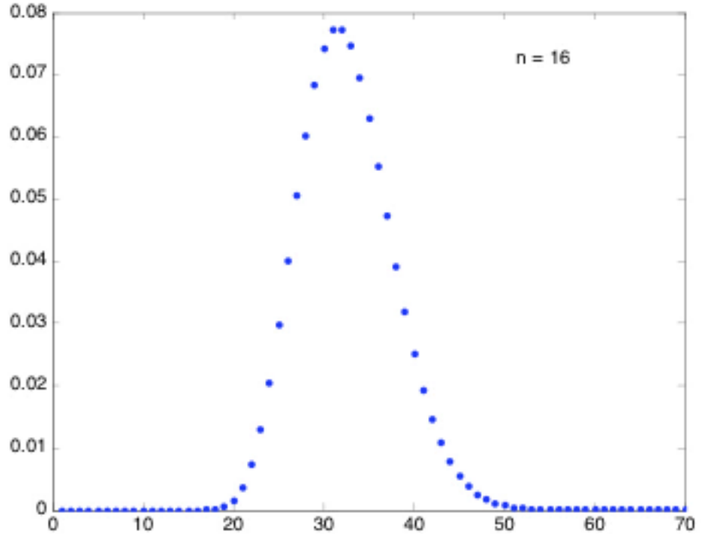
$n = 2$

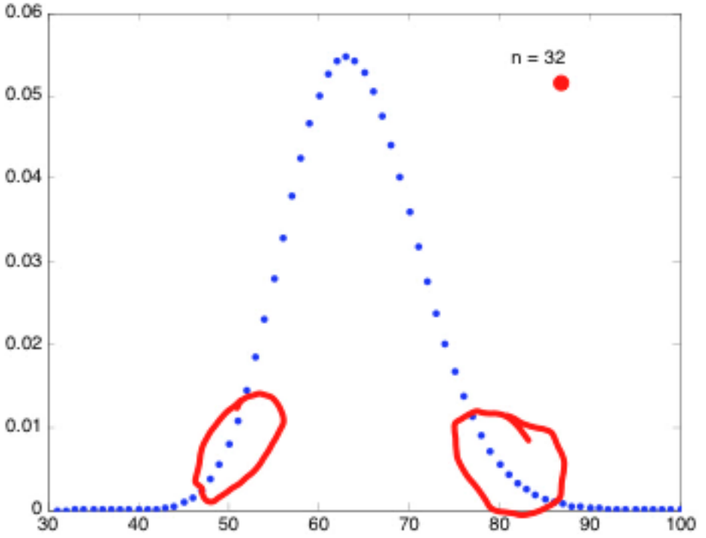

n =4
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
20
25
30
35


n =32
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
100
120
140
160
180
200


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7


0.12
0.1
0.08
0.06
0.04
0.02
0
n = 8
0
5
10
15
20
25
30
35
40


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
n = 16
0
10
20
30
40
50
60
70


0.06
0.05
0.04
0.03
0.02
0.01
0
n = 32
30
40
50
60
70
80
90
100

Example 1

- $\mathbf{P}(S_n \le a) \approx b$ given two parameters, find the third
- Package weights X_i, i.i.d. exponential, $\lambda = 1/2$;
- Load container with $n = 100$ packages

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

$$\mu = \sigma = 2$$

$\mathbf{P}(S_n \ge 210)$

$$= P\left(\frac{S_n - 200}{20} \ge \frac{210 - 200}{20}\right)$$

$$= P(Z_n \ge 0.5) \approx P(Z \ge 0.5)$$

$$= 1 - P(Z < 0.5) = 1 - \Phi(0.5)$$

$$= 1 - .6915 = 0.3085$$

	.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	.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

- $\mathbf{P}(S_n \le a) \approx b$ 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 $\mathbf{P}(S_n \ge a) \approx 0.05$.

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

$$\mu = \sigma = 2$$

$$0.05 \approx P\left(\frac{S_n - 200}{20} \ge \frac{a - 200}{20}\right)$$

$$\approx 1 - \underbrace{\Phi\left(\frac{a-200}{20}\right)}_{0.95}$$

$$\frac{a-200}{20} = 1.645 \qquad a = 232.9$$

	.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

- $\mathbf{P}(S_n \le a) \approx b$ given two parameters, find the third
- Package weights X_i, i.i.d. exponential, $\lambda = 1/2$;
- How large can n be, so that $\mathbf{P}(S_n \ge 210) \approx 0.05$?

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

$$\mu = \sigma = 2$$

$$P\left(\frac{S_n - 2n}{2\sqrt{n}} \ge \frac{210 - 2n}{2\sqrt{n}}\right)$$

$$\approx 1 - \underbrace{\Phi\left(\frac{210 - 2n}{2\sqrt{n}}\right)}_{0.95} \approx 0.05$$

$$\frac{210 - 2n}{2\sqrt{n}} = 1.645$$

$$n = 89$$

	.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

- $\mathbf{P}(S_n \le a) \approx b$ given two parameters, find the third
- Package weights X_i, i.i.d. exponential, $\lambda = 1/2$;

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

$$\mu = \sigma = 2$$

- Load container until weight exceeds 210
  N: number of packages loaded
- $\mathbf{P}(N > 100)$

$$= P\left(\sum_{i=1}^{100} X_i \le 210\right)$$

$$\approx \Phi\left(\frac{210 - 200}{20}\right) = \Phi(0.5)$$

$$= 0.6915$$

	.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); $0 < p < 1$
- $S_n = X_1 + \cdots + X_n$: Binomial(n, p)
  - mean np, variance $np(1-p)$
- $n = 36$, $p = 0.5$; find $\mathbf{P}(S_n \leq 21)$

$np = 18$ $\qquad \sqrt{np(1-p)} = 3$

$$P\left(\frac{S_n - 18}{3} \leq \frac{21-18}{3}\right)$$

$$= P(Z_n \leq 1) \approx \Phi(1) = .8413$$

- CDF of $\dfrac{S_n - np}{\sqrt{np(1-p)}} \longrightarrow$ 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

- $0.8413 \approx \mathbf{P}(S_n \leq 21) = \mathbf{P}(S_n < 22)$, because S_n is integer

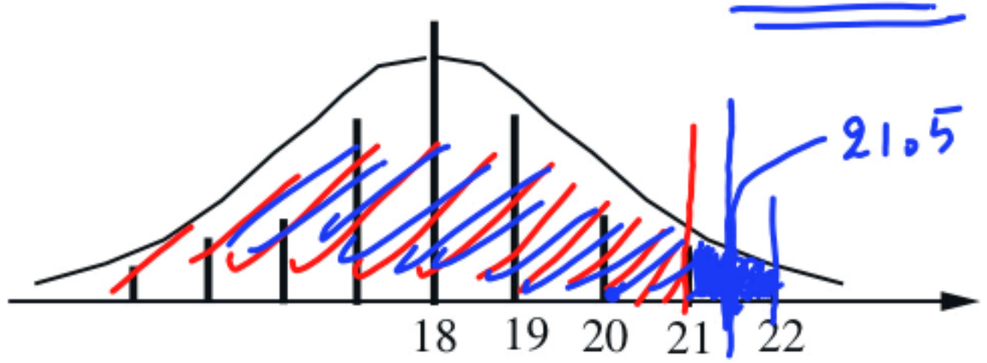
$= P\left(\frac{S_n - 18}{3} < \frac{22-18}{3}\right)$

$= P(Z_n < 1.33) \approx \Phi(1.33) = 0.9082$

true value 0.8785

$\mathbf{P}(S_n \leq 21.5) = P\left(Z_n \leq \frac{21.5-18}{3}\right)$

$\approx \Phi(1.17) = .8790$



	.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

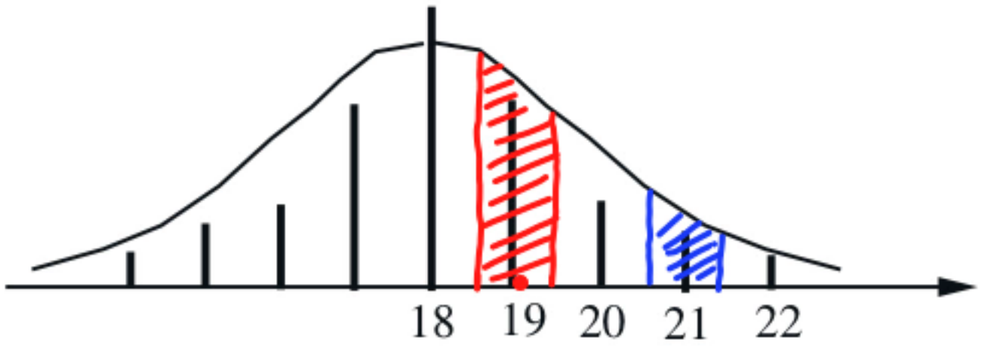
$$\mathbf{P}(S_n = 19) = P(18.5 \le S_n \le 19.5)$$

$$= P\left(\frac{18.5-18}{3} \le Z_n \le \frac{19.5-18}{3}\right)$$

$$= P(0.17 \le Z_n \le 0.5)$$

$$\approx \Phi(0.5) - \Phi(0.17)$$

$$= 0.6915 - 0.56475 = 0.124$$



- 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
- ith (randomly selected) person polled: $X_i = \begin{cases} 1, & \text{if yes,} \\ 0, & \text{if no.} \end{cases}$

$E[X_i] = p = \mu$

$\sigma = \sqrt{p(1-p)}$

- $M_n = (X_1 + \cdots + X_n)/n$: fraction of "yes" in our sample
- Would like "small error," e.g.: $|M_n - p| < 0.01$

$\mathbf{P}\big(|M_n - p| \geq .01\big) = P\left(|Z_n| \geq \frac{.01\sqrt{n}}{\sigma}\right) \approx P\left(|Z| \geq \frac{.01\sqrt{n}}{\sigma}\right)$

$N(0,1)$

$Z_n = \dfrac{S_n - n\mu}{\sqrt{n}\,\sigma}$

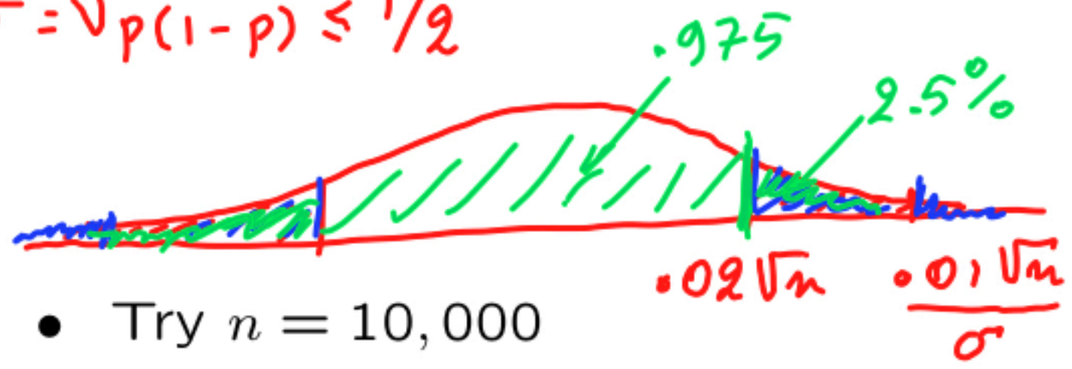
$\left|\dfrac{S_n - np}{n}\right| \geq .01$

$\left|\dfrac{\overbrace{S_n - np}^{Z_n}}{\sqrt{n}\,\sigma}\right| \geq \dfrac{.01\sqrt{n}}{\sigma}$

The pollster's problem revisited

$$\mathbf{P}\Big(|M_n - p| \geq .01\Big) \approx \mathbf{P}\left(|Z| \geq \frac{.01\sqrt{n}}{\sigma}\right) \leq \mathbf{P}\Big(|Z| \geq .02\sqrt{n}\Big) = 2\Big(1 - \Phi\big(\underbrace{.02\sqrt{n}}_{2}\big)\Big) = 0.05$$

$\sigma = \sqrt{p(1-p)} \leq 1/2$



- Try $n = 10,000$

prob $\leq 2(1 - \Phi(2)) =$

$= 2(1 - .9772) = 0.046$

- Specs: $\mathbf{P}\Big(|M_n - p| \geq .01\Big) \leq .05$

$\Phi(.02\sqrt{n}) = 0.975$

$.02\sqrt{n} = 1.96 \Rightarrow n = 9604$

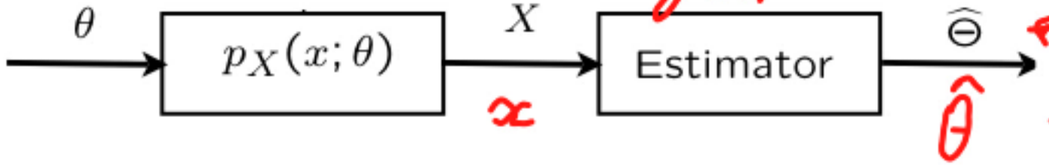
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.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

LECTURE 20: An introduction to classical statistics

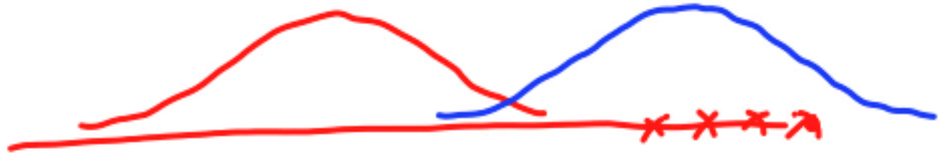
- Unknown constant θ (not a r.v.)
- if $\theta = \mathbf{E}[X]$: estimate using the sample mean $(X_1 + \cdots + X_n)/n$
  - terminology and properties
- Confidence intervals (CIs)
  - CIs using the CLT
  - CIs when the variance is unknown
- Other uses of sample means
- Maximum Likelihood estimation

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

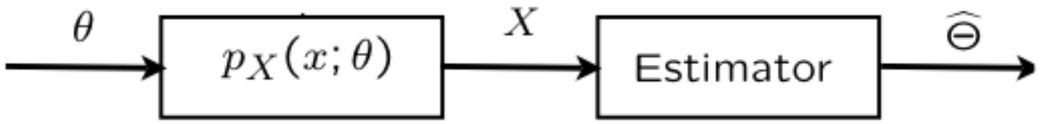


  - also for vectors X and θ: $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)$
  - $p_X(x;\theta)$ are NOT conditional probabilities; θ is NOT random
  - mathematically: many models, one for each possible value of θ



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

- $X_1, \ldots, X_n$: i.i.d., mean θ, variance σ^2

$\widehat{\Theta}_n$ = sample mean $= M_n = \dfrac{X_1 + \cdots + X_n}{n}$ $\quad$ $\widehat{\Theta}_n$: estimator (a random variable)

Properties and terminology:

- $\mathbf{E}[\widehat{\Theta}_n] = \theta$ (unbiased)
  for all θ
- WLLN: $\widehat{\Theta}_n \to \theta$ (consistency)
  i.p.
  for all θ
- mean squared error (MSE): $\mathbf{E}\big[(\widehat{\Theta}_n - \theta)^2\big] = \mathrm{var}(\hat{\Theta}_n) = \dfrac{\sigma^2}{n}$.

$\hat{\Theta} = g(X)$

$E[\hat{\Theta}] = \sum_x g(x)\, p_X(x;\theta)$

On the mean squared error of an estimator

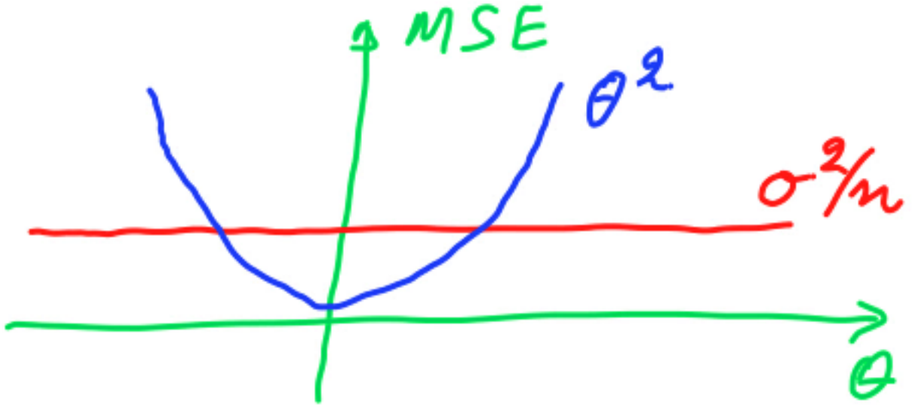
- For any estimator, using $\mathbf{E}[Z^2] = \text{var}(Z) + (\mathbf{E}[Z])^2$:

$$\mathbf{E}[(\widehat{\Theta} - \theta)^2] = \text{var}(\widehat{\Theta} - \theta) + \big(\underbrace{\mathbf{E}[\widehat{\Theta} - \theta]}_{\text{bias}}\big)^2 = \text{var}(\widehat{\Theta}) + (\text{bias})^2$$

$Z = \hat{\Theta} - \theta$

$\hat{\Theta}_n = M_n$: $MSE = \sigma^2/n + 0$

$\hat{\Theta} = 0$: $MSE = 0 + \theta^2$



- $\sqrt{\text{var}(\widehat{\Theta})}$ is called the standard error

Confidence intervals (CIs)

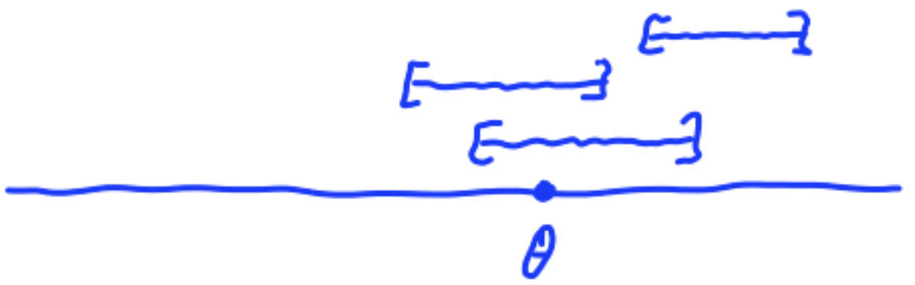
- The value of an estimator $\widehat{\Theta}$ may not be informative enough

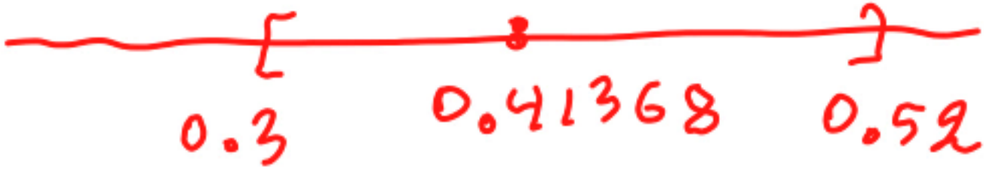
95%

- An $1-\alpha$ **confidence interval** is an interval $[\widehat{\Theta}^-, \widehat{\Theta}^+]$,

  s.t. $\mathbf{P}(\widehat{\Theta}^- \le \theta \le \widehat{\Theta}^+) \ge 1-\alpha$, for all θ

  - often $\alpha = 0.05$, or 0.025, or 0.01
  - interpretation is subtle





$P(0.3 < \theta < 0.52) \ge .95$

CI for the estimation of the mean

iid θ σ^2

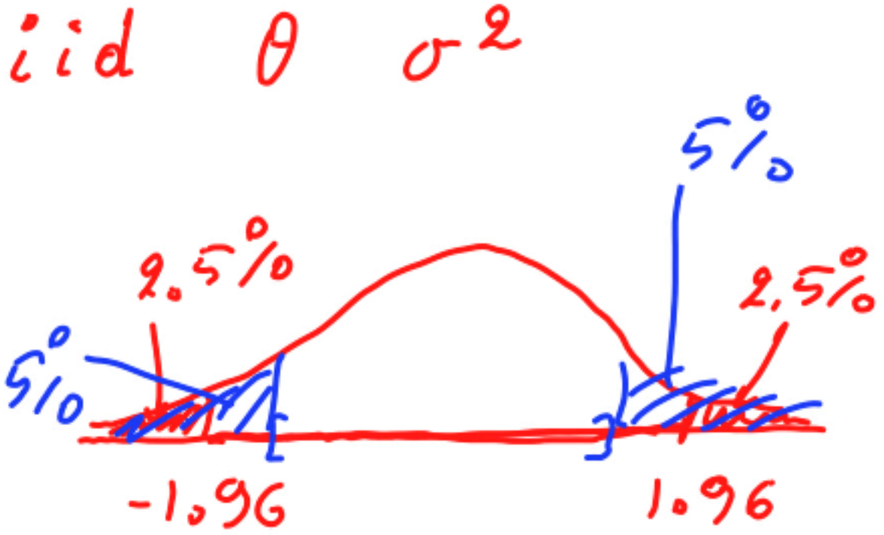
$$\widehat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

95%

normal tables: $\Phi(1.96) = 0.975 = 1 - 0.025$

90%

$\Phi(1.645) = 0.95$



$$\mathbf{P}\left(\frac{|\widehat{\Theta}_n - \theta|}{\sigma/\sqrt{n}} \le 1.96\right) \approx 0.95 \quad \text{(CLT)}$$

$$\mathbf{P}\left(\widehat{\Theta}_n - \frac{1.96\,\sigma}{\sqrt{n}} \le \theta \le \widehat{\Theta}_n + \frac{1.96\,\sigma}{\sqrt{n}}\right) \approx 0.95$$

Θ^- Θ^+

Confidence intervals for the mean when σ is unknown

$$\widehat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

$$\mathbf{P}\left(\widehat{\Theta}_n - \frac{1.96\,\sigma}{\sqrt{n}} \le \theta \le \widehat{\Theta}_n + \frac{1.96\,\sigma}{\sqrt{n}}\right) \approx 0.95$$

- **Option 1:** use upper bound on σ
  - if X_i Bernoulli: $\sigma \le 1/2$
- **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

$$\mathbf{P}\left(\widehat{\Theta}_n - \frac{1.96\,\sigma}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{1.96\,\sigma}{\sqrt{n}}\right) \approx 0.95$$

- **Option 3:** Use sample mean estimate of the variance

Start from $\sigma^2 = \mathbf{E}[(X_i - \theta)^2]$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 \quad \to \quad \sigma^2$$

(but do not know θ)

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\Theta}_n)^2 \quad \to \quad \sigma^2$$

- Two approximations involved here:
  - CLT: approximately normal
  - using estimate of σ
- correction for second approximation (t-tables) used when n is small

Other natural estimators

- $\theta_X = \mathbf{E}[X]$ $\qquad \widehat{\Theta}_X = \dfrac{1}{n}\sum_{i=1}^{n} X_i$
- $\theta = \mathbf{E}[g(X)]$ $\qquad \widehat{\Theta} = \dfrac{1}{n}\sum_{i=1}^{n} g(X_i)$
- $v_X = \mathsf{var}(X) = \mathbf{E}\big[(X-\theta_X)^2\big]$ $\qquad \hat{v}_X = \dfrac{1}{n}\sum_{i=1}^{n}\big(X_i - \widehat{\Theta}_X\big)^2$
- $\mathsf{cov}(X,Y) = \mathbf{E}\Big[(X-\theta_X)(Y-\theta_Y)\Big]$ $\qquad \widehat{\mathsf{cov}}(X,Y) = \dfrac{1}{n}\sum_{i=1}^{n}\big(X_i - \widehat{\Theta}_X\big)\big(Y_i - \widehat{\Theta}_Y\big)$

(X_i, Y_i)

- $\rho = \dfrac{\mathsf{cov}(X,Y)}{\sqrt{v_X}\cdot\sqrt{v_Y}}$ $\qquad \hat{\rho} = \dfrac{\widehat{\mathsf{cov}}(X,Y)}{\sqrt{\hat{v}_X}\cdot\sqrt{\hat{v}_Y}}$
- 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: $\widehat{\Theta}_n \to \theta$
  - asumptotically normal: $\dfrac{\widehat{\Theta}_n - \theta}{\sigma(\widehat{\Theta}_n)} \longrightarrow N(0,1)$ (CDF convergence)
- analytical and simulation methods for calculating $\hat{\sigma} \approx \sigma(\widehat{\Theta}_n)$
  - hence confidence intervals $\mathbf{P}\left(\widehat{\Theta}_n - 1.96\,\hat{\sigma} \le \theta \le \widehat{\Theta}_n + 1.96\,\hat{\sigma}\right) \approx 0.95$
  - asymptotically "efficient" ("best")

ML estimation example: parameter of binomial

- K: binomial with parameters n (known), and θ (unknown)

$$p_K(k;\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

$$\log\left[\binom{n}{k}\right] + k\log\theta + (n-k)\log(1-\theta)$$

$$0 + \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \Rightarrow k - k\theta = n\theta - k\theta$$

$$\hat{\theta}_{\text{ML}} = \frac{k}{n} \qquad \widehat{\Theta}_{\text{ML}} = \frac{K}{n}$$

- same as MAP estimator with uniform prior on θ

ML estimation example — normal mean and variance

- $X_1, \ldots, X_n$: 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 $\quad \frac{n}{2} \log v + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2v}$

– minimize w.r.t. μ: $\quad \hat{\mu} = \frac{x_1 + \cdots + x_n}{n}$

$$\frac{1}{v} \sum_{i=1}^{n} (x_i - \mu) = 0 \quad \Rightarrow \quad \sum x_i = n\mu$$

– minimize w.r.t. v: $\quad \hat{v} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$.

$$\frac{n}{2} \cdot \frac{1}{v} \rightarrow \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2v^2} = 0$$