

LECTURE 1: Probability models and axioms

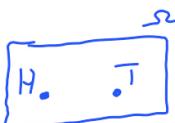
- Sample space
- Probability laws
 - Axioms
 - Properties that follow from the axioms
- Examples
 - Discrete
 - Continuous
- Discussion
 - Countable additivity
 - Mathematical subtleties
- Interpretations of probabilities

Sample space

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

Sample space

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



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

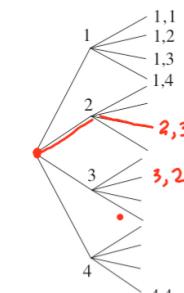
Sample space: discrete/finite example

- Two rolls of a tetrahedral die

4			
3		2,3	
2		3,2	
1	1,1		
	1	2	3

Y = Second roll X = First roll

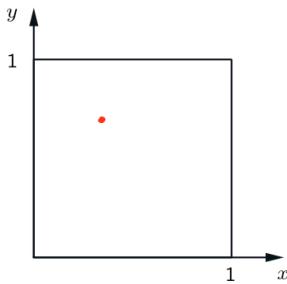
sequential description



Tree

Sample space: continuous example

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



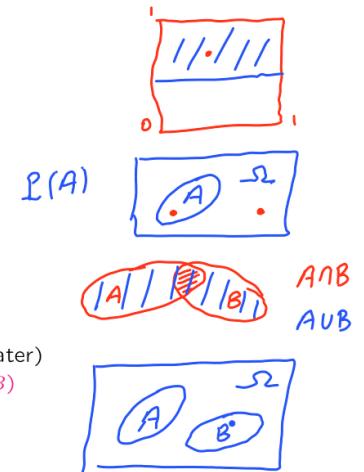
Probability axioms

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

Axioms:

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

empty set



Some simple consequences of the axioms

Axioms	Consequences
$P(A) \geq 0$	$P(A) \leq 1$
$P(\Omega) = 1$	$P(\emptyset) = 0$
For disjoint events:	$P(A) + P(A^c) = 1$
$P(A \cup B) = P(A) + P(B)$	$P(A \cup B \cup C) = P(A) + P(B) + P(C)$
	and similarly for k disjoint events

$$\begin{aligned} P(\{s_1, s_2, \dots, s_k\}) &= P(\{s_1\}) + \dots + P(\{s_k\}) \\ &= P(s_1) + \dots + P(s_k) \end{aligned}$$

Some simple consequences of the axioms

$$\begin{aligned} \text{Axioms} \\ (a) \quad P(A) \geq 0 \\ (b) \quad P(\Omega) = 1 \\ (c) \quad \text{For disjoint events: } P(A \cup B) = P(A) + P(B) \end{aligned}$$

$$\begin{aligned} A \cup A^c &= \Omega \\ A \cap A^c &= \emptyset \\ 1 &= P(\Omega) = P(A \cup A^c) \\ &\stackrel{(c)}{=} P(A) + P(A^c) \\ P(A) &= 1 - \underline{P(A^c)} \stackrel{(a)}{\leq} 1 \end{aligned}$$

$$\begin{aligned} 1 &= P(\Omega) + P(\Omega^c) \\ 1 &= 1 + P(\emptyset) \Rightarrow P(\emptyset) = 0. \end{aligned}$$

Some simple consequences of the axioms

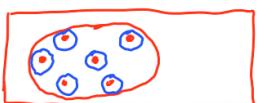
- A, B, C disjoint: $P(A \cup B \cup C) = P(A) + P(B) + P(C)$



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

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

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



$$\begin{aligned} &= P(\{s_1\}) + \dots + P(\{s_k\}) \\ &= P(S_1) + \dots + P(S_k) \end{aligned}$$

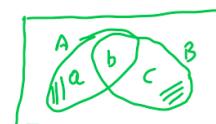
More consequences of the axioms

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



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

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



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

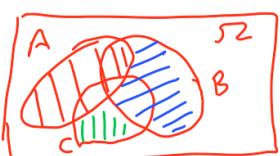
$$P(A \cup B) = a + b + c$$

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

- $P(A \cup B) \leq P(A) + P(B)$ union bound

More consequences of the axioms

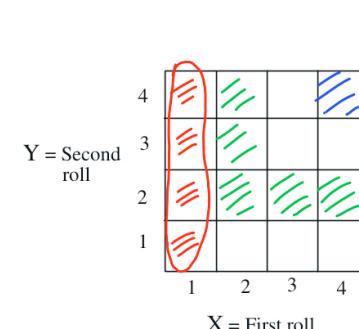
- $P(A \cup B \cup C) = P(A) + \underline{P(A^c \cap B)} + \underline{P(A^c \cap B^c \cap C)}$ • ↗



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

Probability calculation: discrete/finite example

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



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

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

$$\bullet P(Z = 2) = \frac{1}{16}$$

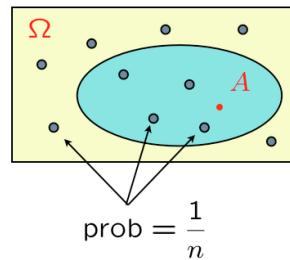
$$\bullet P(Z = 2) = 5 \cdot \frac{1}{16} .$$

Discrete uniform law

finite

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

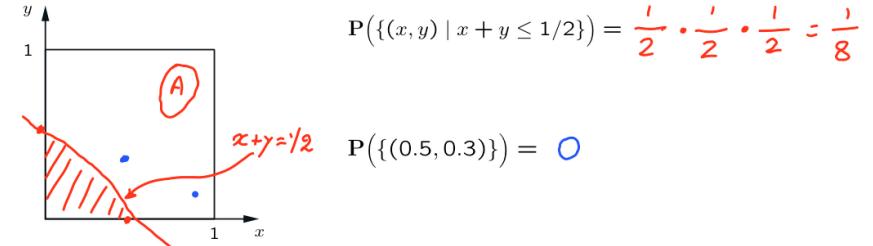
$$P(A) = k \cdot \frac{1}{n}$$



Probability calculation: continuous example

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

- Uniform probability law: Probability = Area



Probability calculation steps

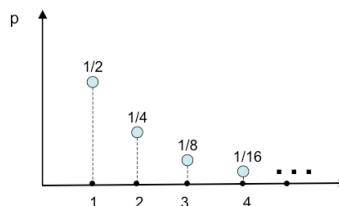
- Specify the sample space
- Specify a probability law
- Identify an event of interest
- Calculate...

Probability calculation: discrete but infinite sample space

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

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

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



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

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

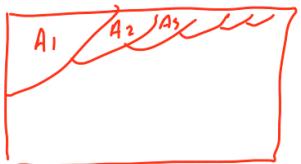
$$= \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6} + \dots = \frac{1}{4} \left(1 + \frac{1}{4} + \frac{1}{4^2} + \dots \right) = \frac{1}{4} \cdot \frac{1}{1 - \frac{1}{4}} = \frac{1}{3}$$

Countable additivity axiom

- Strengthens the finite additivity axiom

Countable Additivity Axiom:

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



Mathematical subtleties

Countable Additivity Axiom:

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

$$I = P(\Omega) = P\left(\bigcup \{\alpha, \gamma\}\right) = \sum P(\{\alpha, \gamma\}) = \sum 0 = 0$$

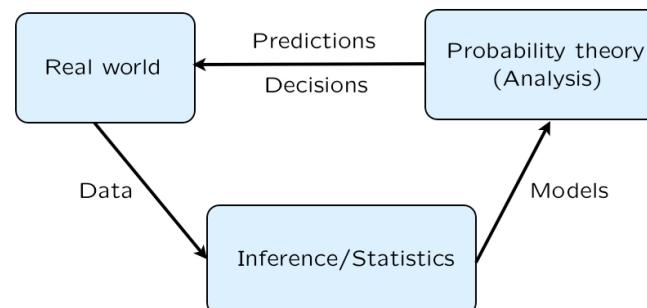
- Additivity holds only for “countable” sequences of events
- The unit square (similarly, the real line, etc.) is **not countable** (its elements cannot be arranged in a sequence)
- “Area” is a legitimate probability law on the unit square, as long as we do not try to assign probabilities/areas to “very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

Interpretations of probability theory

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

The role of probability theory

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

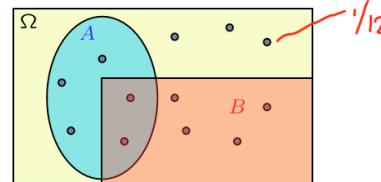


LECTURE 2: Conditioning and Bayes' rule

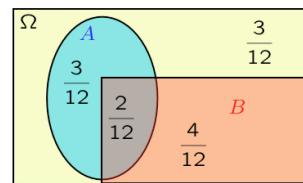
- Conditional probability
- Three **important** tools:
 - Multiplication rule
 - Total probability theorem
 - Bayes' rule (\rightarrow inference)

The idea of conditioning

Assume 12 equally likely outcomes



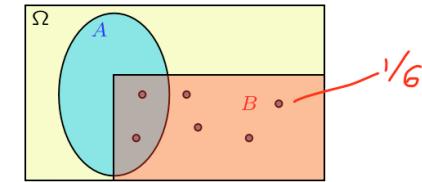
$$P(A) = \frac{5}{12} \quad P(B) = \frac{6}{12}$$



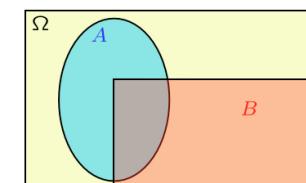
$$P(A) = \frac{3}{12} \quad P(B) = \frac{4}{12}$$

Use new information to revise a model

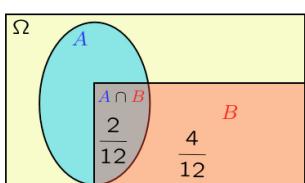
If told B occurred:



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



Definition of conditional probability



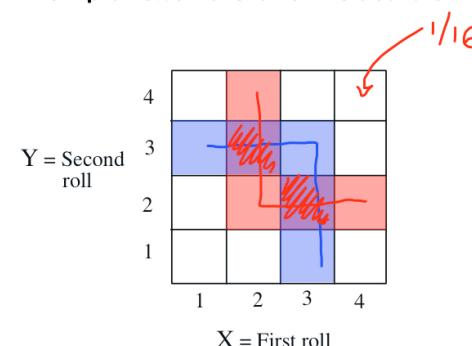
- $P(A | B)$ = "probability of A , given that B occurred"

Def.

$$P(A | B) \stackrel{\Delta}{=} \frac{P(A \cap B)}{P(B)} = \frac{2/12}{4/12} = \frac{1}{2}$$

defined only when $P(B) > 0$

Example: two rolls of a 4-sided die



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

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

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

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

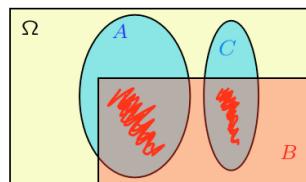
$$= \frac{2/16}{5/16} = \frac{2}{5}$$

Conditional probabilities share properties of ordinary probabilities

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

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

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



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

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

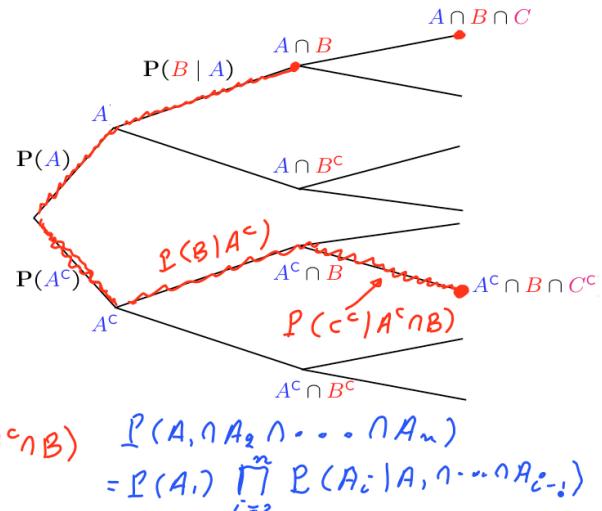
$= P(A|B) + P(C|B)$ also finite countable additivity

The multiplication rule

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

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

$$P(A^c \cap B \cap C^c) = \\ = P(A^c \cap B)P(C^c | A^c \cap B) \\ = P(A^c) \cdot P(B | A^c)P(C^c | A^c \cap B) = P(A^c) \prod_{i=2}^n P(A_i | A_1 \cap \dots \cap A_{i-1}) \\ = P(A^c) \prod_{i=2}^n P(A_i | A_1 \cap \dots \cap A_{i-1})$$



Models based on conditional probabilities

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

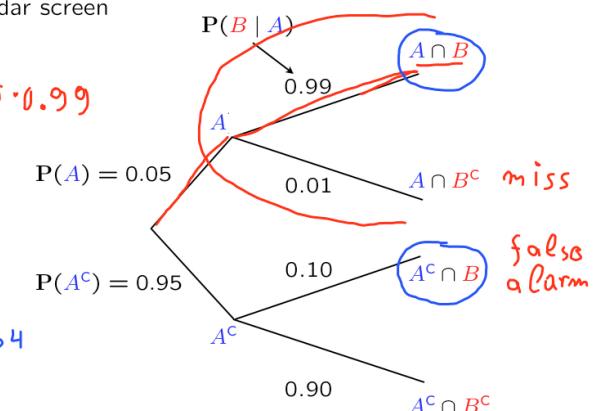
Event A : Airplane is flying above

Event B : Something registers on radar screen

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

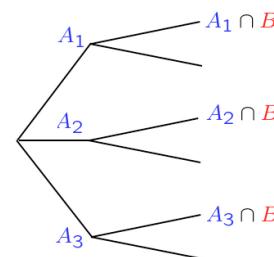
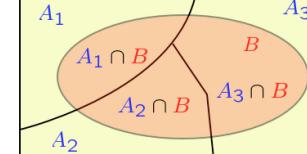
- $P(B) = \frac{0.05 \cdot 0.99}{0.95 \cdot 0.1} = 0.1445$

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



Total probability theorem

$$\Omega = A_1 \cup A_2 \cup A_3$$



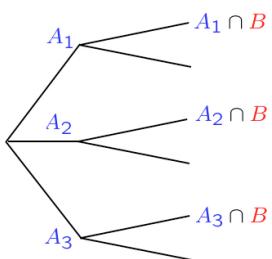
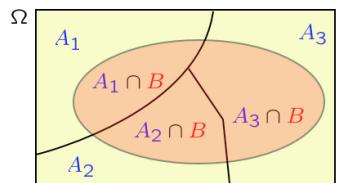
- Partition of sample space into A_1, A_2, A_3, \dots
- Have $P(A_i)$, for every i
- Have $P(B | A_i)$, for every i

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) \\ = P(A_1)P(B | A_1) + \dots + \dots$$

$$\sum_i P(A_i) = 1 \quad \text{weights}$$

$$P(B) = \sum_i P(A_i)P(B | A_i) \quad \text{weighted average of } P(B | A_i)$$

Bayes' rule



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

revised "beliefs," given that B occurred:

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

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

Bayes' rule and inference

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

$$A_i \xrightarrow{\text{model}} B$$

- draw conclusions about causes

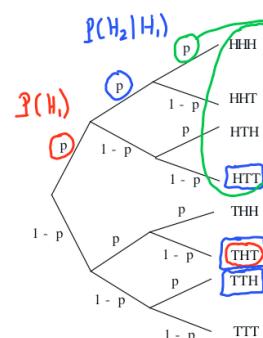
$$B \xrightarrow{\text{inference}} A_i$$

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: $P(H) = p$, $P(T) = 1 - p$



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

- Multiplication rule: $P(THT) = (1-p)p(1-p)$
- Total probability: $P(1 \text{ head}) = 3 p(1-p)^2$
- Bayes rule:

$$P(\text{first toss is } H | 1 \text{ head}) = \frac{P(H, 1 \text{ head})}{P(1 \text{ head})}$$

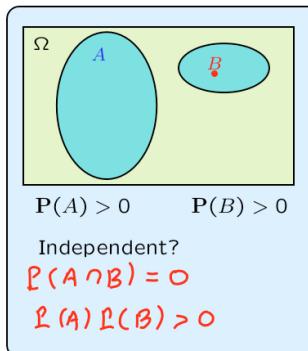
$$= \frac{p(1-p)^2}{3 p(1-p)^2} = \frac{1}{3}$$

Independence of two events

- Intuitive "definition": $P(B | A) = P(B)$
 - occurrence of A provides no new information about B
$$P(A \cap B) = P(A) \quad P(B | A) = P(A) \cdot P(B)$$

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

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



Independence of event complements

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

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

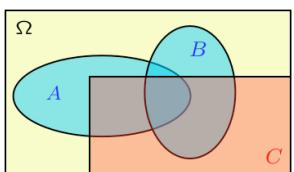
$A = (A \cap B) \cup (A \cap B^c)$

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

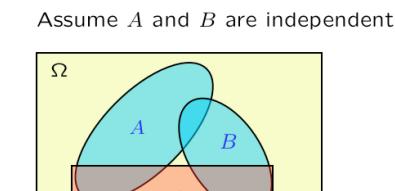
$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) = P(A)(1 - P(B)) \\ &= P(A)P(B^c) \end{aligned}$$

Conditional independence

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



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

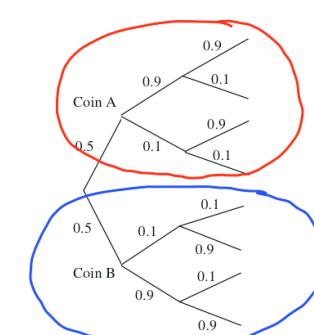


- If we are told that C occurred, are A and B independent? **No**

Conditioning may affect independence

- Two unfair coins, A and B :
 $P(H | \text{coin } A) = 0.9$, $P(H | \text{coin } B) = 0.1$
- choose either coin with equal probability

Given a coin:
independent tosses



Are coin tosses independent? **No!**

Compare:
 $P(\text{toss 11} = H) = P(A)P(H_1 | A) + P(B)P(H_1 | B)$
 $= 0.5 \times 0.9 + 0.5 \times 0.1 = 0.5$

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

$$\approx P(H_1 | A) = 0.9$$

Independence of a collection of events

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

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

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

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

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

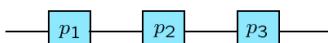
$n = 3:$

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

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

Reliability

p_i : probability that unit i is "up"
independent units



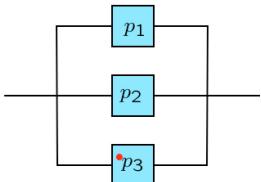
U_i : i th unit up
 U_1, U_2, \dots, U_m independent

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

probability that system is "up"?

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

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



Independence vs. pairwise independence

- Two independent fair coin tosses

– H_1 : First toss is H

– H_2 : Second toss is H

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

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

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

$H_2, C: \text{indep.}$

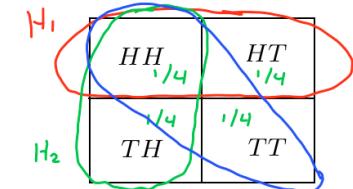
$$P(H_1 \cap H_2 \cap C) = P(HH) = 1/4 \quad \text{diff.}$$

$$P(H_1) P(H_2) P(C) = 1/8$$

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

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

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



The king's sibling

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

boy have precedence

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

independent

BB	BG
GB	GG

2/3

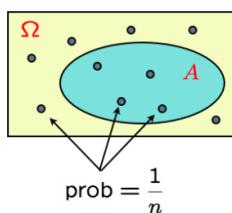
- till 1 boy $\Rightarrow P(G) = 1$
- till 2 boys $\Rightarrow P(G) = 0$

LECTURE 4: Counting

Discrete uniform law

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

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



- Basic counting principle
- Applications

permutations	number of subsets
combinations	binomial probabilities
partitions	

Basic counting principle

4 shirts

3 ties

2 jackets

Number of possible attires?

- r stages
- n_i choices at stage i

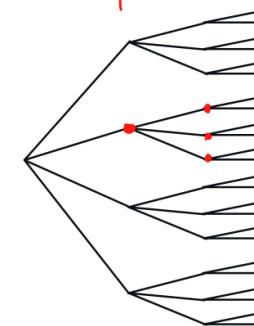
$$4 \cdot 12 \cdot 24 = 4 \cdot 3 \cdot 2$$

$$r = 3$$

$$n_1 = 4$$

$$n_2 = 3$$

$$n_3 = 2$$



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

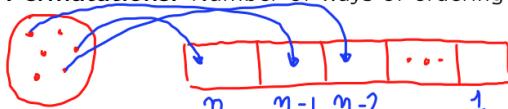
Basic counting principle examples

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

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

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

- Permutations: Number of ways of ordering n elements:



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

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



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

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

$$\{\} \quad \emptyset$$

Example

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

(Assume all outcomes equally likely.)

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

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

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

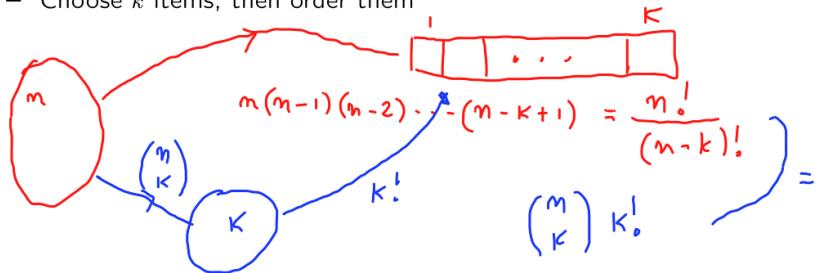
Combinations

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

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

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

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



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

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

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

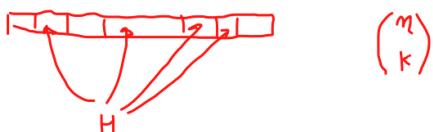
$$\sum_{k=0}^n \binom{n}{k} = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n} = \# \text{all subsets} = 2^n$$

$0! = 1$ convention

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

- $n \geq 1$ independent coin tosses; $P(H) = p$
- $n=6$
- $P(HTTHHH) = p(1-p)(1-p)p p p = p^4(1-p)^2$
- $P(\text{particular sequence}) = p^{\# \text{heads}} (1-p)^{\# \text{tails}}$
- $P(\text{particular } k\text{-head sequence}) = p^k (1-p)^{n-k}$

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



A coin tossing problem

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

Assumptions:

- independence
- $P(H) = p$

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

- First solution:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\text{H}_1, \text{H}_2 \text{ and one H in tosses 3, ..., 10})}{P(B)}$$

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

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

length 10 seg.
3-head segs

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

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

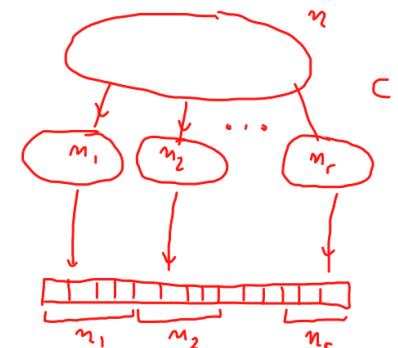
- Assumptions:
- Independence
 - $P(H) = p$

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

Partitions

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

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



$$n=2 \quad n_1=k \quad n_2=n-k$$

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

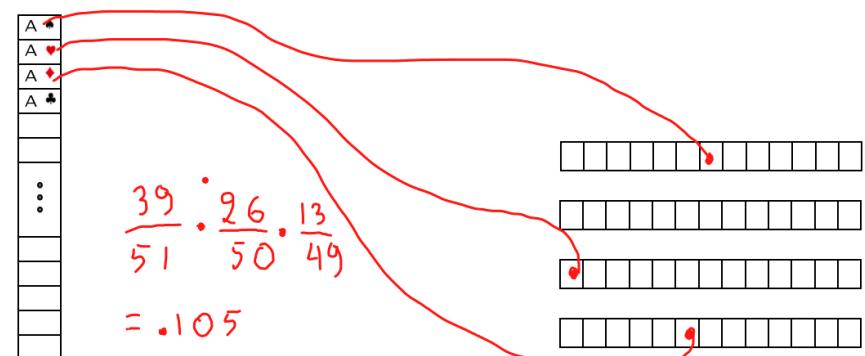
Example: 52-card deck, dealt (fairly) to four players. Find $P(\text{each player gets an ace})$

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

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

A smart solution

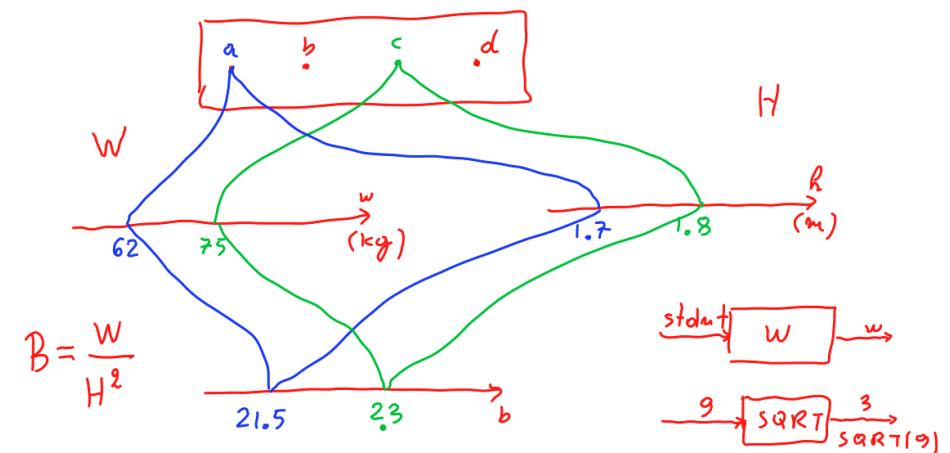
Stack the deck, aces on top



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

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

Random variables: the idea



Random variables: the formalism

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

Notation: random variable X numerical value x

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

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

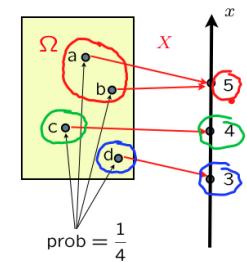
- It is the "probability law" or "probability distribution" of X
- If we fix some x , then " $X = x$ " is an event

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

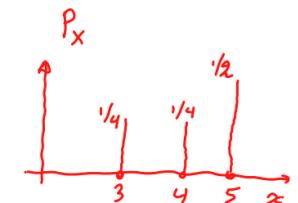
$$p_X(5) = 1/2$$

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

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

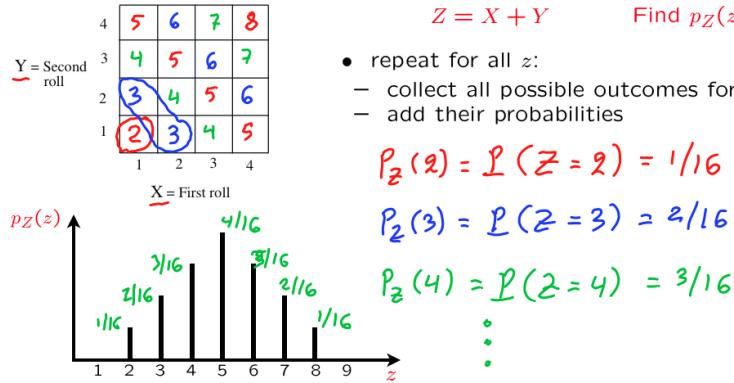


$$P_Y(y)$$



PMF calculation

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



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

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

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

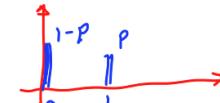
$$P_Z(4) = P(Z=4) = 3/16$$

.

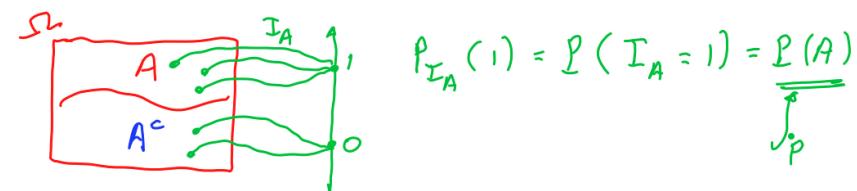
The simplest random variable: Bernoulli with parameter $p \in [0, 1]$

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

$$\begin{aligned} p_x(0) &= 1-p \\ p_x(1) &= p \end{aligned}$$

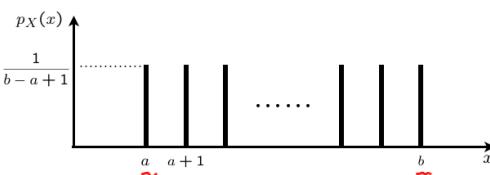


- Models a trial that results in success/failure, Heads/Tails, etc.
- Indicator r.v. of an event A : $I_A = 1$ iff A occurs

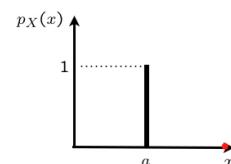


Discrete uniform random variable; parameters a, b

- Parameters: integers a, b ; $a \leq b$
- Experiment: Pick one of $a, a+1, \dots, b$ at random; all equally likely
- Sample space: $\{a, a+1, \dots, b\}$ $b-a+1$ possible values
- Random variable X : $X(\omega) = \omega$ $11:52:26$ $\{0, 1, \dots, 59\}$
- Model of: complete ignorance

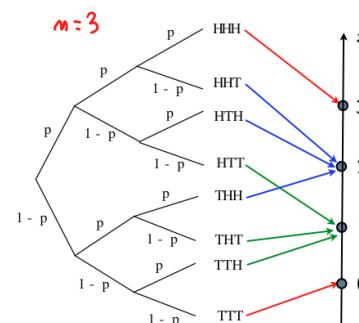


Special case: $a = b$
constant/deterministic r.v.



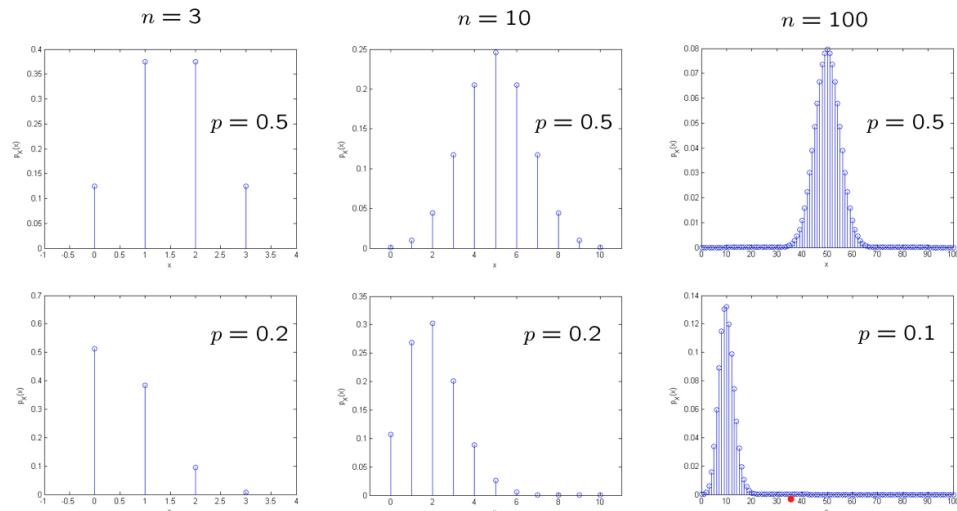
Binomial random variable; parameters: positive integer n ; $p \in [0, 1]$

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



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

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



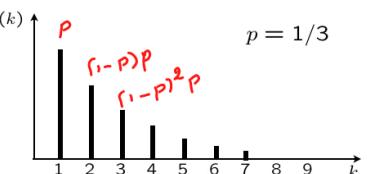
Geometric random variable; parameter p : $0 < p \leq 1$

- Experiment:** infinitely many independent tosses of a coin; $P(\text{Heads}) = p$
- Sample space:** Set of infinite sequences of H and T $\overbrace{\text{TTTTHHT...}}^{\text{X}=5}$
- Random variable X :** number of tosses until the first Heads

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

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

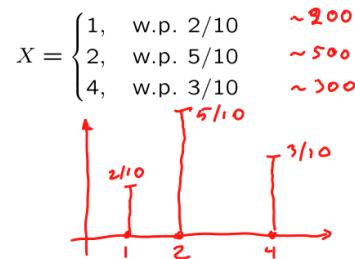
$$\begin{aligned} P(\text{no Heads ever}) &\leq P(\underbrace{\text{T...T}}_k) = (1-p)^k \\ "X=\infty" &\quad \downarrow k \rightarrow \infty \\ &= 0 \end{aligned}$$



Expectation/mean of a random variable

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

$$\begin{aligned} &\underline{1 \cdot 200 + 9 \cdot 500 + 4 \cdot 300} \\ &\quad 1000 \\ &= 1 \cdot \frac{2}{10} + 2 \cdot \frac{5}{10} + 4 \cdot \frac{3}{10} \end{aligned}$$



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

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

- Caution:** If we have an infinite sum, it needs to be well-defined.

$$\text{We assume } \sum_x |x| p_X(x) < \infty$$

Expectation of a Bernoulli r.v.

$$X = \begin{cases} 1, & \text{w.p. } p \\ 0, & \text{w.p. } 1-p \end{cases} \quad E[X] = 1 \cdot p + 0 \cdot (1-p) = p$$

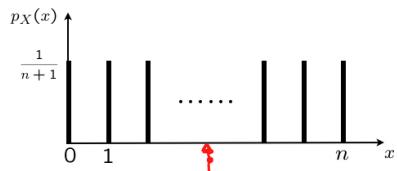
If X is the indicator of an event A , $X = I_A$:

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

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

Expectation of a uniform r.v.

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



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

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

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

Expectation as a population average

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

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

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

Elementary properties of expectations

- If $X \geq 0$, then $E[X] \geq 0$
for all ω : $X(\omega) \geq 0$

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

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

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

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, $E[c] = c$

• $E[c] = c \cdot p(c) = c$

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

- Let X be a r.v. and let $Y = g(X)$

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

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

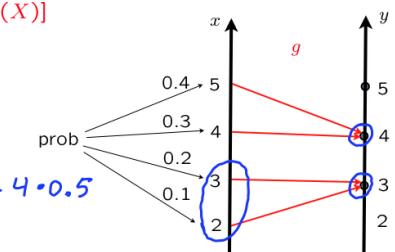
Averaging over x : $3 \cdot 0.1 + 3 \cdot 0.2 + 4 \cdot 0.3 + 4 \cdot 0.5$

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

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

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

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



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

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

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

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

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

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

- Intuitive

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

$$\begin{aligned} g(x) &= ax + b \\ Y &= g(x) \\ E[Y] &= \sum_x g(x) p_x(x) \\ &= \sum_x (ax + b) p_x(x) = a \sum_x x p_x(x) + b \sum_x p_x(x) \\ E[g(x)] &= g(E[x]) = aE[x] + b \end{aligned}$$

exceptional g

Variance — a measure of the spread of a PMF

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

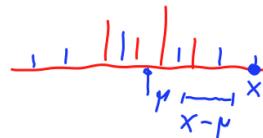
$$E[X - \mu] = E[X] - \mu = \mu - \mu = 0$$

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

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

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

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



≥ 0

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

Properties of the variance

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

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

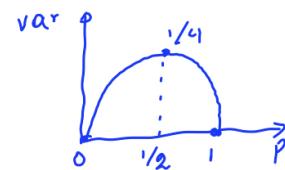
$$\begin{aligned} \text{var}(3 - 4x) &= (-4)^2 \text{var}(x) \\ &= 16 \text{var}(x) \end{aligned}$$

9

Variance of the Bernoulli

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

$$\mathbb{E}[x] = p$$

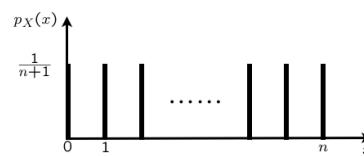


$$\text{var}(X) = \sum_x (x - \mathbb{E}[X])^2 p_X(x) = (1-p)^2 p + (0-p)^2 \cdot (1-p) = p - 2p^2 + p^2 + p^2 - p^2 = p - p^2 = p(1-p)$$

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

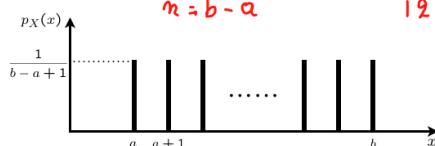
$$X^2 = X$$

Variance of the uniform



$$\text{var}(x) = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 = \frac{1}{n+1} (0^2 + 1^2 + 2^2 + \dots + n^2) - \left(\frac{n}{2}\right)^2 = \frac{1}{12} n(n+2)$$

$n = b - a$



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

Conditional PMF and expectation, given an event

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

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

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

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

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

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

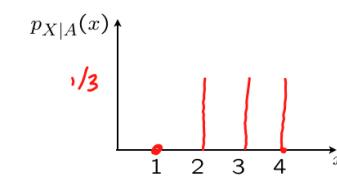
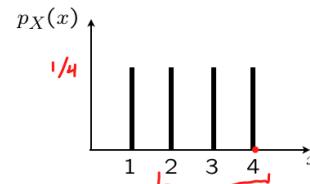
$$\mathbb{E}[X] = \sum_x x p_X(x)$$

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

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

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

Example of conditioning



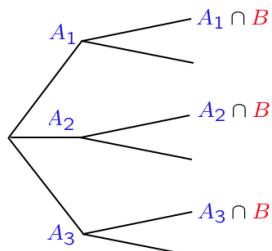
$$\mathbb{E}[X] = 2.5$$

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

$$\mathbb{E}[X | A] = 3$$

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

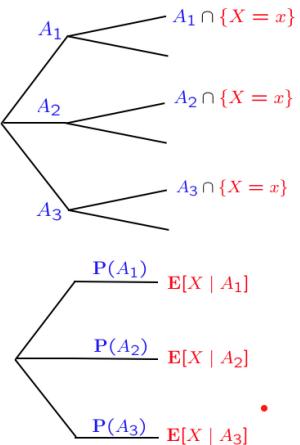
Total expectation theorem



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

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

Total expectation theorem



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

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

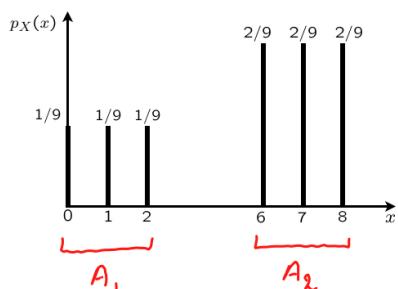
$$p_X(x) = P(A_1)p_{X|A_1}(x) + \dots + P(A_n)p_{X|A_n}(x)$$

for all x

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

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

Total expectation example



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

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

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

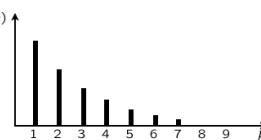
$$E[x|A_2] = 7$$

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

Conditioning a geometric random variable

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

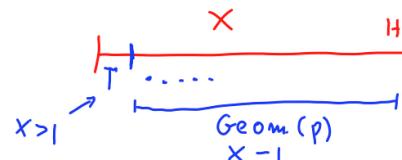
$$p_X(k) = (1-p)^{k-1}p, \quad k = 1, 2, \dots$$



Memorylessness:

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

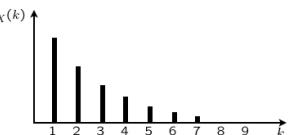
Conditioned on $X > 1$, $X - 1$ is geometric with parameter p



Conditioning a geometric random variable

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

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



Memorylessness:

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

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

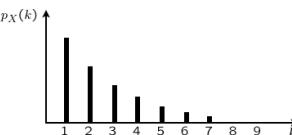
$$\underset{x=1}{\cancel{P_{x-1|X>1}(3)}} = P(X-1=3 | X>1) = P(T_2 T_3 H_4 | T_1) = P(T_2 T_3 H_4)$$

$$P_{x-1|X>1}(k) = p_X(k) = (1-p)^{k-1}p = p_x(3)$$

Conditioning a geometric random variable

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

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



Memorylessness:

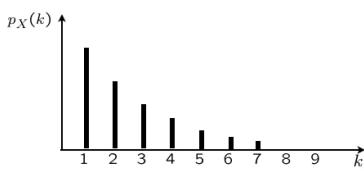
Number of **remaining** coin tosses, conditioned on Tails in the first toss, is **Geometric**, with parameter p

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

$$\underset{x=n}{\cancel{P_{x-n|X>n}(3)}} = P(X-n=3 | X>n) = P(T_2 T_3 H_4 | T_1) = P(T_2 T_3 H_4)$$

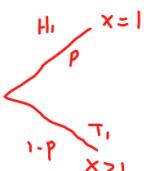
$$P_{x-n|X>n}(k) = p_X(k) = p_{x-n|X>n}(k) = (1-p)^{k-n}p = p_x(3)$$

The mean of the geometric



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

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

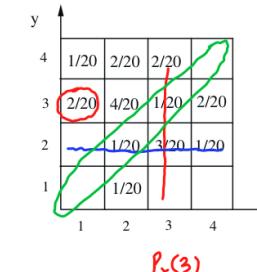


$$\begin{aligned} E[X] &= 1 + E[X-1] \\ &= 1 + p \cdot E[X-1 | X=1] + (1-p)E[X-1 | X>1] \\ &= 1 + 0 + (1-p)E[X] \end{aligned}$$

Multiple random variables and joint PMFs

$$X : p_X \quad Y : p_Y \quad \underset{\text{marginal pmfs}}{\cancel{P(X=Y) = \frac{2}{20}}}$$

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



$$P_{x,y}(1,3) = \frac{2}{20}$$

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

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

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

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

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

More than two random variables

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

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

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

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

Functions of multiple random variables

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

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

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

$$E[g(x)]$$

Linearity of expectations

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

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

$$\begin{aligned} E[x + y] &= E[g(x, y)] \\ (g(x, y) &= x + y) \\ &= \sum_x \sum_y (x + y) p_{X,Y}(x, y) \\ &= \sum_x \underbrace{\sum_y x p_{X,Y}(x, y)}_{\sum_x x p_x(x)} + \sum_y \underbrace{\sum_x y p_{X,Y}(x, y)}_{\sum_y y p_y(y)} \\ &= \sum_x x \underbrace{\sum_y p_{X,Y}(x, y)}_{\sum_x p_x(x)} + \underbrace{\sum_y y \sum_x p_{X,Y}(x, y)}_{\sum_y y p_y(y)} \\ &= \sum_x x p_x(x) + \sum_y y p_y(y) = E[X] + E[Y] \end{aligned}$$

Linearity of expectations

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

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

$$E[X_1 + \dots + X_n] = E[X_1] + \dots + E[X_n]$$

$$E[2X + 3Y - Z] = E[2X] + E[3Y] - E[Z] = 2E[X] + 3E[Y] - E[Z]$$

The mean of the binomial

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

$X_i = 1$ if i th trial is a success; $\overset{p}{\cancel{X_i}}$
 $X_i = 0$ otherwise $\overset{1-p}{\cancel{X_i}}$

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

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

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

$$\boxed{P_X(k)}$$

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

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

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

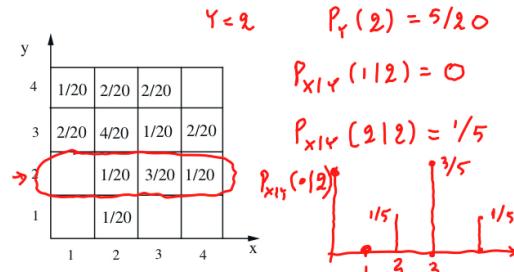
$$p_{X|A}(x | A) = P(X = x | A)$$

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

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

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

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



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

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

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

- Self-explanatory notation

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

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

- Multiplication rule

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

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

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

Conditional expectation

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

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

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

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

- Expected value rule

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

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

Total probability and expectation theorems

$$\bullet \quad A_1, \dots, A_n: \text{partition of } \Omega \quad Y = \{y_1, \dots, y_n\} \quad A_i = \{Y = y_i\}$$

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

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

$$\bullet \quad \mathbb{E}[X] = P(A_1) \mathbb{E}[X | A_1] + \dots + P(A_n) \mathbb{E}[X | A_n]$$

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

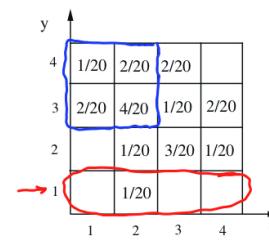
- Fine print:
Also valid when Y is a discrete r.v. that ranges over an infinite set, as long as $\mathbb{E}[|X|] < \infty$

Independence

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

X, Y, Z are independent if:
 $p_{X,Y,Z}(x, y, z) = p_X(x)p_Y(y)p_Z(z), \text{ for all } x, y, z$

Example: independence and conditional independence

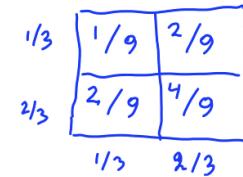


• Independent? No

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

$$p_{x|y}(1|1) = 0$$

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



Yes .

Independence and expectations

- In general: $E[g(X, Y)] \neq g(E[X], E[Y])$
- Exceptions: $E[aX + b] = aE[X] + b$ $E[X + Y + Z] = E[X] + E[Y] + E[Z]$

always true

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

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

$$\begin{aligned} E[g(x, y)] &= g(x, y) = xy \\ &= \sum_x \sum_y xy p_{x,y}(x, y) = \sum_x \sum_y (xy) p_x(x) p_y(y) \\ &= \sum_x x p_x(x) \sum_y y p_y(y) = E[x] E[y] \end{aligned}$$

Independence and variances

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

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

assume
 $E[x] = E[y] = 0$
 $E[xy] = E[x]E[y] = 0$

$$\begin{aligned} \text{var}(x+y) &= E[(x+y)^2] = E[x^2 + 2xy + y^2] \\ &= E[x^2] + 2E[xy] + E[y^2] = \text{var}(x) + \text{var}(y) \end{aligned}$$

- Examples:
 - If $X = Y$: $\text{var}(X + Y) = \text{var}(2X) = 4\text{var}(X)$
 - If $X = -Y$: $\text{var}(X + Y) = \text{var}(0) = 0$
 - If X, Y independent: $\text{var}(X - 3Y) = \text{var}(x) + \text{var}(-3Y) = \text{var}(x) + 9\text{var}(Y)$

Variance of the binomial

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

$X_i = 1$ if i th trial is a success;
 $X_i = 0$ otherwise (indicator variable) *independent*

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

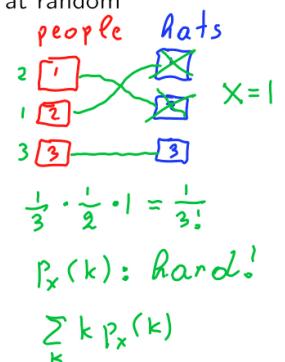
$$\begin{aligned} \text{var}(x) &= \text{var}(X_1) + \dots + \text{var}(X_n) \\ &= n \cdot \text{var}(X_1) = n p(1-p) \end{aligned}$$

The hat problem

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

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

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



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

The variance in the hat problem

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

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

$$\begin{aligned} m &= 2 \\ X_1 &= 1 \Rightarrow X_2 = 1 \\ X_1 &= 0 \Rightarrow X_2 = 0 \end{aligned}$$

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

$$E[X_i^2] = E[X_1^2] = E[X_1] = 1/n \quad E[X^2] = n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n} \cdot \frac{1}{n-1}$$

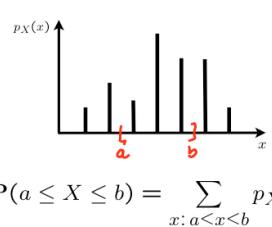
$$\text{For } i \neq j: E[X_i X_j] = E[X_1 X_2] = P(X_1, X_2 = 1) = P(X_1 = 1, X_2 = 1)$$

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

LECTURE 8: Continuous random variables and probability density functions

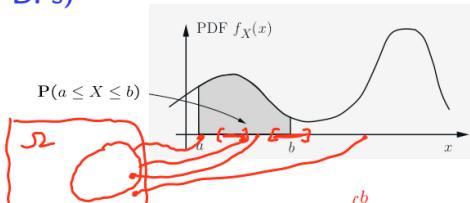
- Probability density functions
 - Properties
 - Examples
- Expectation and its properties
 - The expected value rule
 - Linearity
- Variance and its properties
- Uniform and exponential random variables
- Cumulative distribution functions
- Normal random variables
 - Expectation and variance
 - Linearity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

Probability density functions (PDFs)



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

$$p_X(x) \geq 0 \quad \sum_x p_X(x) = 1$$



$$\begin{aligned} P(a < X < b) &= \int_a^b f_X(x) dx \\ f_X(x) &\geq 0 \quad \int_{-\infty}^{\infty} f_X(x) dx = 1 \end{aligned}$$

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

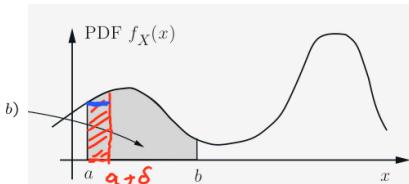
$$P(1 \leq X \leq 3 \text{ or } 4 \leq X \leq 5) = P(1 \leq X \leq 3) + P(4 \leq X \leq 5)$$

Probability density functions (PDFs)

$$\begin{aligned} \delta &> 0, \text{ small} \\ P(a < X < a+\delta) &\approx f_X(a) \cdot \delta \end{aligned}$$

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

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



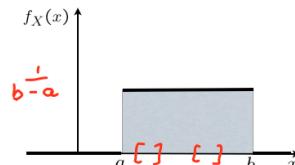
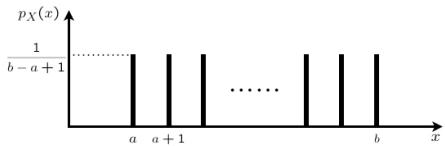
$$P(a < X < a+\delta) = \int_a^{a+\delta} f_X(x) dx$$

$$f_X(x) \geq 0$$

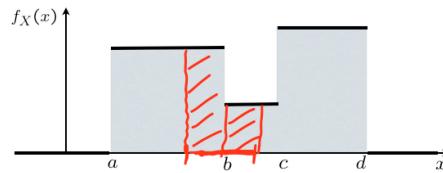
$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

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

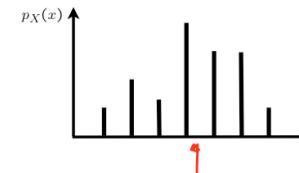
Example: continuous uniform PDF



- Generalization: piecewise constant PDF

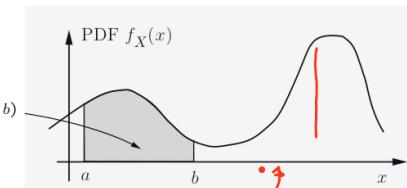


Expectation/mean of a continuous random variable



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

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



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

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

Properties of expectations

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

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

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

- Linearity

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

Variance and its properties

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

$$\mu = E[X]$$

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

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

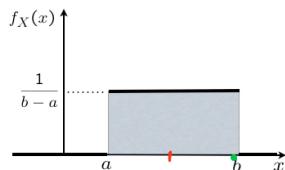
$$g(x) = (x - \mu)^2$$

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

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

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

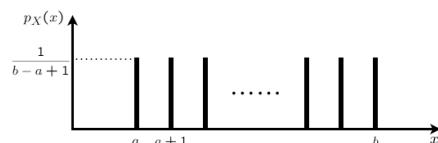
Continuous uniform random variable; parameters a, b



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

$$E[X^2] = \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \left(\frac{b^3}{3} - \frac{a^3}{3} \right) \quad \text{var}(X) = \frac{1}{12}(b-a)(b-a+2)$$

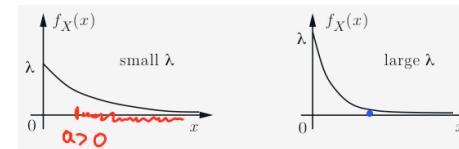
$$\text{var}(X) = E[X^2] - (E[X])^2 = \boxed{\frac{(b-a)^2}{12}} \quad \sigma = \frac{b-a}{\sqrt{12}}$$



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

Exponential random variable; parameter $\lambda > 0$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad \int f_X(x) dx = 1$$



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

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

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

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

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

$$\begin{aligned} P(X \geq a) &= \int_a^{\infty} \lambda e^{-\lambda x} dx \\ \int e^{ax} dx &= \frac{1}{a} e^{ax} \quad a \leftrightarrow -\lambda \\ &= \lambda \cdot \left(-\frac{1}{\lambda} \right) e^{-\lambda x} \Big|_{\infty}^a \\ &= -e^{-\lambda \cdot \infty} + e^{-\lambda a} = \boxed{e^{-\lambda a}} \end{aligned}$$

Cumulative distribution function (CDF)

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

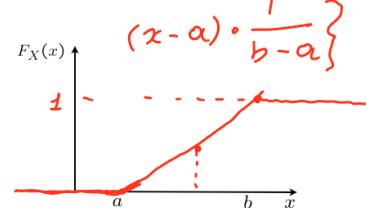
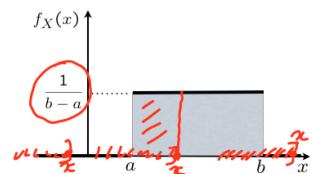
- Continuous random variables:

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

$$\text{area under } f_X(t) \text{ from } 3 \text{ to } 4$$

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

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

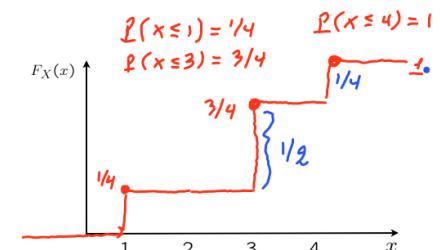
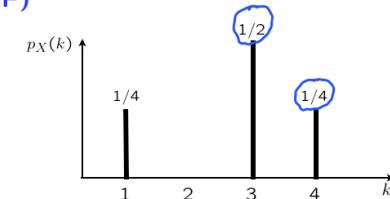


Cumulative distribution function (CDF)

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

- Discrete random variables:

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



General CDF properties

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



- Non-decreasing $\text{if } y \geq x \Rightarrow F_X(y) \geq F_X(x)$

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

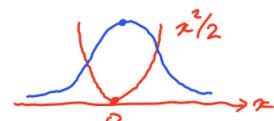
- $F_X(x)$ tends to 0, as $x \rightarrow -\infty$

Normal (Gaussian) random variables

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

Standard normal (Gaussian) random variables

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



calculus:

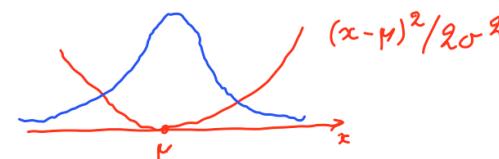
$$\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}$$

- $E[X] = 0$

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

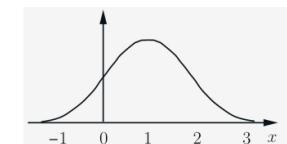
General normal (Gaussian) random variables

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



- $E[X] = \mu$

- $\text{var}(X) = \sigma^2$



Linear functions of a normal random variable

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

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

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

- Fact (will prove later in this course):

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

- Special case: $a = 0$:

$$Y = b \text{ discrete} \\ \sim N(b, 0)$$

Standard normal tables

- No closed form available for CDF

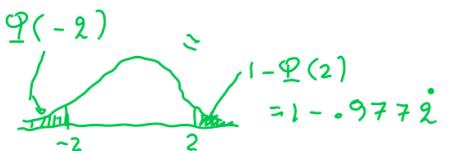
but have tables for the standard normal

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

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

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

$$\Phi(1.0) = 0.8770 \quad \Phi(2.0) = 0.9981$$



	.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 $\sigma^2 > 0$

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

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

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

Calculating normal probabilities

- Express an event of interest in terms of standard normal

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

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

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

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

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

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

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

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

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

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

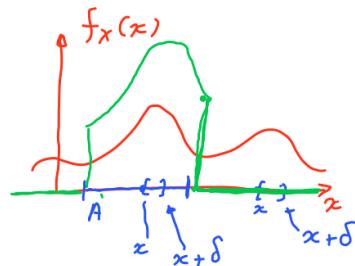
$$P(X \in B) = \sum_{x \in B} p_X(x) \quad P(X \in B) = \int_B f_X(x) dx$$

$$P(X \in B | A) = \sum_{x \in B} p_{X|A}(x) \quad \underline{P(X \in B | A) = \int_B f_{X|A}(x) dx} \quad \text{Def}$$

$$\sum_x p_{X|A}(x) = 1 \quad \int f_{X|A}(x) dx = 1$$

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

$$\begin{aligned} P(x \leq X \leq x + \delta | X \in A) &\approx f_{X|X \in A}(x) \cdot \delta \\ &= \frac{\underline{P(x \leq X \leq x + \delta, X \in A)}}{P(A)} \\ &= \frac{\underline{P(x \leq X \leq x + \delta)}}{P(A)} \approx \frac{f_X(x)}{P(A)} \end{aligned}$$



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

Conditional expectation of X , given an event

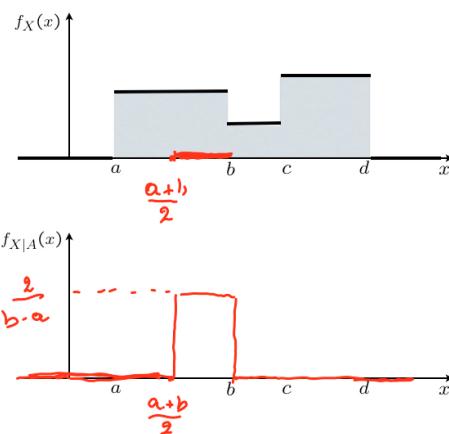
$$\begin{aligned} E[X] &= \sum_x x p_X(x) & E[X] &= \int x f_X(x) dx \\ E[X | A] &= \sum_x x p_{X|A}(x) & E[X | A] &= \int x f_{X|A}(x) dx \quad \text{Def} \end{aligned}$$

Expected value rule:

$$E[g(X)] = \sum_x g(x)p_X(x) \quad E[g(X)] = \int g(x)f_X(x) dx$$

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

Example



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

$$\begin{aligned} E[X | A] &= \frac{1}{2} \cdot \frac{a+b}{2} + \frac{1}{2} b \\ &= \frac{1}{4} a + \frac{3}{4} b \\ E[X^2 | A] &= \left. \frac{2}{b-a} \cdot x^2 dx \right|_{\frac{a+b}{2}}^b \\ &\approx \frac{a+b}{2} \end{aligned}$$

Memorylessness of the exponential PDF

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

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



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

— we are told that $T > t$

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

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

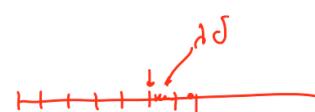
$$\begin{aligned} &= \frac{P(T-t > x, T > t)}{P(T > t)} = \frac{P(T > t+x, T > t)}{P(T > t)} = \frac{P(T > t+x)}{P(T > t)} \\ &= \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x} \end{aligned}$$

Memorylessness of the exponential PDF

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

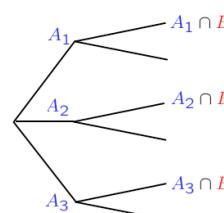
$$P(0 \leq T \leq \delta) \approx f_T(0) \cdot \delta = \lambda \delta$$

$$P(t \leq T \leq t+\delta | T > t) = \approx \lambda \delta$$



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

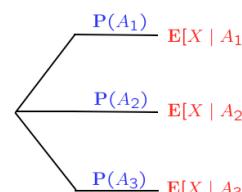
Total probability and expectation theorems



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

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

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



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

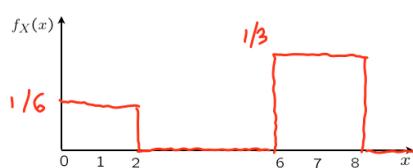
$$\int x f_X(x) dx = P(A_1) \int x f_{X|A_1}(x) dx + \dots$$

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

Example

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

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



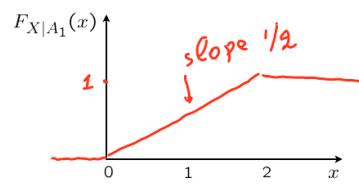
$$f_X(x) = \Pr(A_1)f_{X|A_1}(x) + \dots + \Pr(A_n)f_{X|A_n}(x)$$

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

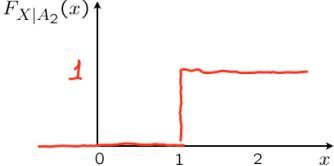
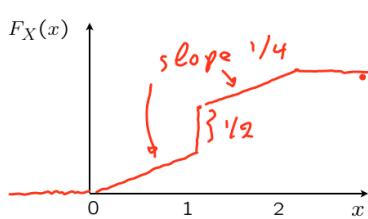
$$\frac{1}{3} \cdot 1 + \frac{2}{3} \cdot 7$$

Mixed distributions

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



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



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**

$$\Pr(X=1) = 1/2$$

X is mixed

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

$$Z \text{ continuous} \quad F_x(x) = p \cdot \Pr(Y \leq x) + (1-p) \Pr(Z \leq x) \\ = p F_Y(x) + (1-p) F_Z(x)$$

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

Jointly continuous r.v.'s and joint PDFs

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

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

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

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

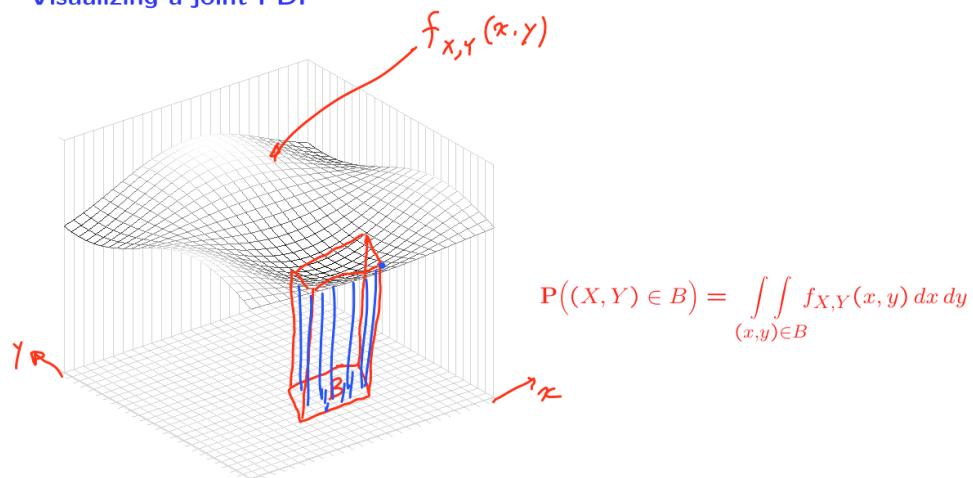
$$\Pr((X,Y) \in B) = \int \int_{(x,y) \in B} f_{X,Y}(x,y) dx dy \cdot$$

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

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

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

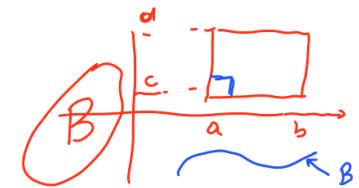
Visualizing a joint PDF



On joint PDFs

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

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

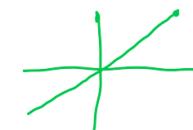


$P(a \leq X \leq a + \delta, c \leq Y \leq c + \delta) \approx f_{X,Y}(a,c) \cdot \delta^2$

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

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

$$Y = X$$



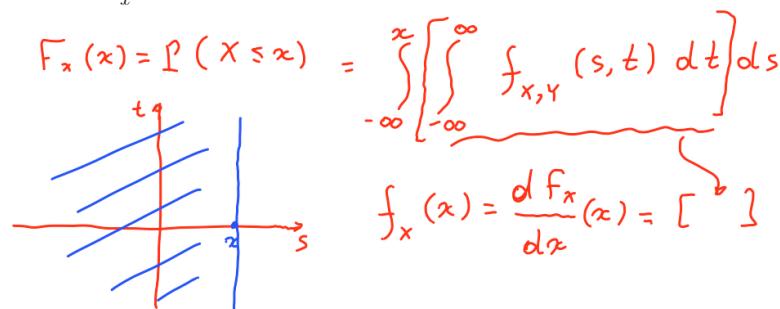
From the joint to the marginals

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

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

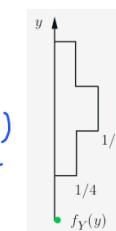
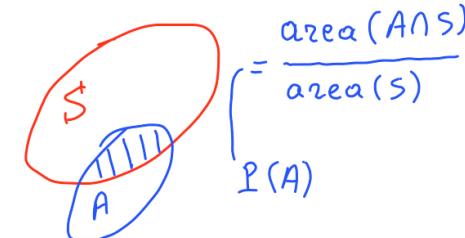
$$p_Y(y) = \sum_x p_{X,Y}(x,y)$$

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

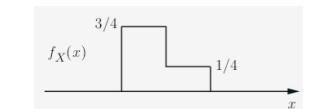
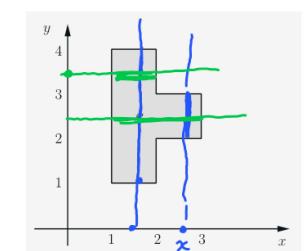


Uniform joint PDF on a set S

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



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



More than two random variables

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

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

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

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

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

Functions of multiple random variables

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

Expected value rule:

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

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

Linearity of expectations

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

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

$$E[X_1 + \dots + X_n] = E[X_1] + \dots + E[X_n]$$

The joint CDF

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

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

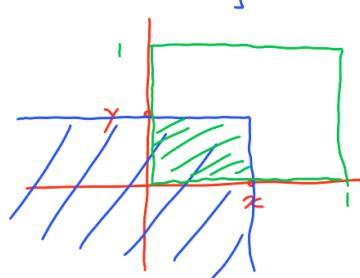
$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

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

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

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

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



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.

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

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

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

$\downarrow Y=y$ $\downarrow Y \approx y$

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

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

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

Comments on conditional PDFs

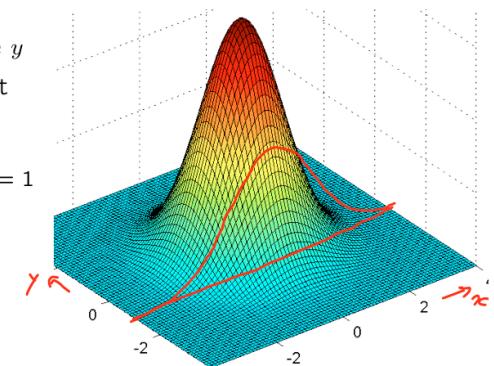
$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \bullet \quad f_{X|Y}(x|y) \geq 0$$

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

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

- Multiplication rule:

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



Total probability and expectation theorems

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

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

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

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

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

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

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

- Expected value rule...

$$E[g(x)|Y=y]$$

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

Independence

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

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

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

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

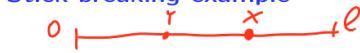
- equivalent to: $f_{X|Y}(x|y) = f_X(x)$, for all y with $f_Y(y) > 0$ and all x

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

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

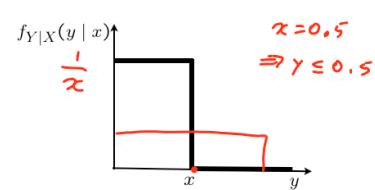
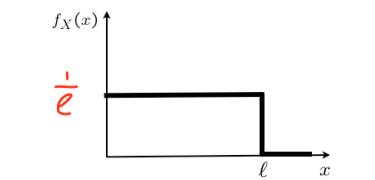
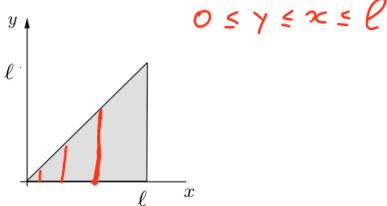
$$g(X) \text{ and } h(Y) \text{ are also independent: } E[g(X)h(Y)] = E[g(X)] \cdot E[h(Y)]$$

Stick-breaking example



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

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



Stick-breaking example

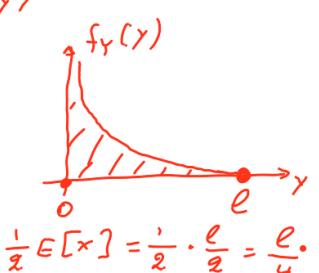
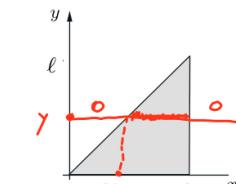
$$f_{X,Y}(x,y) = \frac{1}{\ell x}, \quad 0 \leq y \leq x \leq \ell$$

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

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

- Using total expectation theorem:

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

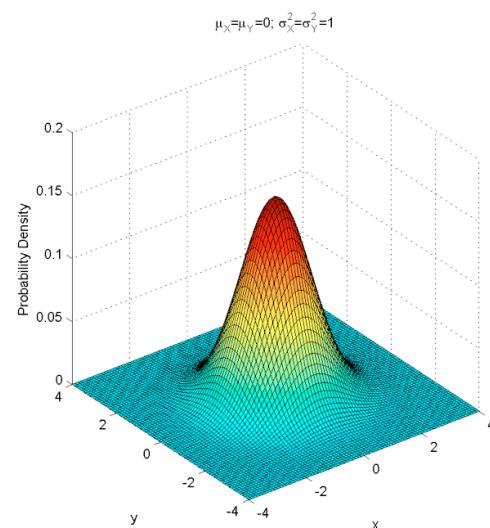
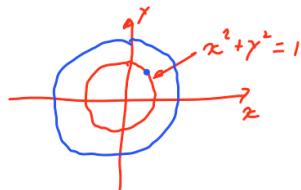


Independent standard normals

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

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

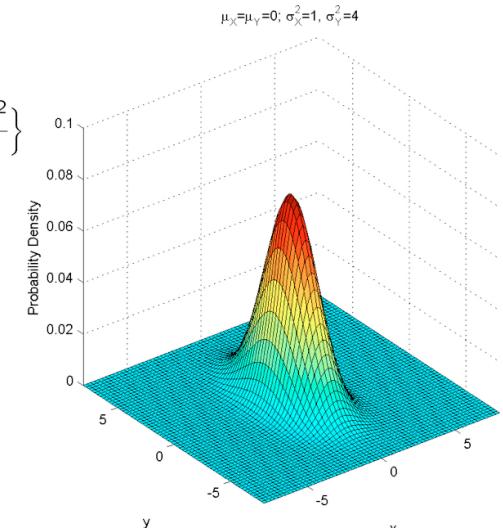
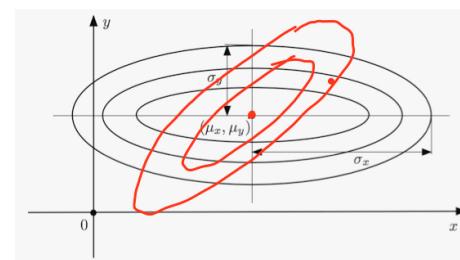
$$= \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(x^2 + y^2)\right\}$$



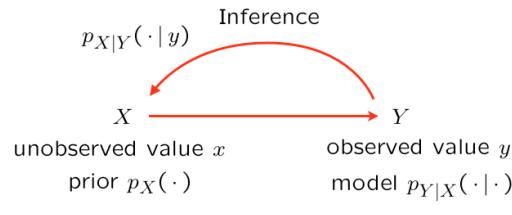
Independent normals

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

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



The Bayes rule — a theme with variations



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

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

posterior

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

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

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

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

The Bayes rule — discrete unknown, continuous measurement

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

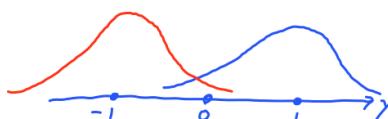
- Probability that $K = 1$, given that $Y = y$?

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

$k = -1, +1$

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

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



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

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

The Bayes rule — one discrete and one continuous random variable

K : discrete Y : continuous

$$\Pr(K=k, y \leq Y \leq y+\delta) \quad \delta > 0, \delta \approx 0$$

$$= \Pr(K=k) \Pr(y \leq Y \leq y+\delta | K=k) \approx p_K(k) f_{Y|K}(y|k) \delta$$

$$= \Pr(y \leq Y \leq y+\delta) \Pr(K=k | y \leq Y \leq y+\delta) \approx f_Y(y) \delta p_{K|Y}(k|y)$$

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

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

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

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

The Bayes rule — continuous unknown, discrete measurement

- measurement K : Bernoulli with parameter Y

$$K \begin{cases} 1 \\ 0 \end{cases} \quad Y \begin{cases} 1 \\ 0 \end{cases}$$

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

$$p_K(k) = \frac{1}{2} \quad f_{Y|K}(y|k) = \begin{cases} 1 & y \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

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

$$f_{Y|K}(y|1) = \frac{1+y}{1+e^{-2y}} = 2y, y \in [0,1]$$

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

The Bayes rule — continuous unknown, discrete measurement

- measurement K : Bernoulli with parameter Y

$$K \begin{cases} 1 \\ 0 \end{cases} \quad Y \begin{cases} 1 \\ 0 \end{cases}$$

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

$$f_{Y|K}(y|k) = \begin{cases} 1 & y \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

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

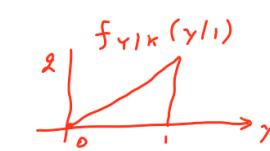
$$f_{Y|K}(y|1) = \frac{1+y}{1+e^{-2y}} = 2y, y \in [0,1]$$

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

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

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

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

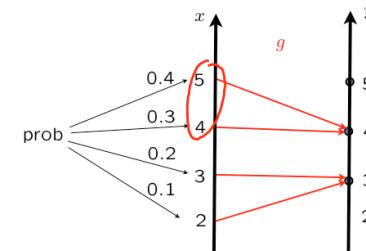


LECTURE 11: Derived distributions

- 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

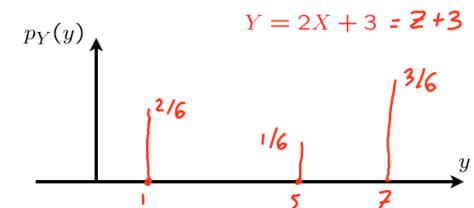
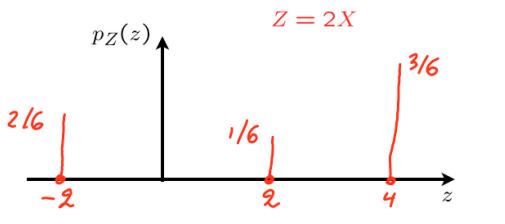
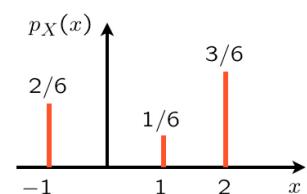
$$Y = g(X)$$



$$\begin{aligned} p_Y(4) &= P(Y=4) \\ &= P(X=4) + P(X=5) \\ &= p_X(4) + p_X(5) = 0.3 + 0.4 \end{aligned}$$

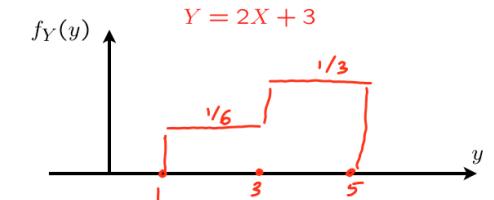
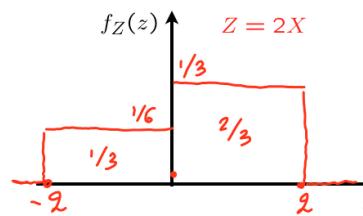
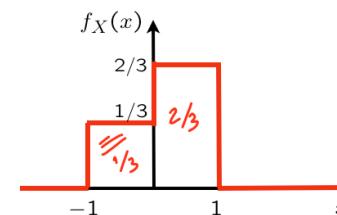
$$\begin{aligned} p_Y(y) &= P(g(X) = y) \\ &= \sum_{x: g(x)=y} p_X(x) \end{aligned}$$

A linear function of a discrete r.v.



$$\begin{aligned} Y = 2X + 3 &= Z + 3 \\ p_Y(y) &= P(Y=y) = P(Z+3=y) \\ &= P\left(Z=\frac{y-3}{2}\right) = p_Z\left(\frac{y-3}{2}\right) \\ Y = aX + b &: p_Y(y) = p_X\left(\frac{y-b}{a}\right) \end{aligned}$$

A linear function of a continuous r.v.



$$Y = 2X + 3$$

A linear function of a continuous r.v.

$$a > 0$$

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

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

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

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

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

$$Y = aX + b$$

$$a < 0$$

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

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

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

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

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

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

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

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

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

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

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

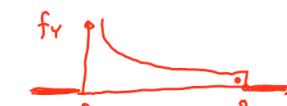
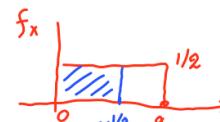
If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

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

- Two-step procedure:

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

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



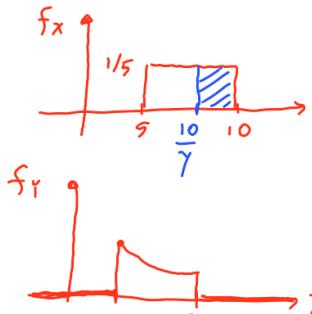
$$0 \leq y \leq 8$$

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

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

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.



$$\text{time} = Y = \frac{10}{X} \quad 0 \leq Y \leq 2$$

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

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

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

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

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

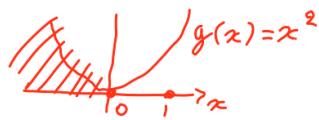
Example: $Y = X^2$; X uniform on $[0, 1]$

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

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

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

$$0 \leq y \leq 1$$



A general formula for the PDF of $Y = g(X)$ when g is monotonic $x^3 \frac{\alpha}{x}$

Assume g strictly increasing

and differentiable

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

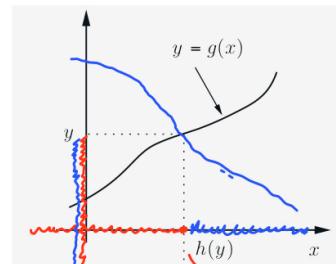
$$f_Y(y) = f_X(h(y)) \left| \frac{dh(y)}{dy} \right|$$

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

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

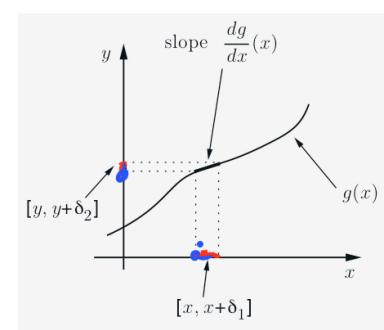
$$f_Y(y) = f_X(h(y)) \left| \frac{dh(y)}{dy} \right|$$

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



inverse function h decreasing

An intuitive explanation for the monotonic case



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

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

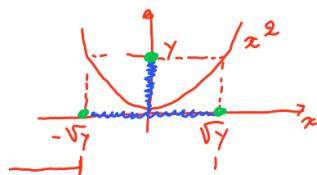
$$f_Y(y) \approx P(y \leq Y \leq y + \delta_2) = P(x \leq X \leq x + \delta_1)$$

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

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

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

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: $y \geq 0$

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) = P(|X| \leq \sqrt{y}) = P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \quad f_Y(y) = f_X(\sqrt{y}) \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \frac{1}{2\sqrt{y}} \end{aligned}$$

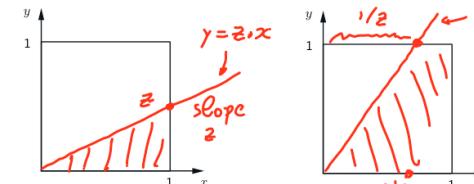
LECTURE 12: Sums of independent random variables; Covariance and correlation

- 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

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

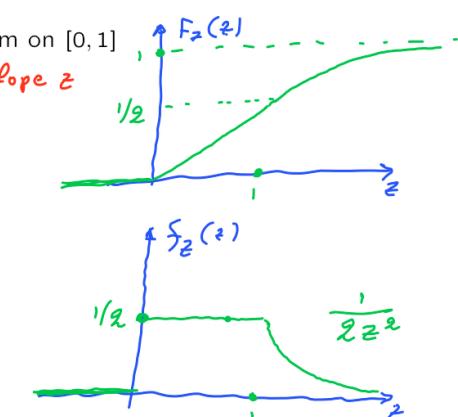
- Same methodology: find CDF of Z

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



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

$Y = g(x)$

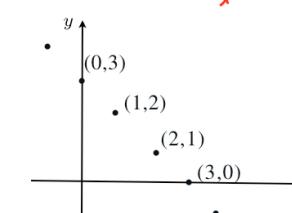
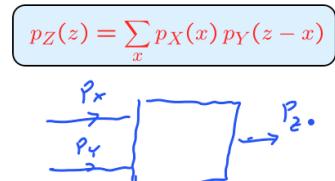


The distribution of $X + Y$: the discrete case

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

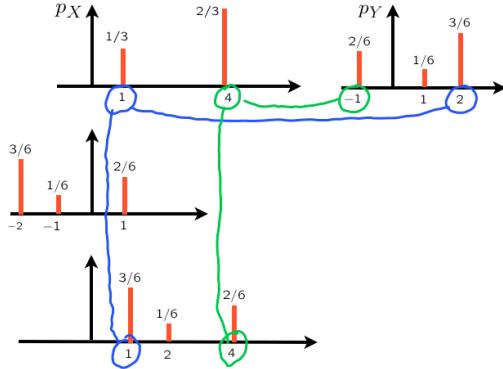
$p_{X,Y}(x,y)$ known PMFs

$$\begin{aligned} p_Z(3) &= \dots + P(X=0, Y=3) + P(X=1, Y=2) + \dots \\ &= \dots + p_X(0)p_Y(3) + p_X(1)p_Y(2) + \dots \end{aligned}$$



$$\begin{aligned} p_Z(z) &= \sum_x P(X=x, Y=z-x) \\ &= \sum_x p_X(x)p_Y(z-x) \end{aligned}$$

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 sum of independent normal r.v.'s

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

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

$$Z = X + Y$$

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

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

$$\underline{X+Y} + \underline{W}$$

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

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

The sum of finitely many independent normals is normal

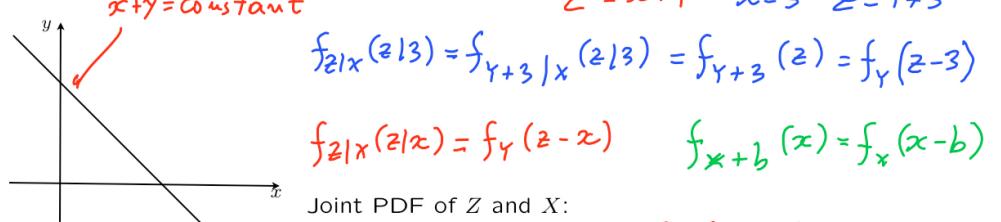
The distribution of $X + Y$: the continuous case

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

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

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

$x+y=\text{constant}$



$$\text{Conditional on } X=x: Z = x+Y \quad x=3 \quad Z=3+Y$$

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

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

Joint PDF of Z and X :

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

From joint to the marginal: $f_Z(z) = \int_{-\infty}^{\infty} f_{X,Z}(x,z) dx$

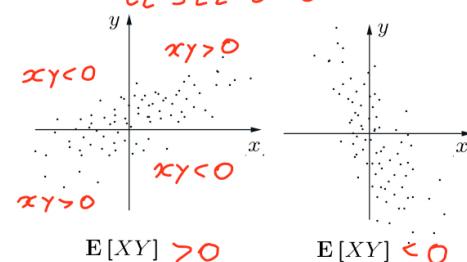
- Same mechanics as in discrete case (flip, shift, etc.)

Covariance

- Zero-mean, discrete X and Y

- if independent: $E[XY] =$

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

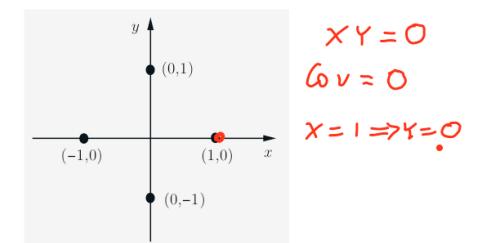


Definition for general case:

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

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

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



Covariance properties

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

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

$$\text{cov}(aX + b, Y) =$$

(assume 0 means)

$$= E[(ax+b)y] = aE[xy] + bE[y]$$

$$= a \cdot \text{cov}(x, y)$$

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

$$= E[XY] + E[XZ] = \text{cov}(x, y) + \text{cov}(x, z)$$

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

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

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

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

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

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

The variance of a sum of random variables

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

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

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

$$+ 2(X_1 - E[X_1])(X_2 - E[X_2])]$$

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

The variance of a sum of random variables

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

$$\text{var}(X_1 + \dots + X_n) = E[(X_1 + \dots + X_n)^2]$$

(assume 0 means)

$$= E\left[\sum_{i=1}^n X_i^2 + \sum_{\substack{i=1, \dots, n \\ j=1, \dots, n \\ i \neq j}} X_i X_j\right]$$

$$= \sum_i \text{var}(X_i) + \sum_{i \neq j} \text{cov}(X_i, X_j)$$

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

The Correlation coefficient

- Dimensionless version of covariance:

$$-1 \leq \rho \leq 1$$

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

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

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

Proof of key properties of the correlation coefficient

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

$$-1 \leq \rho \leq 1$$

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

$$\begin{aligned} E[(X - \rho Y)^2] &= E[X^2] - 2\rho E[XY] + \rho^2 E[Y^2] \\ 0 &\leq = 1 - 2\rho^2 + \rho^2 = \underline{\underline{1 - \rho^2}} \quad 1 - \rho^2 \geq 0 \Rightarrow \rho^2 \leq 1 \end{aligned}$$

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

Interpreting the correlation coefficient

- Association does not imply causation or influence

X : math aptitude

Y : musical ability

- Correlation often reflects underlying, common, hidden factor

- Assume, Z, V, W are independent

$$X = \underline{\underline{Z}} + V \quad Y = \underline{\underline{Z}} + W$$

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

$$\begin{aligned} \text{var}(x) &= \text{var}(z) + \text{var}(v) = 2 \Rightarrow \sigma_x = \sqrt{2} \quad \sigma_y = \sqrt{2} \\ \text{cov}(x, y) &= E[(z+v)(z+w)] = E[z^2] + E[vz] + E[zw] + E[vw] \\ &= 1 + 0 + 0 + 0 \end{aligned}$$

Correlations matter...

- A real-estate investment company invests \$10M in each of 10 states. At each state i , the return on its investment is a random variable X_i , with mean 1 and standard deviation 1.3 (in millions).

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

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

- If the X_i are uncorrelated, then:

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

- If for $i \neq j$, $\rho(X_i, X_j) = 0.9$: $\text{cov}(X_i, X_j) = \rho \sigma_{X_i} \sigma_{X_j} = 0.9 \times 1.3 \times 1.3 = 1.52$

$$\text{var}(X_1 + \dots + X_{10}) = 10 \cdot (1.3)^2 + 90 \cdot 1.52 = 154$$

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

LECTURE 13: Conditional expectation and variance revisited;

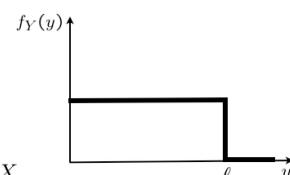
Application: Sum of a random number of independent r.v.'s

- 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 $h(X)$?
 $\not=h^2$
- $h(X)$ is the r.v. that takes the value x^2 , if X happens to take the value x
- $\underline{g(y)} = E[X | Y = y] = \sum_x x p_{X|Y}(x | y)$
(integral in continuous case)
- $\underline{g(Y)}$: is the r.v. that takes the value $E[X | Y = y]$, if Y happens to take the value y
- Remarks:
 - It is a function of Y
 - It is a random variable
 - Has a distribution, mean, variance, etc.

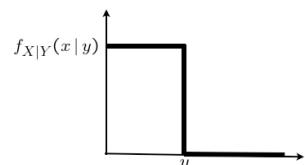
Definition: $\underline{E[X|Y]} = g(Y)$



Stick-breaking example

- Stick example: stick of length ℓ break at uniformly chosen point Y
break what is left at uniformly chosen point X
- $E[X | Y = y] = \underline{y/2}$
- $E[X | Y] = \underline{y/2}$

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



The mean of $E[X | Y]$: Law of iterated expectations

- $\underline{g(y)} = E[X | Y = y]$

$$\boxed{E[E[X | Y]] = E[X]}$$

$$E[x|Y] \stackrel{\Delta}{=} g(Y)$$

$$\underline{E[E[X|Y]]} = E[g(Y)]$$

$$= \sum_y g(y) P_Y(y) \quad \text{exp. value rule}$$

$$= \sum_y E[x|Y=y] P_Y(y)$$

$$= E[x] \quad \bullet \text{total exp thm}$$

Forecast revisions

$$\boxed{E[E[X | Y]] = E[X]}$$

- Suppose forecasts are made by calculating expected value, given any available information



- X : February sales

- Forecast in the beginning of the year: $E[X]$

- End of January: will get new information, value y of Y

Revised forecast: $E[X|Y=y] \quad E[x|Y]$

- Law of iterated expectations:

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

The conditional variance as a random variable

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

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

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

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

$$\text{var}(X | Y = y) = \frac{y^2}{12}$$

$$\text{var}(X | Y) = \frac{Y^2}{12}$$

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

Derivation of the law of total variance

$$\text{var}(X) = E[\text{var}(X | Y)] + \text{var}(E[X | Y])$$

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

$$\text{var}(X | Y = y) = E[X^2 | Y = y] - (E[X | Y = y])^2 \text{ for all } y$$

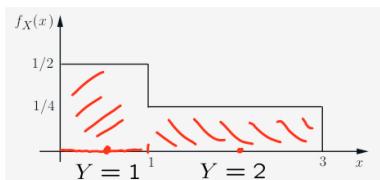
$$\text{var}(X | Y) = E[X^2 | Y] - (E[X | Y])^2$$

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

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

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

A simple example



$$\text{var}(X) = E[\text{var}(X | Y)] + \text{var}(E[X | Y]) = \frac{37}{48}$$

$$= \frac{5}{24} + \frac{9}{16}$$

$$\text{var}(X | Y) = \frac{1/2}{1/2} \text{ var}(X | Y = 1) = \frac{1}{12}$$

$$\frac{1/2}{1/2} \text{ var}(X | Y = 2) = \frac{2^2/12}{1/2} = \frac{4}{12}$$

$$E[\text{var}(X | Y)] = \frac{1}{2} \cdot \frac{1}{12} + \frac{1}{2} \cdot \frac{4}{12} = \frac{5}{24}$$

$$E[X | Y] = \frac{1/2}{1/2} E[X | Y = 1] = \frac{1}{2}$$

$$\frac{1/2}{1/2} E[X | Y = 2] = 2$$

$$E[E[X | Y]] = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot 2 = \frac{5}{4} = E[X]$$

$$\text{var}(E[X | Y]) = \frac{1}{2} \left(\frac{1}{2} - \frac{5}{4} \right)^2$$

$$+ \frac{1}{2} \left(2 - \frac{5}{4} \right)^2 = \frac{9}{16}$$

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: $y = 1: \frac{1}{10} \sum_{i=1}^{10} x_i = 90 \quad y = 2: \frac{1}{20} \sum_{i=11}^{30} x_i = 60$

- $E[X] = \frac{1}{30} \sum_{i=1}^{30} x_i = \frac{1}{30} (90 \cdot 10 + 60 \cdot 20) = 70$

$$E[X | Y = 1] = 90$$

$$E[X | Y] = \frac{1/3 \cdot 90 + 2/3 \cdot 60}{3} = 70$$

$$E[X | Y = 2] = 60$$

- $E[E[X | Y]] = \frac{1}{3} \cdot 90 + \frac{2}{3} \cdot 60 = 70$

Section means and variances (ctd.)

$$E[X | Y] = \begin{cases} 90, & \text{w.p. } 1/3 \\ 60, & \text{w.p. } 2/3 \end{cases}$$

$$E[E[X | Y]] = 70 = E[X]$$

$$\text{var}(E[X | Y]) = \frac{1}{3}(90 - 70)^2 + \frac{2}{3}(60 - 70)^2 = 200$$

- More data: $\frac{1}{10} \sum_{i=1}^{10} (x_i - 90)^2 = 10$ $\frac{1}{20} \sum_{i=11}^{30} (x_i - 60)^2 = 20$

$$\text{var}(X | Y = 1) = 10$$

$$\text{var}(X | Y) = \frac{\cancel{1/3} \cdot 10}{\cancel{2/3} \cdot 20}$$

$$\text{var}(X | Y = 2) = 20$$

$$E[\text{var}(X | Y)] = \frac{1}{3} \cdot 10 + \frac{2}{3} \cdot 20 = \frac{50}{3}$$

$$\text{var}(X) = E[\text{var}(X | Y)] + \text{var}(E[X | Y]) = \frac{50}{3} + 200$$

$\text{var}(X)$ = (average variability **within** sections) + (variability **between** sections)

Sum of a random number of independent r.v.'s

$$E[Y] = E[N] \cdot E[X]$$

- N : number of stores visited (N is a nonnegative integer r.v.)
- X_i : money spent in store i
- X_i independent, identically distributed
- Let $Y = X_1 + \dots + X_N$
- independent of N

$$\begin{aligned} E[Y | N = n] &= E[X_1 + \dots + X_n | N = n] = E[X_1 + \dots + X_n | N = n] \\ &\stackrel{?}{=} E[Y|N] = NE[x] \\ &= E[X_1 + \dots + X_n] = n E[x] \end{aligned}$$

- Total expectation theorem:

$$E[Y] = \sum_n p_N(n) E[Y | N = n] = \underbrace{\sum_n p_n(n)}_{n} n E[x] = E[n] E[x]$$

- Law of iterated expectations:

$$E[Y] = E[E[Y | N]] = E[NE[x]] = E[n] E[x]$$

Variance of sum of a random number of independent r.v.'s

$$Y = X_1 + \dots + X_N$$

- $\text{var}(Y) = E[\text{var}(Y | N)] + \text{var}(E[Y | N])$

- $E[Y | N] = N E[X]$

- $\text{var}(E[Y | N]) = \text{var}(NE[x]) = (E[x])^2 \text{var}(N)$

- $\text{var}(Y | N = n) = \text{var}(X_1 + \dots + X_n | N = n) = \text{var}(X_1 + \dots + X_n) = n \text{var}(x)$

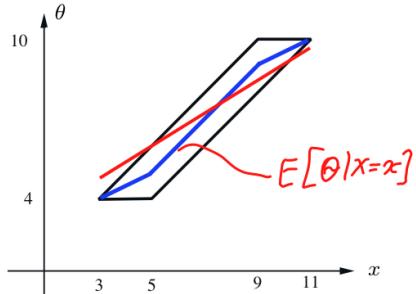
- $E[\text{var}(Y | N)] = E[N \text{var}(x)] = E[N] \text{var}(x)$

LECTURE 17: Linear least mean squares (LLMS) estimation

- Conditional expectation $E[\Theta | X]$ may be hard to compute/implement
- Restrict to estimators $\widehat{\Theta} = aX + b$
 - minimize mean squared error
- Simple solution
- Mathematical properties
- Example

LLMS formulation

- Unknown Θ ; observation X
- Minimize $E[(\hat{\Theta} - \Theta)^2]$
- Estimators $\hat{\Theta} = g(X) \rightarrow \hat{\Theta}_{LLMS} = E[\Theta | X]$
- Consider estimators of Θ , of the form $\hat{\Theta} = aX + b$
- Minimize $E[(\Theta - aX - b)^2]$, w.r.t. a, b
- If $E[\Theta | X]$ is linear in X , then $\hat{\Theta}_{LLMS} = \hat{\Theta}_{LMS}$



Solution to the LLMS problem

- Minimize $E[(\Theta - aX - b)^2]$, w.r.t. a, b

– suppose a has already been found: $b = E[\Theta] - aE[X]$

$$\min E[(\Theta - aX - E[\Theta] + aE[X])^2] = \text{var}(\Theta - aX)$$

$$= \text{var}(\Theta) + a^2 \text{var}(X) - 2a \text{cov}(\Theta, X)$$

$$\frac{d}{da} = 0 : 2a \text{var}(X) - 2 \text{cov}(\Theta, X) = 0 \quad \left| \begin{array}{l} p = \frac{\text{cov}(\Theta, X)}{\sigma_\Theta \sigma_X} \\ a = \frac{p \sigma_\Theta \sigma_X}{\sigma_X^2} \end{array} \right.$$

$$\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X]) = E[\Theta] + \rho \frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

Remarks on the solution and on the error variance

$$\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X]) = E[\Theta] + \rho \frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

• Only means, variances, covariances matter

$$\bullet \rho > 0: X > E[X] \Rightarrow \hat{\Theta}_L > E[\Theta]$$

$$\bullet \rho = 0: \hat{\Theta}_L = E[\Theta]$$

$$\mathbb{E}[(\hat{\Theta}_L - \Theta)^2] = (1 - \rho^2) \text{var}(\Theta)$$

$$\mathbb{E}[(\Theta - \rho \frac{\sigma_\Theta}{\sigma_X} X)^2] = \sigma_\Theta^2 - 2\rho \frac{\sigma_\Theta}{\sigma_X} \rho \sigma_\Theta \sigma_X + \rho^2 \frac{\sigma_\Theta^2}{\sigma_X^2}$$

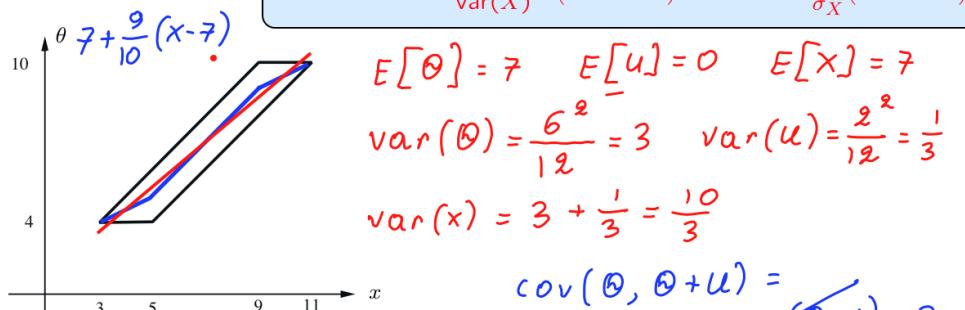
assume $E[\Theta] = E[X] = 0$

$$|\rho| = 1$$

$$\hat{\Theta}_L = \Theta$$

Example

$$\hat{\Theta}_L = E[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - E[X]) = E[\Theta] + \rho \frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$



$$\Theta: \text{uniform}[4, 10]$$

$$X = \Theta + U \quad \text{uniform}[-1, 1]$$

$$\Theta, U \text{ independent}$$

$$\begin{aligned} \text{cov}(\Theta, \Theta + U) &= \\ &= \text{cov}(\Theta, \Theta) + \text{cov}(\Theta, U) = 3 \end{aligned}$$

$$E[\Theta] = 7 \quad E[U] = 0 \quad E[X] = 7$$

$$\text{var}(\Theta) = \frac{6^2}{12} = 3 \quad \text{var}(U) = \frac{2^2}{12} = \frac{1}{3}$$

$$\text{var}(X) = 3 + \frac{1}{3} = \frac{10}{3}$$

LLMS for inferring the parameter of a coin

- Standard example:
 - coin with bias Θ ; prior $f_\Theta(\cdot)$
 - fix n ; $X = \text{number of heads}$
- Assume $f_\Theta(\cdot)$ is uniform in $[0, 1]$

$$\hat{\Theta}_{\text{LLMS}} = \frac{X+1}{n+2} = \hat{\Theta}_{\text{LLMS}}$$

$$\hat{\Theta}_{\text{LLMS}} = \mathbb{E}[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - \mathbb{E}[X])$$

LLMS for inferring the parameter of a coin

$$\hat{\Theta}_{\text{LLMS}} = \mathbb{E}[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}(X - \mathbb{E}[X])$$

$$\text{cov}(\Theta, X) = \frac{n}{12} \quad \text{var}(X) = \frac{n(n+2)}{12} \quad \mathbb{E}[X] = \frac{n}{2}$$

$$\hat{\Theta}_{\text{LLMS}} = \frac{X+1}{n+2} = \hat{\Theta}_{\text{LLMS}}$$

LLMS for inferring the parameter of a coin

- Θ : uniform on $[0, 1]$ $\mathbb{E}[\Theta] = \frac{1}{2}$ $\text{var}(\Theta) = \frac{1}{12}$ $\mathbb{E}[\Theta^2] = \frac{1}{12} + \frac{1}{2^2} = \frac{1}{3}$
- $p_{X|\Theta}$: $\text{Bin}(n, \Theta)$ $\mathbb{E}[X | \Theta] = n\Theta$ $\text{var}(X | \Theta) = n\Theta(1 - \Theta)$
- $\mathbb{E}[X] = \mathbb{E}[n\Theta] = n/2$ $\mathbb{E}[X^2 | \Theta] = n\Theta(1-\Theta) + n^2\Theta^2$
- $\mathbb{E}[X^2] = \mathbb{E}[\mathbb{E}[X^2 | \Theta]] = \mathbb{E}[n\Theta + (n^2-n)\Theta^2] = \frac{n}{2} + \frac{n^2-n}{3} = \frac{n}{6} + \frac{n^2}{3}$
- $\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{n}{6} + \frac{n^2}{3} - \frac{n^2}{4} = \frac{n}{6} + \frac{n^2}{12} = \frac{n(n+2)}{12}$
- $\mathbb{E}[\Theta X | \Theta] = \Theta \mathbb{E}[X | \Theta] = n\Theta^2$
- $\mathbb{E}[\Theta X] = \mathbb{E}[\mathbb{E}[\Theta X | \Theta]] = \mathbb{E}[n\Theta^2] = n/3$
- $\text{cov}(\Theta, X) = \mathbb{E}[\Theta X] - \mathbb{E}[\Theta]\mathbb{E}[X] = \frac{n}{3} - \frac{n}{4} = \frac{n}{12}$

LLMS with multiple observations

- Unknown Θ ; observations $X = (X_1, \dots, X_n)$
- Consider estimators of the form: $\hat{\Theta} = a_1X_1 + \dots + a_nX_n + b$
- Find best choices of a_1, \dots, a_n, b
 $\text{minimize: } \mathbb{E}[(a_1X_1 + \dots + a_nX_n + b - \Theta)^2] = a_1^2\mathbb{E}[X_1^2] + 2a_1a_2\mathbb{E}[X_1X_2] + \dots + a_n^2\mathbb{E}[X_n^2] + \dots + 2a_1b\mathbb{E}[X_1] + \dots + 2a_nb\mathbb{E}[X_n] + nb^2$
- If $\mathbb{E}[\Theta | X]$ is linear in X , then $\hat{\Theta}_{\text{LMS}} = \hat{\Theta}_{\text{LLMS}}$
- 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

$$\begin{aligned} X_1 &= \Theta + W_1 & \Theta &\sim x_0, \sigma_0^2 & W_i &\sim 0, \sigma_i^2 \\ &\vdots \\ X_n &= \Theta + W_n & \Theta, W_1, \dots, W_n &\text{uncorrelated} \end{aligned}$$

- Suppose Θ, W_1, \dots, W_n are independent normal

$$\hat{\theta}_{\text{LMS}} = \mathbb{E}[\Theta | X = x] = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}} \quad \widehat{\Theta}_{\text{LMS}} = \mathbb{E}[\Theta | X] = \frac{\frac{x_0}{\sigma_0^2} + \sum_{i=1}^n \frac{X_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}} = \widehat{\Theta}_{\text{LLMS}}$$

- Suppose general (not normal) distributions, but same means, variances, as in normal example
 - all covariances also the same
 - solution must be the same

The representation of the data matters in LLMS

- Estimation based on X versus X^3
 - LMS: $\underline{\mathbb{E}[\Theta | X]}$ is the same as $\underline{\mathbb{E}[\Theta | X^3]}$
 - LLMS is different: estimator $\widehat{\Theta} = aX + b$ versus $\widehat{\Theta} = aX^3 + b$
- $\text{cov}(\Theta, X^3)$ $\text{var}(X^3)$
- can also consider $\widehat{\Theta} = a_1 \widehat{X} + a_2 \widehat{X^2} + a_3 \widehat{X^3} + b$
 - 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 $P(X \geq a)$ based on limited information about a distribution
 - Markov inequality (based on the mean)
 - Chebyshev inequality (based on the mean and variance)
- WLLN: X, X_1, \dots, X_n i.i.d.

$$\frac{X_1 + \dots + X_n}{n} \rightarrow \mathbb{E}[X]$$
 - 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 $\mathbb{E}[X]$ is small, then X is unlikely to be very large"

Markov inequality: If $X \geq 0$ and $a > 0$, then $P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$.

$$Y = \begin{cases} 0, & \text{if } X < a \\ a, & \text{if } X \geq a \end{cases} \quad \text{and } P(X \geq a) = E[Y] \leq E[X]$$

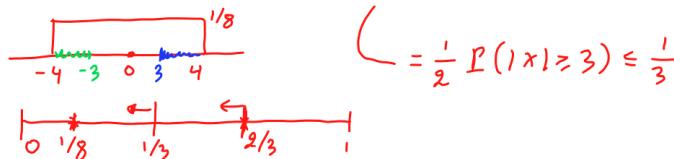
The Markov inequality

Markov inequality: If $X \geq 0$ and $a > 0$, then $P(X \geq a) \leq \frac{E[X]}{a}$

- Example: X is Exponential($\lambda = 1$): $P(X \geq a) \leq \frac{1}{a}$



- Example: X is Uniform[-4, 4]: $P(X \geq 3) \leq P(|X| \geq 3) \leq \frac{E[|X|]}{3} = \frac{2}{3}$



The Chebyshev inequality

- Random variable X , with finite mean μ and variance σ^2
- "If the variance is small, then X is unlikely to be too far from the mean"

Chebyshev inequality: $P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$

Markov inequality: If $X \geq 0$ and $a > 0$, then $P(X \geq a) \leq \frac{E[X]}{a}$

$$P(|X - \mu| \geq c) = P((X - \mu)^2 \geq c^2) \leq \frac{E[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$$

The Chebyshev inequality

Chebyshev inequality: $P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$

$$P(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2} \quad k=3 \quad \leq \frac{1}{9}$$

- Example: X is Exponential($\lambda = 1$): $P(X \geq a) \leq \frac{1}{a}$ (Markov)



$$P(X \geq a) = P(X - 1 \geq a - 1) \leq P(|X - 1| \geq a - 1) \leq \frac{1}{(a-1)^2} \sim \frac{1}{a^2}$$

The Weak Law of Large Numbers (WLLN)

- X_1, X_2, \dots i.i.d.; finite mean μ and variance σ^2

Sample mean: $M_n = \frac{X_1 + \dots + X_n}{n}$ $\mu = E[X_i]$

$$E[M_n] = \frac{E[X_1 + \dots + X_n]}{n} = \frac{n\mu}{n} = \mu$$

$$\text{Var}(M_n) = \frac{\text{Var}(X_1 + \dots + X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$P(|M_n - \mu| \geq \epsilon) \leq \frac{\text{Var}(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \xrightarrow[n \rightarrow \infty]{\text{f/x real } \epsilon > 0} 0$$

WLLN: For $\epsilon > 0$, $P(|M_n - \mu| \geq \epsilon) = P\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| \geq \epsilon\right) \rightarrow 0$, as $n \rightarrow \infty$

Interpreting the WLLN

$$M_n = (X_1 + \dots + X_n)/n$$

WLLN: For $\epsilon > 0$, $P(|M_n - \mu| \geq \epsilon) = P\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| \geq \epsilon\right) \rightarrow 0$, as $n \rightarrow \infty$

- One experiment
 - many measurements $X_i = \mu + W_i$
 - W_i : measurement noise; $E[W_i] = 0$; independent W_i
 - sample mean M_n is unlikely to be far off from true mean μ
- Many independent repetitions of the same experiment
 - event A , with $p = P(A)$
 - X_i : indicator of event A
 - the sample mean M_n is the empirical frequency of event A

$$X_i = 1, \text{ if } A \text{ occurs}$$

o.o.w.

$$E[X_i] = p$$

Convergence "in probability"

WLLN: For any $\epsilon > 0$, $P(|M_n - \mu| \geq \epsilon) \rightarrow 0$, as $n \rightarrow \infty$

- 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

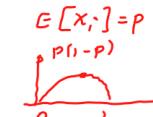
$$M_n \xrightarrow[n \rightarrow \infty]{i.p.} \mu$$

Definition: A sequence Y_n converges in probability to a number a if:

$$\text{for any } \epsilon > 0, \lim_{n \rightarrow \infty} P(|Y_n - a| \geq \epsilon) = 0$$

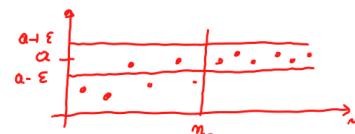
The pollster's problem

- p : fraction of population that will vote "yes" in a referendum
- i th (randomly selected) person polled: $X_i = \begin{cases} 1, & \text{if yes,} \\ 0, & \text{if no.} \end{cases}$
- uniformly, independently
- $M_n = (X_1 + \dots + X_n)/n$: fraction of "yes" in our sample
- Would like "small error," e.g.: $|M_n - p| < 0.01$
- Try $n = 10,000$
- $P(|M_{10,000} - p| \geq 0.01) \leq \frac{\sigma^2}{n\epsilon^2} = \frac{p(1-p)}{10^4 \cdot 10^{-4}} \leq \frac{1}{4} \leftarrow \text{want } 5\%$
- $\frac{1/4}{n \cdot 10^{-4}} \leq \frac{5}{10^2} \Leftrightarrow n \geq \frac{10^6}{20} = 50,000 \leftarrow \text{will suffice}$

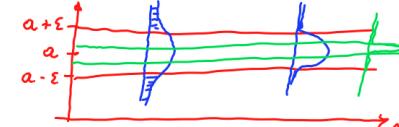


Understanding convergence "in probability"

- Ordinary convergence
 - Sequence a_n ; number a
 - $a_n \rightarrow a$
 - " a_n eventually gets and stays (arbitrarily) close to a "
- Convergence in probability
 - Sequence Y_n ; number a
 - $Y_n \rightarrow a$
 - for any $\epsilon > 0$, $P(|Y_n - a| \geq \epsilon) \rightarrow 0$



- For every $\epsilon > 0$, there exists n_0 , such that for every $n \geq n_0$, we have $|a_n - a| \leq \epsilon$



- (almost all) of the PMF/PDF of Y_n eventually gets concentrated (arbitrarily) close to a .

Some properties

- Suppose that $X_n \rightarrow a$, $Y_n \rightarrow b$, in probability

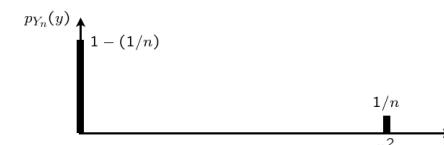
- If g is continuous, then $g(X_n) \rightarrow g(a)$

$$X_n^2 \rightarrow a^2$$

- $X_n + Y_n \rightarrow a + b$

- But:** $E[X_n]$ need not converge to a

Convergence in probability examples



$$\epsilon > 0 \quad P(|Y_n - 0| \geq \epsilon) = 1/n \xrightarrow[n \rightarrow \infty]{} 0$$

$$E[Y_n] = n^2 \cdot \frac{1}{n} = n \xrightarrow[n \rightarrow \infty]{} \infty$$

$$Y_n \xrightarrow[n \rightarrow \infty]{i.p.} 0.$$

- convergence in probability does **not** imply convergence of expectations

Convergence in probability examples

- X_i : i.i.d., uniform on $[0, 1]$



$$P(|Y_n - 0| \geq \epsilon) = P(Y_n \geq \epsilon).$$

$$\begin{aligned} \epsilon > 0 &= P(X_1 \geq \epsilon, \dots, X_n \geq \epsilon) & Y_n \xrightarrow[n \rightarrow \infty]{i.p.} 0 \\ \epsilon > 1 &= P(X_1 \geq \epsilon) \cdots P(X_n \geq \epsilon) \\ \epsilon \leq 1 &= (1 - \epsilon)^n \xrightarrow[n \rightarrow \infty]{} 0 \end{aligned}$$

Related topics

- Better bounds/approximations on tail probabilities
 - Markov and Chebyshev inequalities
 - Chernoff bound $P(|M_n - \mu| \geq a) \leq e^{-n\frac{\lambda(a)}{2}}$
 - Central limit theorem " $M_n \sim N(\mu, \sigma^2/n)$ "
- Different types of convergence
 - Convergence in probability
 - Convergence "with probability 1" $P(\{w : Y_n(w) \xrightarrow[n \rightarrow \infty]{} Y(w)\}) = 1$
 - Strong law of large numbers $M_n \xrightarrow[n \rightarrow \infty]{w.p.} \mu$
 - Convergence of a sequence of distributions (CDFs) to a limiting CDF

LECTURE 19: The Central Limit Theorem (CLT)

- WLLN: $\frac{X_1 + \dots + X_n}{n} \rightarrow E[X]$

- CLT: $X_1 + \dots + X_n \approx \text{normal}$
 - precise statement
 - universality, usefulness
 - many examples
 - refinement for discrete r.v.s
 - application to polling

Different scalings of the sum of i.i.d. random variables

- X_1, \dots, X_n i.i.d., finite mean μ and variance σ^2



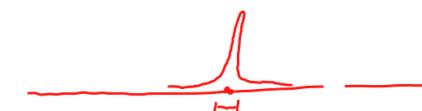
- $S_n = X_1 + \dots + X_n$

variance: $n\sigma^2$



- $M_n = \frac{S_n}{n} = \frac{X_1 + \dots + X_n}{n}$

variance: $\frac{\sigma^2}{n} \rightarrow 0$



- $\frac{S_n}{\sqrt{n}} = \frac{X_1 + \dots + X_n}{\sqrt{n}}$

variance: $\sigma^2 = \frac{n\sigma^2}{n}$



The Central Limit Theorem (CLT)

- X_1, \dots, X_n i.i.d., finite mean μ and variance σ^2

- $S_n = X_1 + \dots + X_n$ variance: $n\sigma^2$

- $\frac{S_n}{\sqrt{n}} = \frac{X_1 + \dots + X_n}{\sqrt{n}}$ variance: σ^2

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

$$E[Z_n] = 0$$

$$\text{var}(Z_n) = 1$$

- Let Z be a standard normal r.v. (zero mean, unit variance)

Central Limit Theorem: For every z : $\lim_{n \rightarrow \infty} P(\underline{Z_n \leq z}) = P(Z \leq \underline{z})$

- $P(Z \leq z)$ is the standard normal CDF, $\Phi(z)$, available from the normal tables

Usefulness of the CLT

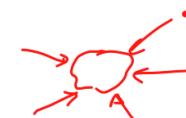
$$S_n = X_1 + \dots + X_n$$

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

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

Central Limit Theorem: For every z : $\lim_{n \rightarrow \infty} P(\underline{Z_n \leq z}) = P(Z \leq \underline{z})$

- universal and easy to apply; only means, variances matter
- fairly accurate computational shortcut
- justification of normal models



What exactly does the CLT say? — Theory

$$S_n = X_1 + \dots + X_n \quad Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} \quad Z \sim N(0, 1)$$

Central Limit Theorem: For every z : $\lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z)$

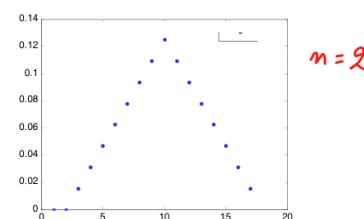
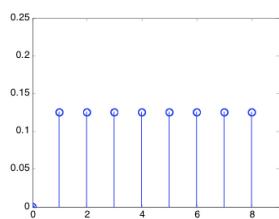
- CDF of Z_n converges to normal CDF
- results for convergence of PDFs or PMFs (with more assumptions)
- results without assuming that the X_i are identically distributed
- results under “weak dependence”
- proof: uses “transforms”: $E[e^{sZ_n}] \rightarrow E[e^{sZ}]$, for all s

What exactly does the CLT say? — Practice

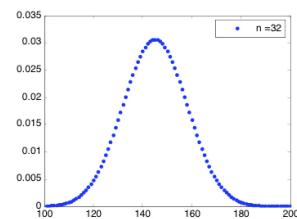
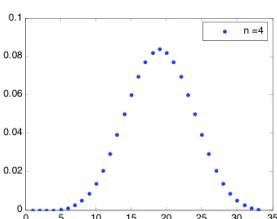
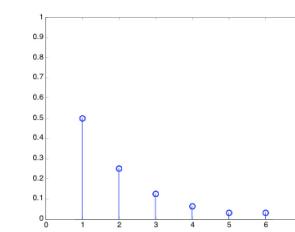
$$S_n = X_1 + \dots + X_n \quad Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} \quad Z \sim N(0, 1)$$

Central Limit Theorem: For every z : $\lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z)$

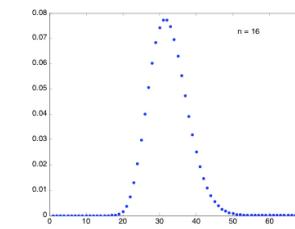
- The **practice** of normal approximations:
 - treat Z_n as if it were normal
 - hence treat S_n as if normal: $N(n\mu, n\sigma^2)$
- Can we use the CLT when n is “moderate”? $n = 30$?
- usually, yes
- symmetry and unimodality help



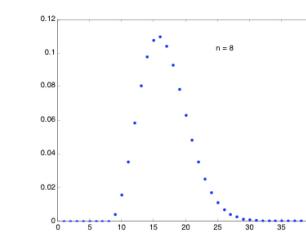
$n = 9$



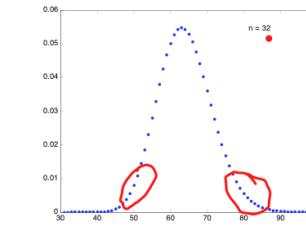
$n = 32$



$n = 16$



$n = 8$



$n = 32$

Example 1

- $P(S_n \leq a) \approx b$ given two parameters, find the third
- Package weights X_i , i.i.d. exponential, $\lambda = 1/2$;
- Load container with $n = 100$ packages

$$P(S_n \geq 210)$$

$$= P\left(\frac{S_n - 200}{20} > \frac{210 - 200}{20}\right)$$

$$= P(Z_n > 0.5) \approx P(Z > 0.5)$$

$$= 1 - P(Z < 0.5) = 1 - \Phi(0.5)$$

$$= 1 - 0.6915 = 0.3085$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
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

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

$$\mu = \sigma = 2$$

Example 2

Example 2

- $P(S_n \leq a) \approx b$ given two parameters, find the third

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

$$\mu = \sigma = 2$$

- Package weights X_i , i.i.d. exponential, $\lambda = 1/2$;
- Let $n = 100$. Choose the "capacity" a , so that $P(S_n \geq a) \approx 0.05$.

$$0.05 \approx P\left(\frac{S_n - 200}{20} > \frac{a - 200}{20}\right)$$

$$\approx 1 - \Phi\left(\frac{a - 200}{20}\right)$$

$$0.95$$

$$\frac{a - 200}{20} = 1.645 \quad a = 232.9$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Example 3

- $P(S_n \leq a) \approx b$ given two parameters, find the third
- Package weights X_i , i.i.d. exponential, $\lambda = 1/2$;
- How large can n be, so that $P(S_n \geq 210) \approx 0.05$?

$$P\left(\frac{S_n - 2n}{2\sqrt{n}} > \frac{210 - 2n}{2\sqrt{n}}\right)$$

$$\approx 1 - \Phi\left(\frac{210 - 2n}{2\sqrt{n}}\right) \approx 0.05$$

$$\frac{210 - 2n}{2\sqrt{n}} = 1.645$$

$$n = 89$$

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

$$\mu = \sigma = 2$$

Example 4

- $P(S_n \leq a) \approx b$ given two parameters, find the third

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

$$\mu = \sigma = 2$$

- Load container until weight exceeds 210 N: number of packages loaded

$$= P\left(\sum_{i=1}^{100} X_i \leq 210\right)$$

$$\approx \Phi\left(\frac{210 - 200}{20}\right) = \Phi(0.5) = 0.6915$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Normal approximation to the binomial

- X_i : independent, Bernoulli(p); $0 < p < 1$

$$S_n = X_1 + \dots + X_n: \text{Binomial}(n, p)$$

- mean np , variance $np(1-p)$

- $n = 36$, $p = 0.5$; find $P(S_n \leq 21)$

$$np = 18 \quad \sqrt{np(1-p)} = 3$$

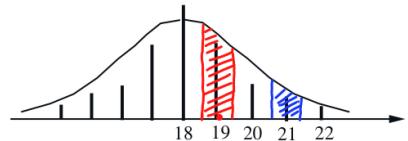
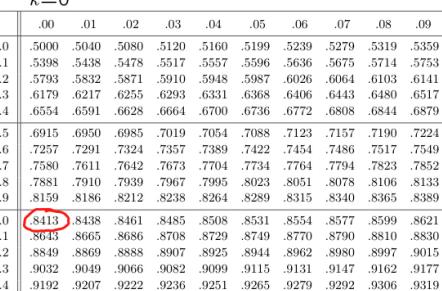
$$P\left(\frac{S_n - 18}{3} \leq \frac{21 - 18}{3}\right)$$

$$= P(Z_n \leq 1) \approx \Phi(1) = .8413$$

- CDF of $\frac{S_n - np}{\sqrt{np(1-p)}}$ \rightarrow 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	.6738	.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	.7938	.7967	.7994	.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



• 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 1/2 correction for integer random variables

- $0.8413 \approx P(S_n \leq 21) = P(S_n < 22)$, because S_n is integer

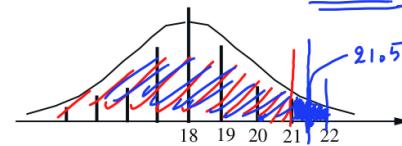
$$= P\left(\frac{S_n - 18}{3} < \frac{22 - 18}{3}\right)$$

$$= P(Z_n < 1.33) \approx \Phi(1.33) = 0.9082$$

true value 0.8785

$$P(S_n \leq 21.5) = P(Z_n \leq \frac{21.5 - 18}{3})$$

$$\approx \Phi(1.17) = \underline{\underline{.8790}}$$



	.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	.6738	.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	.7938	.7967	.7994	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

De Moivre–Laplace CLT to the binomial

$$P(S_n = 19) = P(18.5 \leq S_n \leq 19.5)$$

$$= P\left(\frac{18.5 - 18}{3} \leq Z_n \leq \frac{19.5 - 18}{3}\right)$$

$$= P(0.17 \leq Z_n \leq 0.5)$$

$$\approx \Phi(0.5) - \Phi(0.17)$$

$$= 0.6915 - 0.5675 = 0.124$$

• Exact answer:

$$\binom{36}{19} \left(\frac{1}{2}\right)^{36} = 0.1251$$

The pollster's problem revisited

- p : fraction of population that will vote "yes" in a referendum

- i th (randomly selected) person polled: $X_i = \begin{cases} 1, & \text{if yes,} \\ 0, & \text{if no.} \end{cases}$ $E[X_i] = p = \mu$
 $\sigma = \sqrt{p(1-p)}$

- $M_n = (X_1 + \dots + X_n)/n$: fraction of "yes" in our sample

- Would like "small error," e.g.: $|M_n - p| < 0.01$

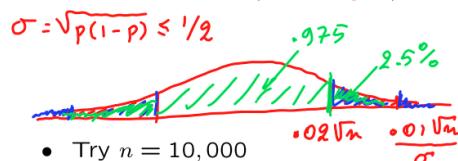
$$P(|M_n - p| \geq 0.01) = P\left(|Z_n| \geq \frac{0.01\sqrt{n}}{\sigma}\right) \approx P\left(|Z| \geq \frac{0.01\sqrt{n}}{\sigma}\right)$$

$$Z_n = \frac{S_n - np}{\sqrt{n}\sigma} \quad \left| \frac{S_n - np}{n} \right| \geq 0.01$$

$$\left| \frac{S_n - np}{\sqrt{n}\sigma} \right| \geq \frac{0.01\sqrt{n}}{\sigma}$$

The pollster's problem revisited

$$P(|M_n - p| \geq .01) \approx P\left(|Z| \geq \frac{.01\sqrt{n}}{\sigma}\right) \leq P\left(|Z| \geq .02\sqrt{n}\right) = 2(1 - \Phi(.02\sqrt{n})) = 0.05$$



$$\text{prob} \leq 2(1 - \Phi(2)) =$$

$$= 2(1 - 0.9772) = 0.046$$

$$\text{Specs: } P(|M_n - p| \geq .01) \leq .05$$

$$\Phi(.02\sqrt{n}) = 0.975$$

$$.02\sqrt{n} = 1.96 \Rightarrow n = 9604$$

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9775	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

Classical statistics

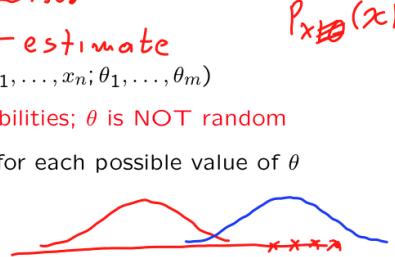
- Inference using the Bayes rule:
unknown Θ and observation X are both random variables
– Find $p_{\Theta|X}$

$$P_\theta \quad P_{X|\theta}$$

- Classical statistics: unknown constant θ

- also for vectors X and θ : $p_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta_1, \dots, \theta_m)$
- $p_X(x; \theta)$ are NOT conditional probabilities; θ is NOT random
- mathematically: many models, one for each possible value of θ

$$P_{X|\theta}(x|\theta)$$

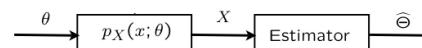


LECTURE 20: An introduction to classical statistics

- Unknown constant θ (not a r.v.)
- if $\theta = E[X]$: estimate using the sample mean $(X_1 + \dots + X_n)/n$
 - terminology and properties
- Confidence intervals (CIs)
 - CIs using the CLT
 - CIs when the variance is unknown
- Other uses of sample means
- Maximum Likelihood estimation

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 $\hat{\theta}$, to "keep estimation error $\hat{\theta} - \theta$ small"

Art! •

Estimating a mean

- X_1, \dots, X_n : i.i.d., mean θ , variance σ^2

$$\hat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \dots + X_n}{n} \quad \hat{\Theta}_n: \text{estimator (a random variable)}$$

Properties and terminology:

- $E[\hat{\Theta}_n] = \theta$ (unbiased)
for all θ *i.p.*
- WLLN: $\hat{\Theta}_n \rightarrow \theta$ (consistency)
for all θ
- mean squared error (MSE): $E[(\hat{\Theta}_n - \theta)^2] = \text{var}(\hat{\Theta}_n) = \frac{\sigma^2}{n}$.

$$\hat{\Theta} = g(x)$$

$$E[\hat{\Theta}] = \sum_x g(x) P_x(x; \theta)$$

$$\hat{\Theta} = O : MSE = O + \theta^2$$

On the mean squared error of an estimator

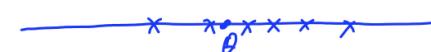
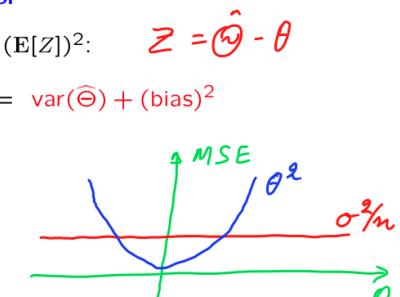
- For any estimator, using $E[Z^2] = \text{var}(Z) + (E[Z])^2$:

$$E[(\hat{\Theta} - \theta)^2] = \text{var}(\hat{\Theta} - \theta) + (E[\hat{\Theta} - \theta])^2 = \text{var}(\hat{\Theta}) + (\text{bias})^2$$

$$\hat{\Theta}_n = M_n : MSE = \sigma^2/n + 0$$

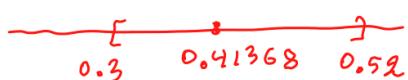
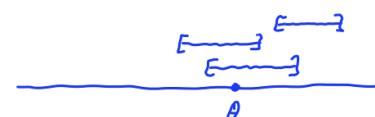
$$\hat{\Theta} = O : MSE = O + \theta^2$$

- $\sqrt{\text{var}(\hat{\Theta})}$ is called the standard error



Confidence intervals (CIs)

- The value of an estimator $\hat{\Theta}$ may not be informative enough
95%
- An $1 - \alpha$ confidence interval is an interval $[\hat{\Theta}^-, \hat{\Theta}^+]$,
s.t. $P(\hat{\Theta}^- \leq \theta \leq \hat{\Theta}^+) \geq 1 - \alpha$, for all θ
 - often $\alpha = 0.05$, or 0.025 , or 0.01
 - interpretation is subtle



$$P(0.3 < \theta < 0.52) \geq 0.95$$

CI for the estimation of the mean

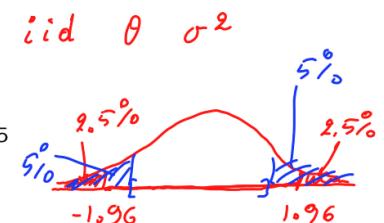
iid θ σ^2

$$\hat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \dots + X_n}{n}$$

95% normal tables: $\Phi(1.96) = 0.975 = 1 - 0.025$

90% $\Phi(1.645) = 0.95$

$$P\left(\frac{|\hat{\Theta}_n - \theta|}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95 \quad (\text{CLT})$$



$$P\left(\hat{\Theta}_n - \frac{1.96 \sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + \frac{1.96 \sigma}{\sqrt{n}}\right) \approx 0.95$$

$\hat{\Theta}^-$ $\hat{\Theta}^+$

Confidence intervals for the mean when σ is unknown

$$\widehat{\theta}_n = \text{sample mean} = M_n = \frac{X_1 + \dots + X_n}{n}$$

$$P\left(\widehat{\theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \widehat{\theta}_n + \frac{1.96\sigma}{\sqrt{n}}\right) \approx 0.95$$

- **Option 1:** use upper bound on σ
 - if X_i Bernoulli: $\sigma \leq 1/2$
- **Option 2:** use ad hoc estimate of σ
 - if X_i Bernoulli: $\hat{\sigma} = \sqrt{\widehat{\theta}_n(1 - \widehat{\theta}_n)}$

$$\sigma = \sqrt{\theta(1-\theta)}$$

Confidence intervals for the mean when σ is unknown

$$P\left(\widehat{\theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \widehat{\theta}_n + \frac{1.96\sigma}{\sqrt{n}}\right) \approx 0.95$$

- **Option 3:** Use sample mean estimate of the variance

Start from $\sigma^2 = E[(X_i - \theta)^2]$

$$\frac{1}{n} \sum_{i=1}^n (X_i - \theta)^2 \rightarrow \sigma^2$$

(but do not know θ)

$$\frac{1}{n-1} \sum_{i=1}^n (X_i - \widehat{\theta}_n)^2 \rightarrow \sigma^2$$

- Two approximations involved here:
 - CLT: approximately normal
 - using estimate of σ
- correction for second approximation (t -tables) used when n is small

Other natural estimators

$$\bullet \quad \theta_X = E[X] \quad \widehat{\theta}_X = \frac{1}{n} \sum_{i=1}^n X_i \quad \bullet \quad \theta = E[g(X)] \quad \widehat{\theta} = \frac{1}{n} \sum_{i=1}^n g(X_i)$$

$$\bullet \quad v_X = \text{var}(X) = E[(X - \theta_X)^2] \quad \widehat{v}_X = \frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\theta}_X)^2$$

$$\bullet \quad \text{cov}(X, Y) = E[(X - \theta_X)(Y - \theta_Y)] \quad \widehat{\text{cov}}(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\theta}_X)(Y_i - \widehat{\theta}_Y)$$

$$\bullet \quad \rho = \frac{\text{cov}(X, Y)}{\sqrt{v_X} \cdot \sqrt{v_Y}} \quad \widehat{\rho} = \frac{\widehat{\text{cov}}(X, Y)}{\sqrt{\widehat{v}_X} \cdot \sqrt{\widehat{v}_Y}}$$

• next steps: find the distribution of $\widehat{\theta}$, MSE, confidence intervals,...

Maximum Likelihood (ML) estimation

- Pick θ that "makes data most likely"

$$\widehat{\theta}_{\text{ML}} = \arg \max_{\theta} p_X(x; \theta)$$

– also applies when x, θ are vectors or x is continuous

$$\bullet \quad \text{compare to Bayesian posterior: } p_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{p_X(x)}$$

constant

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: $\hat{\Theta}_n \rightarrow \theta$
 - asymptotically normal: $\frac{\hat{\Theta}_n - \theta}{\sigma(\hat{\Theta}_n)} \rightarrow N(0, 1)$ (CDF convergence)
- analytical and simulation methods for calculating $\hat{\sigma} \approx \sigma(\hat{\Theta}_n)$
 - hence confidence intervals $P(\hat{\Theta}_n - 1.96 \hat{\sigma} \leq \theta \leq \hat{\Theta}_n + 1.96 \hat{\sigma}) \approx 0.95$
 - asymptotically “efficient” (“best”)

ML estimation example: parameter of binomial

- K : binomial with parameters n (known), and θ (unknown)

$$p_K(k; \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

$$\log \left[\binom{n}{k} \right] + k \log \theta + (n-k) \log (1-\theta)$$

$$0 + \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \Rightarrow k - k\theta = n\theta - k\theta$$

$$\hat{\theta}_{\text{ML}} = \frac{k}{n} \quad \hat{\Theta}_{\text{ML}} = \frac{K}{n}$$

- same as MAP estimator with uniform prior on θ

ML estimation example — normal mean and variance

$$\bullet X_1, \dots, X_n: \text{i.i.d., } N(\mu, v) \quad f_X(x; \mu, v) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi v}} \exp \left\{ -\frac{(x_i - \mu)^2}{2v} \right\}$$

$$\text{minimize } \frac{n}{2} \log v + \sum_{i=1}^n \frac{(x_i - \mu)^2}{2v}$$

$$\text{– minimize w.r.t. } \mu: \quad \hat{\mu} = \frac{x_1 + \dots + x_n}{n}$$

$$\frac{1}{v} \sum_{i=1}^n (x_i - \mu) = 0 \Rightarrow \sum x_i = n\mu$$

$$\text{– minimize w.r.t. } v: \quad \hat{v} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

$$\cancel{\frac{n}{2} \cdot \frac{1}{v}} \Rightarrow \sum_{i=1}^n \frac{(x_i - \mu)^2}{v} = 0$$

LECTURE 21: The Bernoulli process

- Definition of Bernoulli process
- Stochastic processes
- Basic properties (memorylessness)
- The time of the k th success/arrival
- Distribution of interarrival times
- Merging and splitting
- Poisson approximation

The Bernoulli process

- A sequence of independent Bernoulli trials, X_i
- At each trial, i :

$$P(X_i = 1) = P(\text{success at the } i\text{th trial}) = p$$

$$P(X_i = 0) = P(\text{failure at the } i\text{th trial}) = 1 - p$$
- Key assumptions:
 - Independence
 - Time-homogeneity
- Model of:
 - Sequence of lottery wins/losses
 - Arrivals (each second) to a bank
 - Arrivals (at each time slot) to server
 - ...



• Jacob Bernoulli
(1655–1705)

$$0 < p < 1$$

Number of successes/arrivals S in n time slots

- $S = X_1 + \dots + X_n$
- $P(S = k) = \binom{n}{k} p^k (1-p)^{n-k}$ $k = 0, \dots, n$
- $E[S] = np$
- $\text{var}(S) = np(1-p)$

Stochastic processes

infinite

- First view: sequence of random variables X_1, X_2, \dots

{ Interested in: $E[X_i] = p$ $\text{var}(X_i) = p(1-p)$ $p_{X_i}(x) = \begin{cases} p & x=1 \\ 1-p & x=0 \end{cases}$

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = p_{X_1}(x_1) \dots p_{X_n}(x_n)$$

for all n

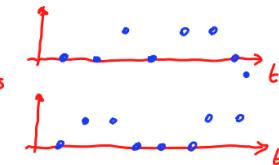
- Second view – sample space:

{ $\Omega = \text{set of infinite sequences}$
 of 0's and 1's

- Example (for Bernoulli process):

$$P(X_i = 1 \text{ for all } i) = 0 \quad (p < 1)$$

$$\leq P(X_1 = 1, \dots, X_n = 1) = p^n, \text{ for all } n$$



Time until the first success/arrival

- $T_1 = \min \{ i : X_i = 1 \}$
- $P(T_1 = k) = P(\underbrace{0, 0, \dots, 0}_{k-1}, 1) = (1-p)^{k-1} p$
 $k = 1, 2, \dots$
- $E[T_1] = \frac{1}{p}$
- $\text{var}(T_1) = \frac{1-p}{p^2}$

Independence, memorylessness, and fresh-start properties

$$\{X_i\} \sim \text{Ber}(p)$$

$$Y_1 = X_6^{\text{X}_{n+1}} \quad \{Y_i\} \quad \textcircled{1} \quad \{\text{Y}_i\} \text{ independent of } X_1, \dots, X_{\textcolor{red}{n}}$$

$$Y_2 = X_7^{\text{X}_{n+2}} \quad \{i=1,2,\dots\} \quad \textcircled{2} \quad \text{Ber}(p)$$

- Fresh-start after time n



$$Y_1 = X_{T_1+1} \quad \textcircled{1} \quad \{\text{Y}_i\} \text{ independent of } X_1, \dots, X_{T_1}$$

$$Y_2 = X_{T_1+2} \quad \textcircled{2} \quad \text{Ber}(p)$$

- Fresh-start after time T_1

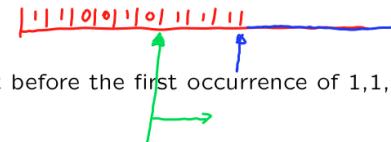
Independence, memorylessness, and fresh-start properties

- Fresh-start after a random time N ?

N = time of 3rd success



N = first time that 3 successes in a row have been observed



$\left. \begin{array}{l} N \text{ is causally determined} \\ N \text{ not causally determined} \end{array} \right\}$

The process X_{N+1}, X_{N+2}, \dots is:

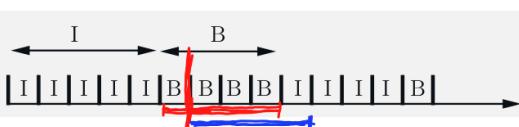
- a Bernoulli process (as long as N is determined "causally")
- independent of N, X_1, \dots, X_N

The distribution of busy periods

- At each slot, a server is busy or idle (Bernoulli process) P

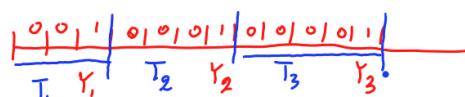
- First busy period: $\text{Geo}(1-p)$

- starts with first busy slot
- ends just before the first subsequent idle slot



$\text{Geo}(1-p)$

Time of the k th success/arrival



- Y_k = time of k th arrival

$$Y_k = T_1 + \dots + T_k$$

- T_k = k th inter-arrival time = $Y_k - Y_{k-1}$ ($k \geq 2$)

- The process starts fresh after time T_1

- T_2 is independent of T_1 ; Geometric(p); etc.

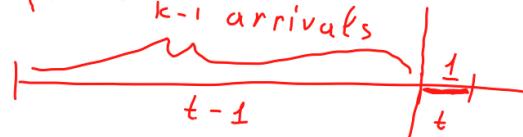
Time of the k th success/arrival

$$P(Y_k = t)$$

= $P(k-1 \text{ arrivals in time } t-1)$

$\cdot P(\text{arrival at time } t)$

$$= \binom{t-1}{k-1} p^{k-1} (1-p)^{t-k} \cdot p$$



$$Y_k = T_1 + \dots + T_k$$

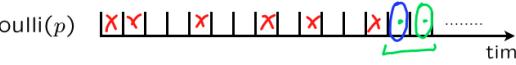
the T_i are i.i.d., Geometric(p)

$$E[Y_k] = \frac{k}{p} \quad \text{var}(Y_k) = \frac{k(1-p)}{p^2}$$

$$p_{Y_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \quad t = k, k+1, \dots$$

Merging of independent Bernoulli processes

$$X_t \quad \text{Bernoulli}(p)$$



$$Z_t \quad \text{merged process}$$

$$\text{Bernoulli}(p+q-pq)$$

(collisions are counted as one arrival)

$$Y_t \quad \text{Bernoulli}(q)$$

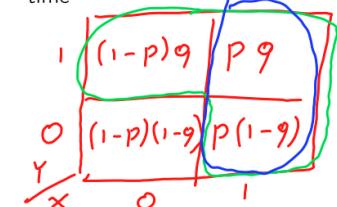


time

$$Z_t = g(X_t, Y_t) \quad (Z_1, \dots, Z_t)$$

$$Z_{t+1} = g(X_{t+1}, Y_{t+1}) \quad 1 - (1-p)(1-q)$$

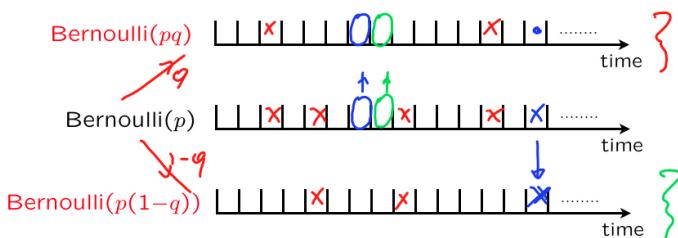
$$P(\text{arrival in first process} \mid \text{arrival}) = \frac{p}{p+q-pq}$$



Splitting of a Bernoulli process



- Split successes into two streams, using independent flips of a coin with bias q
 - assume that coin flips are independent from the original Bernoulli process



- Are the two resulting streams independent? **No**

Poisson approximation to binomial

- Interesting regime: large n , small p , moderate $\lambda = np$
- Number of arrivals S in n slots: $\underline{p_S(k) = \frac{n!}{(n-k)!k!} \cdot p^k (1-p)^{n-k}}$, $k = 0, \dots, n$

$$\text{For fixed } k = 0, 1, \dots, \\ p_S(k) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda},$$

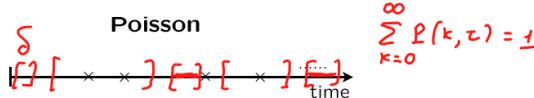
$$\begin{aligned} &= \frac{n \cdot (n-1) \cdots (n-k+1)}{k!} \cdot \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} \cdot \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &\xrightarrow{n \rightarrow \infty} 1 \cdot 1 \cdots 1 \cdot \frac{\lambda^k}{k!} e^{-\lambda} \cdot 1 \end{aligned}$$

- Fact: $\lim_{n \rightarrow \infty} (1 - \lambda/n)^n = e^{-\lambda}$

LECTURE 22: The Poisson process

- Definition of the Poisson process
 - applications
- Distribution of number of arrivals
- The time of the k th arrival
- Memorylessness
- Distribution of interarrival times

Definition of the Poisson process



- Numbers of arrivals in disjoint time intervals are **independent**

$P(k, \tau)$ = Prob. of k arrivals in interval of duration τ

- **Small interval probabilities:**

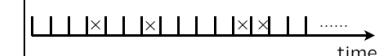
For VERY small δ :

$$P(k, \delta) \approx \begin{cases} 1 - \lambda\delta & \text{if } k = 0 \\ \lambda\delta & \text{if } k = 1 \\ 0 & \text{if } k > 1 \end{cases} \quad P(k, \delta) = \begin{cases} 1 - \lambda\delta + O(\delta^2) & \text{if } k = 0 \\ \lambda\delta + O(\delta^2) & \text{if } k = 1 \\ 0 + O(\delta^2) & \text{if } k > 1 \end{cases}$$

$$\frac{O(\delta^2)}{\delta} \xrightarrow{\delta \rightarrow 0} 0$$

(λ: "arrival rate") •

Bernoulli



- Independence

- **Time homogeneity:**
Constant p at each slot

Applications of the Poisson process



- Deaths from horse kicks in the Prussian army (1898)
- Particle emissions and radioactive decay
- Photon arrivals from a weak source
- Financial market shocks
- Placement of phone calls, service requests, etc. •



Siméon Denis Poisson
(1781-1840)

The Poisson PMF for the number of arrivals



- N_τ : arrivals in $[0, \tau]$ $P(k, \tau) = P(N_\tau = k)$

$n = \tau/\delta$ intervals/slots of length δ ← small

$P(\text{some slot contains two or more arrivals})$

$$\leq \sum_i P(\text{slot } i \text{ has } \geq 2 \text{ arrivals})$$

$$= \frac{\tau}{\delta} O(\delta^2) \xrightarrow{\delta \rightarrow 0} 0$$

$P(k \text{ arrivals in Poisson}) \approx P(k \text{ slots have arrivals})$

$N_\tau \approx \text{binomial}$

$$p = \lambda\tau + O(\delta^2) \quad np = \lambda\tau + O(\delta) \approx \lambda\tau$$

$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

Bernoulli

$$p_S(k) = \frac{n!}{(n-k)! k!} \cdot p^k (1-p)^{n-k}, \quad k = 0, \dots, n$$

$$\lambda = np \quad n \rightarrow \infty \quad p \rightarrow 0$$

For fixed $k = 0, 1, \dots$,

$$p_S(k) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda},$$

Mean and variance of the number of arrivals

$$P(k, \tau) = P(N_\tau = k) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

$$E[N_\tau] = \sum_{k=0}^{\infty} k \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!} = \dots = \lambda\tau$$

$N_\tau \approx \text{Binomial}(n, p)$

$$n = \tau/\delta, \quad p = \lambda\delta + O(\delta^2)$$

$$E[N_\tau] \approx np \approx \lambda\tau$$

$$\text{var}(N_\tau) \approx np(1-p) \approx \lambda\tau$$

$$E[N_\tau] = \lambda\tau$$

$$\text{var}(N_\tau) = \lambda\tau$$

$$\lambda = \frac{E[N_\tau]}{\tau}$$

Example

- You get email according to a Poisson process, at a rate of $\lambda = 5$ messages per hour.

$$E[N_\tau] = \lambda\tau$$

$$\text{var}(N_\tau) = \lambda\tau$$

- Mean and variance of mails received during a day = $5 \cdot 24$
- $P(\text{one new message in the next hour}) = P(1, 1) = 5e^{-5}$

$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

- $P(\text{exactly two messages during each of the next three hours}) =$

$$\underbrace{2+2+2}_{\text{3 hours}} \quad (P(2, 1))^3 = \left(\frac{5^2 e^{-5}}{2}\right)^3$$

The time T_1 until the first arrival



- Find the CDF: $P(T_1 \leq t) =$

$$= 1 - P(T_1 > t) = 1 - P(0, t) = 1 - e^{-\lambda t}$$

$$f_{T_1}(t) = \lambda e^{-\lambda t}, \quad \text{for } t \geq 0$$

Exponential(λ)

$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

Memorylessness: conditioned on $T_1 > t$, the PDF of $T_1 - t$ is again exponential

The time Y_k of the k th arrival

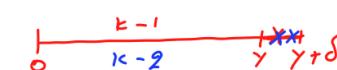
$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \dots$$

- Can derive its PDF by first finding the CDF

$$P(Y_k \leq y) = \sum_{n=k}^{\infty} P(n, y)$$

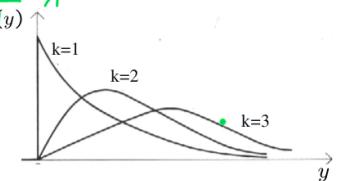
- More intuitive argument:

$$\begin{aligned} f_{Y_k}(y) &\approx P(y \leq Y_k \leq y + \delta) = \\ &\approx P(k-1, y) \lambda \delta \\ &\quad + P(k-2, y) O(\delta^2) \\ &\quad + P(k-3, y) O(\delta^3) \end{aligned}$$



$$\frac{(\lambda y)^{k-1} e^{-\lambda y}}{(k-1)!}$$

$$\text{Erlang distribution: } f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0$$



Memorylessness and the fresh-start property

- Analogous to the properties for the Bernoulli process
 - plausible, given the relation between the two processes
 - use intuitive reasoning
 - can be proved rigorously

Memorylessness and the fresh-start property

- If we start watching at time t ,
we see Poisson process, independent of the history until time t
time until next arrival: $\text{Exp}(\lambda)$, independent of past
- If we start watching at time T_1 , $T_1 = 3$
we see Poisson process, independent of the history until time T_1
hence: time between first and second arrival, $T_2 = Y_2 - Y_1$ is: $\text{Exp}(\lambda)$
similarly for all $T_k = Y_k - Y_{k-1}$, $k \geq 2$



$Y_k = T_1 + \dots + T_k$ is sum of i.i.d. exponentials

$$\mathbb{E}[Y_k] = k/\lambda \quad \text{var}(Y_k) = k/\lambda^2$$

- An equivalent definition
- A simulation method

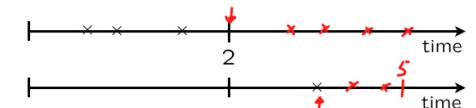
Bernoulli/Poisson relation

$$\begin{array}{c} \text{Bernoulli: } \text{ discrete events over time} \\ \text{Poisson: } \text{ continuous events over time} \end{array} \quad n = \tau/\delta, \quad np = \lambda\tau$$

	POISSON	BERNOULLI
Times of Arrival	Continuous	Discrete
Arrival Rate	$\lambda/\text{unit time}$	$p/\text{per trial}$
PMF of # of Arrivals	Poisson	Binomial
Interarrival Time Distr.	Exponential	Geometric
Time to k -th arrival	Erlang	Pascal

Example: Poisson fishing

- Fish are caught as a Poisson process, $\lambda = 0.6/\text{hour}$
 - fish for two hours;
 - if you caught at least one fish, stop
 - else continue until first fish is caught



$$P(\text{fish for more than two hours}) = P(0, 2)$$

$$P(T_1 > 2) = \int_2^\infty f_{T_1}(t) dt$$

$$P(\text{fish for more than two and less than five hours}) =$$

$$P(0, 2) (1 - P(0, 3))$$

$$P(2 < T_1 \leq 5) = \int_2^5 f_{T_1}(t) dt$$

$$P(k, \tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}$$

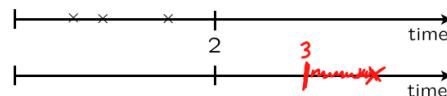
$$\mathbb{E}[N_\tau] = \lambda\tau$$

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}$$

Example: Poisson fishing

- Fish are caught as a Poisson process, $\lambda = 0.6/\text{hour}$

- fish for two hours;
- if you caught at least one fish, stop
- else continue until first fish is caught



$P(\text{catch at least two fish}) =$

$$\sum_{k=2}^{\infty} P(k, 2) = 1 - P(0, 2) - P(1, 2)$$

$$P(Y_2 \leq 2) = \int_0^2 f_{Y_2}(y) dy$$

$E[\text{future fishing time} \mid \text{already fished for three hours}] = \frac{1}{\lambda}$

$$P(k, \tau) = \frac{(\lambda \tau)^k e^{-\lambda \tau}}{k!}$$

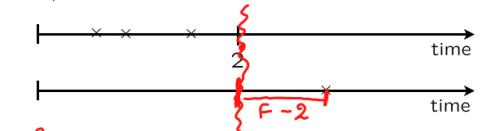
$$E[N_\tau] = \lambda \tau$$

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}$$

Example: Poisson fishing

- Fish are caught as a Poisson process, $\lambda = 0.6/\text{hour}$

- fish for two hours;
- if you caught at least one fish, stop
- else continue until first fish is caught



$$E[\text{total fishing time}] = E[F] = 2 + E[F-2]$$

$$= 2 + P(F=2) \cdot 0 + P(F>2) E[F-2 | F>2]$$

$$= 2 + P(0, 2) \cdot 1/1$$

$$E[\text{number of fish}] = \lambda \tau + P(0, 2) \cdot 1$$

$$0.6 \times 2$$

$$P(k, \tau) = \frac{(\lambda \tau)^k e^{-\lambda \tau}}{k!}$$

$$E[N_\tau] = \lambda \tau$$

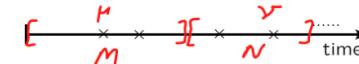
$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}$$

LECTURE 23: More on the Poisson process

- The sum of independent Poisson r.v.s
- Merging and splitting
- Random incidence

The sum of independent Poisson random variables

- Poisson process of rate $\lambda = 1$



- Consecutive intervals of length μ and ν

$$P(k, \tau) = \frac{(\lambda \tau)^k e^{-\lambda \tau}}{k!}$$

$\text{Poisson}(\lambda \tau)$

- Numbers of arrivals during these intervals: M and N

$M: \text{Poisson}(\mu)$

• Independent? Yes

$N: \text{Poisson}(\nu)$

• $M + N: \text{Poisson}(\mu + \nu)$

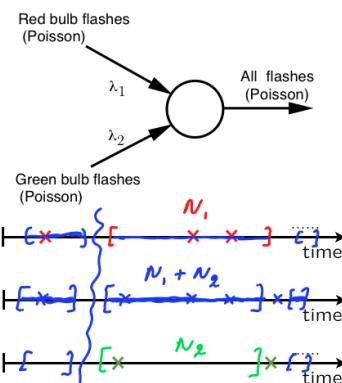
The sum of independent Poisson random variables, with means/parameters μ and ν , is Poisson with mean/parameter $\mu + \nu$

Merging of independent Poisson processes

	$1 - \lambda_1\delta$	$\lambda_1\delta$	$O(\delta^2)$
$1 - \lambda_2\delta$	0 $(1-\lambda_1\delta), \lambda_1\delta(1-\lambda_2\delta)$ $(1-\lambda_2\delta)$	1	≥ 2
$\lambda_2\delta$	1 $\lambda_2\delta(1-\lambda_1\delta)$	$\lambda_1\lambda_2\delta^2$	*
$O(\delta^2)$	≥ 2	*	*

$O: 1 - (\lambda_1 + \lambda_2)\delta$

$1: (\lambda_1 + \lambda_2)\delta$



Merged process: $\text{Poisson}(\lambda_1 + \lambda_2)$

Where is an arrival of the merged process coming from?

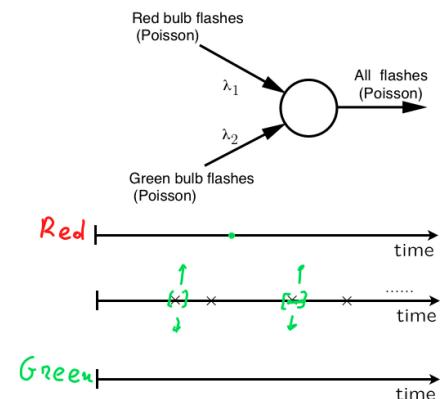
$$P(\text{Red} \mid \text{arrival at time } t) = \lambda_1 / (\lambda_1 + \lambda_2)$$

	$1 - \lambda_1\delta$	$\lambda_1\delta$	$O(\delta^2)$
$1 - \lambda_2\delta$	0 $1 - (\lambda_1 + \lambda_2)\delta$	$\lambda_1\delta$	$O(\delta^2)$
$\lambda_2\delta$	1 $\lambda_2\delta$	$O(\delta^2)$	≥ 2

$$P(k\text{th arrival is Red}) = \lambda_1 / (\lambda_1 + \lambda_2)$$

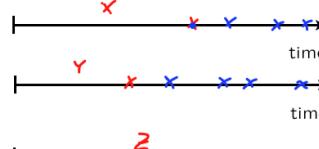
- Independence for different arrivals

$$P(4 \text{ out of first 10 arrivals are Red}) = \binom{10}{4} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^4 \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^6$$



The time the first (or the last) lightbulb burns out

- Three lightbulbs
 - independent lifetimes X, Y, Z ; exponential(λ)



$$\begin{aligned} E[\min\{X, Y, Z\}] &= \int_0^\infty \min\{x, y, z\} \lambda e^{-\lambda x} \lambda e^{-\lambda y} \lambda e^{-\lambda z} dx dy dz \\ &= \frac{1}{3\lambda} \end{aligned}$$

$$P(\min\{X, Y, Z\} \geq t) = P(X \geq t, Y \geq t, Z \geq t) = e^{-\lambda t} e^{-\lambda t} e^{-\lambda t} = e^{-3\lambda t}$$

- X, Y, Z : first arrivals in independent Poisson processes

Merged process: $\text{Poisson}(3\lambda)$

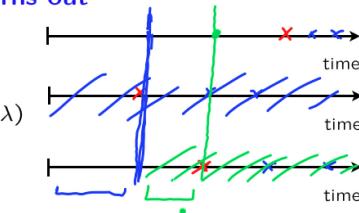
- $\min\{X, Y, Z\}$: 1st arrival in merged process

The time the first (or the last) lightbulb burns out

- Three lightbulbs
 - independent lifetimes X, Y, Z ; exponential(λ)

- Find expected time until all burn out

$$\max\{X, Y, Z\}$$

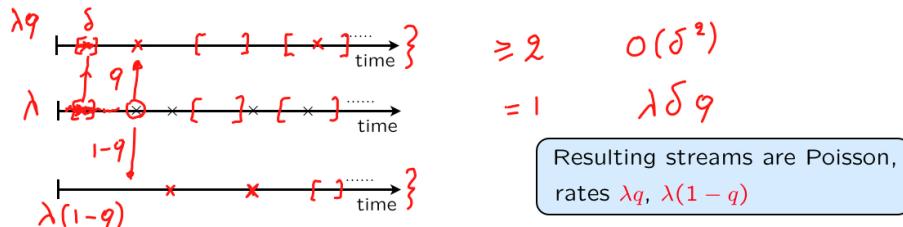


$$\frac{1}{3\lambda} + \frac{1}{2\lambda} + \frac{1}{\lambda}$$

Splitting of a Poisson process

- Split arrivals into two streams, using independent coin flips of a coin with bias q

— assume that coin flips are independent from the original Poisson process



- Are the two resulting streams independent?

Surprisingly, yes! •

"Random incidence" in the Poisson process

- Poisson process that has been running forever

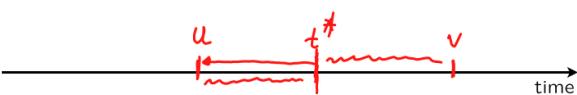


- Believe that $\lambda = 4/\text{hour}$, so that $E[T_k] = \frac{1}{\lambda} \text{ hrs} = 15 \text{ mins}$

- Show up at some time and measure interarrival time

— do it many times, average results, see something around 30 mins! Why?

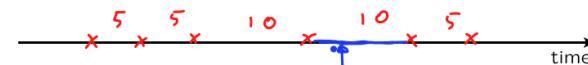
"Random incidence" in the Poisson process — analysis



- Arrive at time t^*
- U : last arrival time • V : next arrival time
- $V - U = \frac{(V-t^*)}{Exp(\lambda)} + \frac{(t^*-U)}{Exp(\lambda)}$
- $E[V - U] = \frac{1}{\lambda} + \frac{1}{\lambda} = \frac{2}{\lambda}$
- $V - U$: interarrival time you see, versus k th interarrival time

$$\bullet \frac{1}{\lambda}$$

Random incidence "paradox" is not special to the Poisson process



- Example: interarrival times, i.i.d., equally likely to be 5 or 10 minutes

expected value of k th interarrival time: $\frac{1}{2} \cdot 5 + \frac{1}{2} \cdot 10 = 7.5$

- you show up at a "random time"

$P(\text{arrive during a 5-minute interarrival interval}) = \frac{1}{3}$

expected length of interarrival interval during which you arrive = $\frac{1}{3} \cdot 5 + \frac{2}{3} \cdot 10 \approx 8.3$

- Calculation generalizes to "renewal processes:" i.i.d. interarrival times, from some general distribution

- "Sampling method" matters

Different sampling methods can give different results

- Average family size?

– look at a “random” family (uniformly chosen)

$$\frac{3}{4} \cdot 1 + \frac{1}{4} \cdot 6$$

– look at a “random” person's (uniformly chosen) family $\frac{3}{9} \cdot 1 + \frac{6}{9} \cdot 6$

- Average bus occupancy?

– look at a “random” bus (uniformly chosen)

$$0 \quad 50$$

– look at a “random” passenger's bus

- Average class size?