6.041 Probabilistic Systems Analysis
6.431 Applied Probability

- Staff:
  - Lecturer: John Tsitsiklis
- Pick up **and read** course information handout
- **Turn in recitation and tutorial scheduling form** (last sheet of course information handout)
- Pick up copy of slides

Coursework

- Quiz 1 (October 12, 12:05-12:55pm) 17%
- Quiz 2 (November 2, 7:30-9:30pm) 30%
- Final exam (scheduled by registrar) 40%
- Weekly homework (best 9 of 10) 10%
- Attendance/participation/enthusiasm in recitations/tutorials 3%

- **Collaboration policy** described in course info handout
- Text: *Introduction to Probability*, 2nd Edition, D. P. Bertsekas and J. N. Tsitsiklis, Athena Scientific, 2008
  Read the text!

LECTURE 1

- **Readings:** Sections 1.1, 1.2

Lecture outline

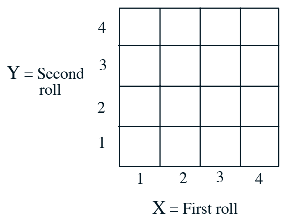
- Probability as a mathematical framework for reasoning about uncertainty
- Probabilistic models
  - sample space
  - probability law
- Axioms of probability
- Simple examples

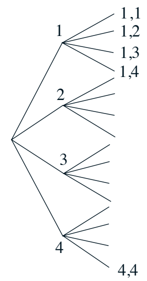
Sample space Ω

- "List" (set) of possible outcomes
- List must be:
  - Mutually exclusive
  - Collectively exhaustive
- Art: to be at the "right" granularity

Sample space: Discrete example

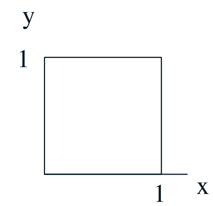
- Two rolls of a tetrahedral die
  - Sample space vs. sequential description





Sample space: Continuous example

$\Omega = \{(x, y) \mid 0 \le x, y \le 1\}$



Probability axioms

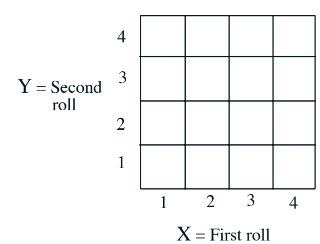
- **Event:** a subset of the sample space
- Probability is assigned to events

Axioms:

1. **Nonnegativity:** $\mathbf{P}(A) \ge 0$
2. **Normalization:** $\mathbf{P}(\Omega) = 1$
3. **Additivity:** If $A \cap B = \varnothing$, then $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$

- $\mathbf{P}(\{s_1, s_2, \ldots, s_k\}) = \mathbf{P}(\{s_1\}) + \cdots + \mathbf{P}(\{s_k\})$
  $= \mathbf{P}(s_1) + \cdots + \mathbf{P}(s_k)$
- Axiom 3 needs strengthening
- Do weird sets have probabilities?

Probability law: Example with finite sample space



- Let every possible outcome have probability 1/16
  - $\mathbf{P}((X, Y)$ is (1,1) or (1,2)$) =$
  - $\mathbf{P}(\{X = 1\}) =$
  - $\mathbf{P}(X + Y$ is odd$) =$
  - $\mathbf{P}(\min(X, Y) = 2) =$

Discrete uniform law

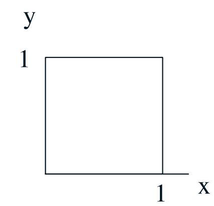
- Let all outcomes be equally likely
- Then,

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

- Computing probabilities $\equiv$ counting
- Defines fair coins, fair dice, well-shuffled decks

Continuous uniform law

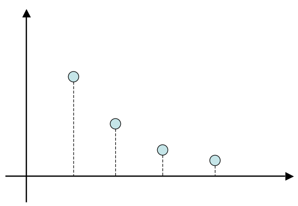
- Two "random" numbers in $[0, 1]$.



- **Uniform** law: Probability = Area
  - $\mathbf{P}(X + Y \leq 1/2) = ?$
  - $\mathbf{P}(\,(X, Y) = (0.5, 0.3)\,)$

Probability law: Ex. w/countably infinite sample space

- Sample space: $\{1, 2, \ldots\}$
  - We are given $\mathbf{P}(n) = 2^{-n}$, $n = 1, 2, \ldots$
  - Find $\mathbf{P}$(outcome is even)



$$\mathbf{P}(\{2, 4, 6, \ldots\}) \;=\; \mathbf{P}(2) + \mathbf{P}(4) + \cdots = \frac{1}{2^2} + \frac{1}{2^4} + \frac{1}{2^6} + \cdots = \frac{1}{3}$$

- Countable additivity axiom (needed for this calculation):
  If $A_1, A_2, \ldots$ are disjoint events, then:

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots$$

Remember!

- Turn in recitation/tutorial scheduling form **now**
- Tutorials start next week

LECTURE 2

- **Readings:** Sections 1.3-1.4

Lecture outline

- Review
- Conditional probability
- Three **important** tools:
  - Multiplication rule
  - Total probability theorem
  - Bayes' rule

Review of probability models

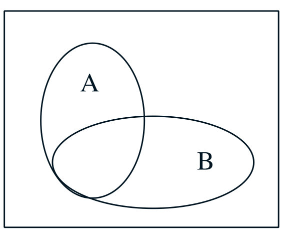
- Sample space Ω
  - Mutually exclusive
    Collectively exhaustive
  - Right granularity
- Event: Subset of the sample space
- Allocation of probabilities to events

1. $\mathbf{P}(A) \geq 0$
2. $\mathbf{P}(\Omega) = 1$
3. If $A \cap B = \varnothing$,
   then $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$

3'. If $A_1, A_2, \ldots$ are disjoint events, then:
$\mathbf{P}(A_1 \cup A_2 \cup \cdots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots$

- Problem solving:
  - Specify sample space
  - Define probability law
  - Identify event of interest
  - Calculate...

Conditional probability

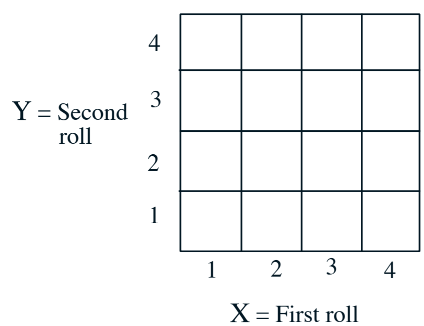


- $\mathbf{P}(A \mid B) =$ probability of A,
  given that B occurred
  - B is our new universe
- **Definition:** Assuming $\mathbf{P}(B) \neq 0$,

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

$\mathbf{P}(A \mid B)$ undefined if $\mathbf{P}(B) = 0$

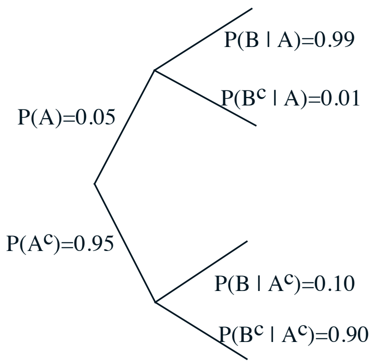
Die roll example



- Let B be the event: $\min(X, Y) = 2$
- Let $M = \max(X, Y)$
- $\mathbf{P}(M = 1 \mid B) =$
- $\mathbf{P}(M = 2 \mid B) =$

Models based on conditional probabilities

- Event A: Airplane is flying above
  Event B: Something registers on radar screen



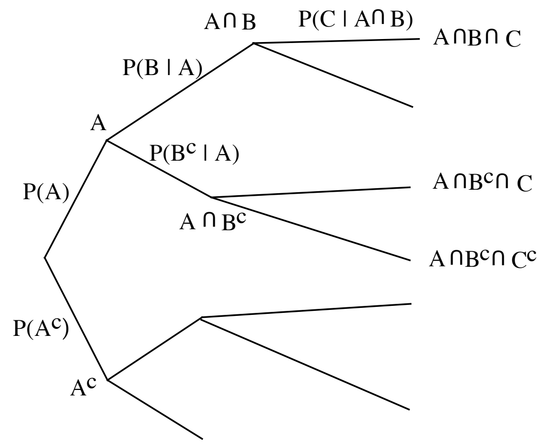
$\mathbf{P}(A \cap B) =$

$\mathbf{P}(B) =$

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

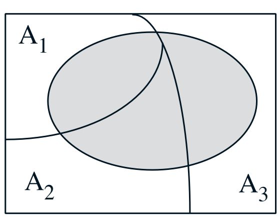
Multiplication rule

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



Total probability theorem

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

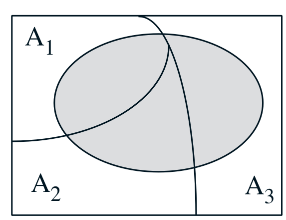


- One way of computing $\mathbf{P}(B)$:

$$\begin{aligned}\mathbf{P}(B) = \; & \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) \\ + \; & \mathbf{P}(A_2)\mathbf{P}(B \mid A_2) \\ + \; & \mathbf{P}(A_3)\mathbf{P}(B \mid A_3)\end{aligned}$$

Bayes' rule

- "Prior" probabilities $\mathbf{P}(A_i)$
  – initial "beliefs"
- We know $\mathbf{P}(B \mid A_i)$ for each i
- Wish to compute $\mathbf{P}(A_i \mid B)$
  – revise "beliefs", given that B occurred



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

LECTURE 3

- **Readings:** Section 1.5
- Review
- Independence of two events
- Independence of a collection of events

Review

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}, \qquad \text{assuming } \mathbf{P}(B) > 0$$

- Multiplication rule:

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

- Total probability theorem:

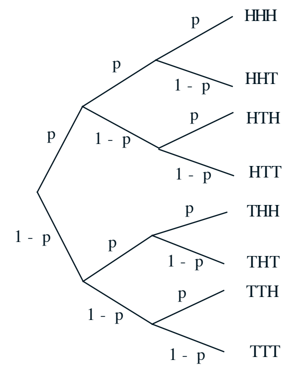
$$\mathbf{P}(B) = \mathbf{P}(A)\mathbf{P}(B \mid A) + \mathbf{P}(A^c)\mathbf{P}(B \mid A^c)$$

- Bayes rule:

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

Models based on conditional probabilities

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



$\mathbf{P}(THT) =$

$\mathbf{P}(\text{1 head}) =$

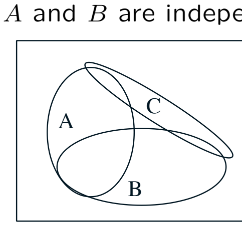
$\mathbf{P}(\text{first toss is H} \mid \text{1 head}) =$

Independence of two events

- **"Defn:"** $\mathbf{P}(B \mid A) = \mathbf{P}(B)$
  - "occurrence of A provides no information about B's occurrence"
- Recall that $\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B \mid A)$
- **Defn:** $\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B)$
- Symmetric with respect to A and B
  - applies even if $\mathbf{P}(A) = 0$
  - implies $\mathbf{P}(A \mid B) = \mathbf{P}(A)$

Conditioning may affect independence

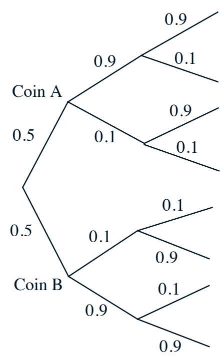
- Conditional independence, given C, is defined as independence under probability law $\mathbf{P}(\cdot \mid C)$
- Assume A and B are independent



- If we are told that C occurred, are A and B independent?

Conditioning may affect independence

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



- Once we know it is coin A, are tosses independent?
- If we do not know which coin it is, are tosses independent?
  - Compare:
    $\mathbf{P}(\text{toss } 11 = H)$
    $\mathbf{P}(\text{toss } 11 = H \mid \text{first 10 tosses are heads})$

Independence of a collection of events

- Intuitive definition:
  Information on some of the events tells us nothing about probabilities related to the remaining events
  - E.g.:

$$\mathbf{P}(A_1 \cap (A_2^c \cup A_3) \mid A_5 \cap A_6^c) = \mathbf{P}(A_1 \cap (A_2^c \cup A_3))$$

- Mathematical definition:
  Events $A_1, A_2, \ldots, A_n$
  are called **independent** if:

$$\mathbf{P}(A_i \cap A_j \cap \cdots \cap A_q) = \mathbf{P}(A_i)\mathbf{P}(A_j)\cdots\mathbf{P}(A_q)$$

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

Independence vs. pairwise independence

- Two independent fair coin tosses
  - A: First toss is H
  - B: Second toss is H
  - $\mathbf{P}(A) = \mathbf{P}(B) = 1/2$

HH	HT
TH	TT

  - C: First and second toss give same result
  - $\mathbf{P}(C) =$
  - $\mathbf{P}(C \cap A) =$
  - $\mathbf{P}(A \cap B \cap C) =$
  - $\mathbf{P}(C \mid A \cap B) =$
- Pairwise independence **does not** imply independence

The king's sibling

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

LECTURE 4

- **Readings:** Section 1.6

Lecture outline

- Principles of counting
- Many examples
  - permutations
  - k-permutations
  - combinations
  - partitions
- Binomial probabilities

Discrete uniform law

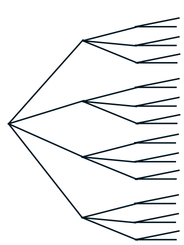
- Let all sample points be equally likely
- Then,

$$\mathbf{P}(A) = \frac{\text{number of elements of } A}{\text{total number of sample points}} = \frac{|A|}{|\Omega|}$$

- Just count...

Basic counting principle

- r stages
- n_i choices at stage i



- Number of choices is: $n_1 n_2 \cdots n_r$
- Number of license plates with 3 letters and 4 digits =
- ... if repetition is prohibited =
- **Permutations:** Number of ways of ordering n elements is:
- Number of subsets of $\{1, \ldots, n\}$ =

Example

- Probability that six rolls of a six-sided die all give different numbers?
  - Number of outcomes that make the event happen:
  - Number of elements in the sample space:
  - Answer:

Combinations

- $\binom{n}{k}$: number of k-element subsets of a given n-element set
- Two ways of constructing an ordered sequence of k **distinct** items:
  - Choose the k items one at a time: $n(n-1)\cdots(n-k+1) = \dfrac{n!}{(n-k)!}$ choices
  - Choose k items, then order them ($k!$ possible orders)
- Hence:

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

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

$$\sum_{k=0}^{n} \binom{n}{k} =$$

Binomial probabilities

- n independent coin tosses
  - $\mathbf{P}(H) = p$
- $\mathbf{P}(HTTHHH) =$
- $\mathbf{P}(\text{sequence}) = p^{\#\text{ heads}}(1-p)^{\#\text{ tails}}$

$$\mathbf{P}(k \text{ heads}) = \sum_{k\text{-head seq.}} \mathbf{P}(\text{seq.})$$

$$= (\#\text{ of } k\text{-head seqs.}) \cdot p^k(1-p)^{n-k}$$

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

Coin tossing problem

- event B: 3 out of 10 tosses were "heads".
  - Given that B occurred, what is the (conditional) probability that the first 2 tosses were heads?
- All outcomes in set B are equally likely: probability $p^3(1-p)^7$
  - Conditional probability law is uniform
- Number of outcomes in B:
- Out of the outcomes in B, how many start with HH?

Partitions

- 52-card deck, dealt to 4 players
- Find $\mathbf{P}(\text{each gets an ace})$
- Outcome: a partition of the 52 cards
  - number of outcomes:

$$\frac{52!}{13!\,13!\,13!\,13!}$$

- Count number of ways of distributing the four aces: $4 \cdot 3 \cdot 2$
- Count number of ways of dealing the remaining 48 cards

$$\frac{48!}{12!\,12!\,12!\,12!}$$

- Answer:

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

LECTURE 5

- **Readings:** Sections 2.1-2.3, start 2.4

Lecture outline

- Random variables
- Probability mass function (PMF)
- Expectation
- Variance

Random variables

- An assignment of a value (number) to every possible outcome
- Mathematically: A function from the sample space Ω to the real numbers
  - discrete or continuous values
- Can have several random variables defined on the same sample space
- Notation:
  - random variable X
  - numerical value x

Probability mass function (PMF)

- ("probability law", "probability distribution" of X)
- Notation:

$$\begin{aligned} p_X(x) &= \mathbf{P}(X = x) \\ &= \mathbf{P}(\{\omega \in \Omega \text{ s.t. } X(\omega) = x\}) \end{aligned}$$

- $p_X(x) \geq 0 \qquad \sum_x p_X(x) = 1$
- **Example:** X=number of coin tosses until first head
  - assume independent tosses, $\mathbf{P}(H) = p > 0$

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

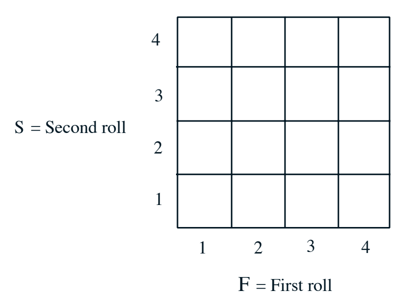
  - **geometric** PMF

How to compute a PMF $p_X(x)$

- collect all possible outcomes for which X is equal to x
- add their probabilities
- repeat for all x

- **Example:** Two independent rools of a fair tetrahedral die

  F: outcome of first throw
  S: outcome of second throw
  $X = \min(F, S)$



$p_X(2) =$

Binomial PMF

- X: number of heads in n independent coin tosses
- $\mathbf{P}(H) = p$
- Let $n = 4$

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

In general:

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

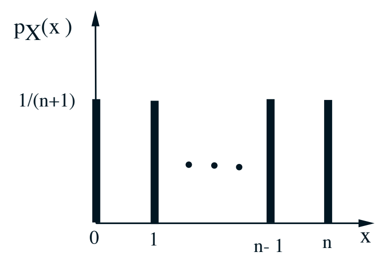
Expectation

- Definition:

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

- Interpretations:
  - Center of gravity of PMF
  - Average in large number of repetitions of the experiment (to be substantiated later in this course)

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



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

Properties of expectations

- Let X be a r.v. and let $Y = g(X)$
  - Hard: $\mathbf{E}[Y] = \sum_y y p_Y(y)$
  - Easy: $\mathbf{E}[Y] = \sum_x g(x) p_X(x)$
- Caution: In general, $\mathbf{E}[g(X)] \neq g(\mathbf{E}[X])$

Properties: If α, β are constants, then:

- $\mathbf{E}[\alpha] =$
- $\mathbf{E}[\alpha X] =$
- $\mathbf{E}[\alpha X + \beta] =$

Variance

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

- **Second moment:** $\mathbf{E}[X^2] = \sum_x x^2 p_X(x)$
- **Variance**

$$
\begin{aligned}
\text{var}(X) &= \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] \\
&= \sum_x (x - \mathbf{E}[X])^2 p_X(x) \\
&= \mathbf{E}[X^2] - (\mathbf{E}[X])^2
\end{aligned}
$$

Properties:

- $\text{var}(X) \geq 0$
- $\text{var}(\alpha X + \beta) = \alpha^2 \text{var}(X)$

LECTURE 6

- **Readings:** Sections 2.4-2.6

Lecture outline

- Review: PMF, expectation, variance
- Conditional PMF
- Geometric PMF
- Total expectation theorem
- Joint PMF of two random variables

Review

- Random variable X: function from sample space to the real numbers
- PMF (for discrete random variables): $p_X(x) = \mathbf{P}(X = x)$
- Expectation:

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

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

$$\mathbf{E}[\alpha X + \beta] = \alpha \mathbf{E}[X] + \beta$$

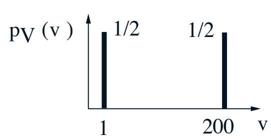
- $\mathbf{E}\Big[X - \mathbf{E}[X]\Big] =$

$$\begin{aligned} \text{var}(X) &= \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] \\ &= \sum_x (x - \mathbf{E}[X])^2 p_X(x) \\ &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \end{aligned}$$

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

Random speed

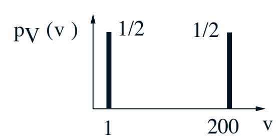
- Traverse a 200 mile distance at constant but random speed V



- $d = 200$, $T = t(V) = 200/V$
- $\mathbf{E}[V] =$
- $\text{var}(V) =$
- $\sigma_V =$

Average speed vs. average time

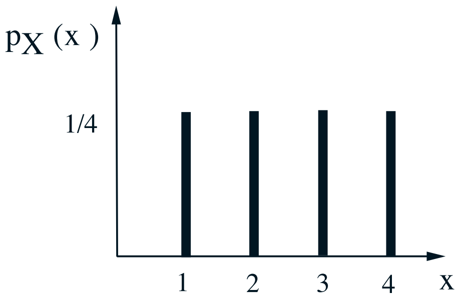
- Traverse a 200 mile distance at constant but random speed V



- time in hours $= T = t(V) =$
- $\mathbf{E}[T] = \mathbf{E}[t(V)] = \sum_v t(v) p_V(v) =$
- $\mathbf{E}[TV] = 200 \neq \mathbf{E}[T] \cdot \mathbf{E}[V]$
- $\mathbf{E}[200/V] = \mathbf{E}[T] \neq 200/\mathbf{E}[V]$.

Conditional PMF and expectation

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



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

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

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

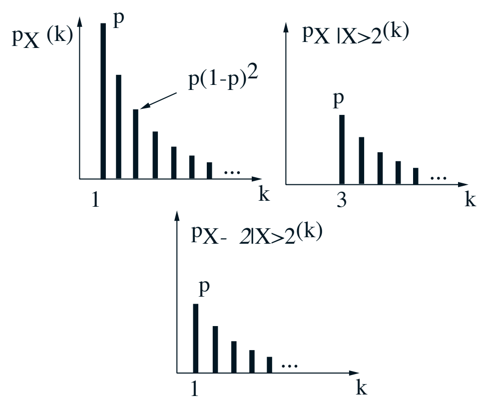
Geometric PMF

- X: number of independent coin tosses until first head

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

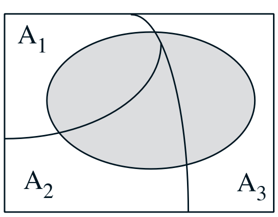
$$\mathbf{E}[X] = \sum_{k=1}^{\infty} k p_X(k) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p$$

- Memoryless property: Given that $X > 2$, the r.v. $X - 2$ has same geometric PMF



Total Expectation theorem

- Partition of sample space into disjoint events $A_1, A_2, \ldots, A_n$



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

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

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

- Geometric example:
  $A_1: \{X = 1\}, \quad A_2: \{X > 1\}$

$$\mathbf{E}[X] = \mathbf{P}(X = 1)\mathbf{E}[X \mid X = 1] + \mathbf{P}(X > 1)\mathbf{E}[X \mid X > 1]$$

- Solve to get $\mathbf{E}[X] = 1/p$

Joint PMFs

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

y \ x	1	2	3	4
4	1/20	2/20	2/20	
3	2/20	4/20	1/20	2/20
2		1/20	3/20	1/20
1		1/20		

- $\sum_x \sum_y p_{X,Y}(x, y) =$
- $p_X(x) = \sum_y p_{X,Y}(x, y)$
- $p_{X|Y}(x \mid y) = \mathbf{P}(X = x \mid Y = y) = \dfrac{p_{X,Y}(x, y)}{p_Y(y)}$
- $\sum_x p_{X|Y}(x \mid y) =$

LECTURE 7

- **Readings:** Finish Chapter 2

Lecture outline

- Multiple random variables
  - Joint PMF
  - Conditioning
  - Independence
- More on expectations
- Binomial distribution revisited
- A hat problem

Review

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

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

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

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

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

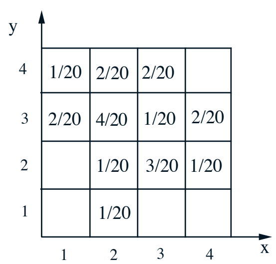
Independent random variables

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

- Random variables X, Y, Z are independent if:

  $$p_{X,Y,Z}(x, y, z) = p_X(x) \cdot p_Y(y) \cdot p_Z(z)$$

  for all x, y, z



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

Expectations

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

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

- In general: $\mathbf{E}[g(X, Y)] \neq g\big(\mathbf{E}[X], \mathbf{E}[Y]\big)$
- $\mathbf{E}[\alpha X + \beta] = \alpha \mathbf{E}[X] + \beta$
- $\mathbf{E}[X + Y + Z] = \mathbf{E}[X] + \mathbf{E}[Y] + \mathbf{E}[Z]$
- If X, Y are independent:
  - $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$
  - $\mathbf{E}[g(X)h(Y)] = \mathbf{E}[g(X)] \cdot \mathbf{E}[h(Y)]$

Variances

- $\text{Var}(aX) = a^2\text{Var}(X)$
- $\text{Var}(X + a) = \text{Var}(X)$
- Let $Z = X + Y$.
  If X, Y are independent:

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

- Examples:
  - If $X = Y$, $\text{Var}(X + Y) =$
  - If $X = -Y$, $\text{Var}(X + Y) =$
  - If X, Y indep., and $Z = X - 3Y$, $\text{Var}(Z) =$

Binomial mean and variance

- $X = \#$ of successes in n independent trials
  - probability of success p

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

- $X_i = \begin{cases} 1, & \text{if success in trial } i, \\ 0, & \text{otherwise} \end{cases}$
- $\mathbf{E}[X_i] =$
- $\mathbf{E}[X] =$
- $\text{Var}(X_i) =$
- $\text{Var}(X) =$

The hat problem

- n people throw their hats in a box and then pick one at random.
  - X: number of people who get their own hat
  - Find $\mathbf{E}[X]$

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

- $X = X_1 + X_2 + \cdots + X_n$
- $\mathbf{P}(X_i = 1) =$
- $\mathbf{E}[X_i] =$
- Are the X_i independent?
- $\mathbf{E}[X] =$

Variance in the hat problem

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

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

- $\mathbf{E}[X_i^2] =$

$$\mathbf{P}(X_1 X_2 = 1) = \mathbf{P}(X_1 = 1) \cdot \mathbf{P}(X_2 = 1 \mid X_1 = 1)$$
$$=$$

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

LECTURE 8

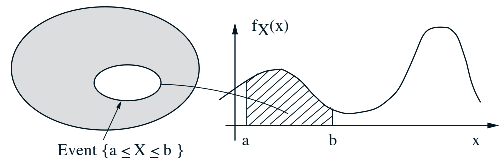
- **Readings:** Sections 3.1-3.3

Lecture outline

- Probability density functions
- Cumulative distribution functions
- Normal random variables

Continuous r.v.'s and pdf's

- **A continuous r.v.** is described by a **probability density function** f_X



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

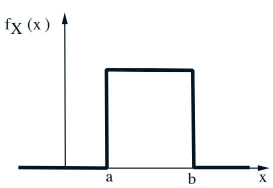
$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$$

$$\mathbf{P}(x \le X \le x+\delta) = \int_x^{x+\delta} f_X(s)\,ds \approx f_X(x)\cdot\delta$$

$$\mathbf{P}(X \in B) = \int_B f_X(x)\,dx, \quad \text{for "nice" sets } B$$

Means and variances

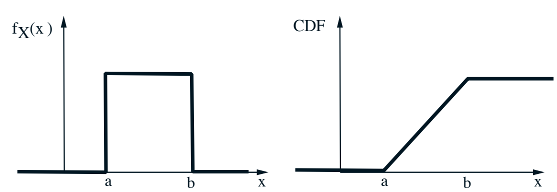
- $\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$
- $\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$
- $\text{var}(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mathbf{E}[X])^2 f_X(x)\,dx$
- **Continuous Uniform r.v.**



- $f_X(x) = \qquad a \le x \le b$
- $\mathbf{E}[X] =$
- $\sigma_X^2 = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a}\,dx = \frac{(b-a)^2}{12}$

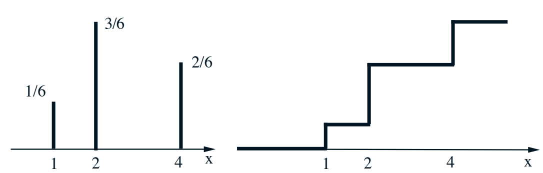
Cumulative distribution function (CDF)

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



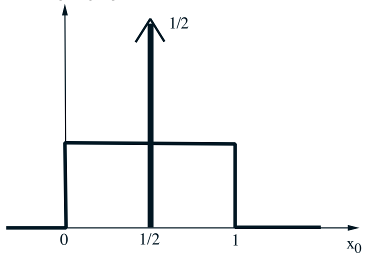
- Also for discrete r.v.'s:

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



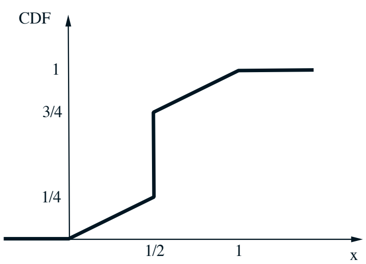
Mixed distributions

- Schematic drawing of a combination of a PDF and a PMF



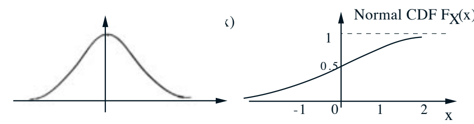
- The corresponding CDF:

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



Gaussian (normal) PDF

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



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

- General normal $N(\mu, \sigma^2)$:

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

- It turns out that:
  $\mathbf{E}[X] = \mu$ and Var$(X) = \sigma^2$.

- Let $Y = aX + b$
  - Then: $\mathbf{E}[Y] =$ $\quad$ Var$(Y) =$
  - Fact: $Y \sim N(a\mu + b, a^2\sigma^2)$

Calculating normal probabilities

- No closed form available for CDF
  - but there are tables (for standard normal)

- If $X \sim N(\mu, \sigma^2)$, then $\frac{X - \mu}{\sigma} \sim N(\quad\quad)$

- If $X \sim N(2, 16)$:

$$\mathbf{P}(X \leq 3) = \mathbf{P}\left(\frac{X-2}{4} \leq \frac{3-2}{4}\right) = \text{CDF}(0.25)$$

	.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

The constellation of concepts

$p_X(x)$ $\qquad\qquad$ $f_X(x)$

$F_X(x)$

$\mathbf{E}[X]$, var(X)

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

$p_{X|Y}(x \mid y)$ $\qquad\qquad$ $f_{X|Y}(x \mid y)$

LECTURE 9

- **Readings:** Sections 3.4-3.5

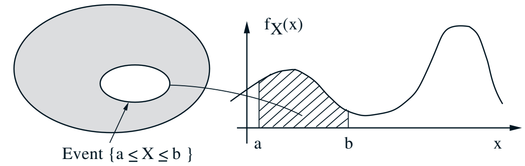
Outline

- PDF review
- Multiple random variables
  - conditioning
  - independence
- Examples

Summary of concepts

$p_X(x)$		$f_X(x)$
	$F_X(x)$	
$\sum_x x p_X(x)$	$\mathbf{E}[X]$	$\int x f_X(x)\,dx$
	var(X)	
$p_{X,Y}(x,y)$		$f_{X,Y}(x,y)$
$p_{X\|A}(x)$		$f_{X\|A}(x)$
$p_{X\|Y}(x \mid y)$		$f_{X\|Y}(x \mid y)$

Continuous r.v.'s and pdf's



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

- $\mathbf{P}(x \le X \le x+\delta) \approx f_X(x) \cdot \delta$
- $\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$

Joint PDF $f_{X,Y}(x,y)$

$$\mathbf{P}((X,Y) \in S) = \int \int_S f_{X,Y}(x,y)\,dx\,dy$$

- Interpretation:

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

- Expectations:

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

- From the joint to the marginal:

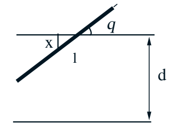
$f_X(x) \cdot \delta \approx \mathbf{P}(x \le X \le x+\delta) =$

- X and Y are called independent if

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

Buffon's needle

- Parallel lines at distance d
  Needle of length ℓ (assume $\ell < d$)
- Find $\mathbf{P}$(needle intersects one of the lines)



- $X \in [0, d/2]$: distance of needle midpoint to nearest line
- **Model:** X, Θ uniform, independent

$f_{X,\Theta}(x,\theta) = \qquad 0 \le x \le d/2,\ 0 \le \theta \le \pi/2$

- Intersect if $X \le \frac{\ell}{2} \sin \Theta$

$$\begin{aligned}
\mathbf{P}\left(X \le \frac{\ell}{2} \sin \Theta\right) &= \int \int_{x \le \frac{\ell}{2} \sin\theta} f_X(x) f_\Theta(\theta)\,dx\,d\theta \\
&= \frac{4}{\pi d} \int_0^{\pi/2} \int_0^{(\ell/2)\sin\theta} dx\,d\theta \\
&= \frac{4}{\pi d} \int_0^{\pi/2} \frac{\ell}{2} \sin\theta\,d\theta = \frac{2\ell}{\pi d}
\end{aligned}$$

Conditioning

- Recall

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

- By analogy, would like:

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

- This leads us to the **definition**:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad \text{if } f_Y(y) > 0$$

- For given y, conditional PDF is a (normalized) "section" of the joint PDF

- If independent, $f_{X,Y} = f_X f_Y$, we obtain

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

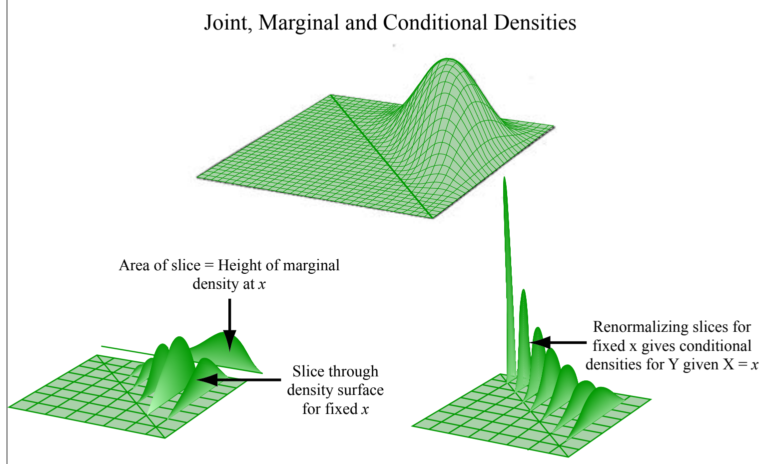
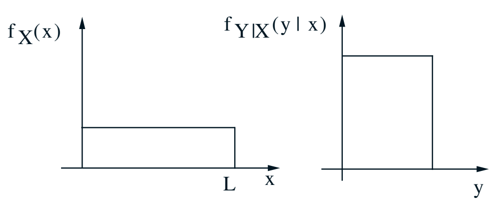


Image by MIT OpenCourseWare, adapted from *Probability*, by J. Pittman, 1999.

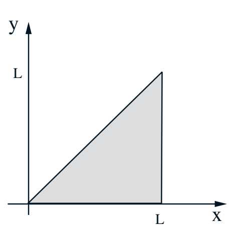
Stick-breaking example

- Break a stick of length ℓ twice:
  break at X: uniform in $[0,1]$;
  break again at Y, uniform in $[0, X]$



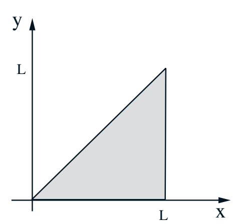
$$f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y \mid x) =$$

on the set:



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

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



$$\begin{aligned} f_Y(y) &= \int f_{X,Y}(x,y)\, dx \\ &= \int_y^{\ell} \frac{1}{\ell x}\, dx \\ &= \frac{1}{\ell} \log \frac{\ell}{y}, \qquad 0 \le y \le \ell \end{aligned}$$

$$\mathbf{E}[Y] = \int_0^{\ell} y f_Y(y)\, dy = \int_0^{\ell} y \frac{1}{\ell} \log \frac{\ell}{y}\, dy = \frac{\ell}{4}$$

LECTURE 10

Continuous Bayes rule; Derived distributions

- **Readings:**
  Section 3.6; start Section 4.1

Review

$$p_X(x) \qquad f_X(x)$$

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

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

$$p_X(x) = \sum_y p_{X,Y}(x,y) \qquad f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$

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

$$\mathbf{E}[X], \quad \text{var}(X)$$

The Bayes variations

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

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

Example:

- $X = 1, 0$: airplane present/not present
- $Y = 1, 0$: something did/did not register on radar

Continuous counterpart

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

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

Example: X: some signal; "prior" $f_X(x)$
Y: noisy version of X
$f_{Y|X}(y \mid x)$: model of the noise

Discrete X, Continuous Y

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

$$f_Y(y) = \sum_x p_X(x)f_{Y|X}(y \mid x)$$

Example:

- X: a discrete signal; "prior" $p_X(x)$
- Y: noisy version of X
- $f_{Y|X}(y \mid x)$: continuous noise model

Continuous X, Discrete Y

$$f_{X|Y}(x \mid y) = \frac{f_X(x)p_{Y|X}(y \mid x)}{p_Y(y)}$$

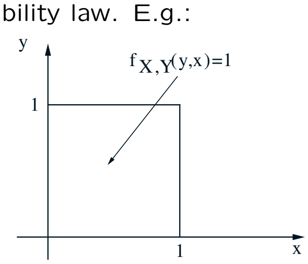
$$p_Y(y) = \int_x f_X(x)p_{Y|X}(y \mid x)\,dx$$

Example:

- X: a continuous signal; "prior" $f_X(x)$ (e.g., intensity of light beam);
- Y: discrete r.v. affected by X (e.g., photon count)
- $p_{Y|X}(y \mid x)$: model of the discrete r.v.

What is a derived distribution

- It is a PMF or PDF of a function of one or more random variables with known probability law. E.g.:



  – Obtaining the PDF for

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

  involves deriving a distribution.
  Note: $g(X,Y)$ is a random variable

When not to find them

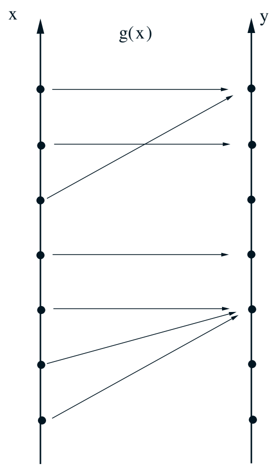
- Don't need PDF for $g(X,Y)$ if only want to compute expected value:

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

How to find them

- **Discrete case**
  - Obtain probability mass for each possible value of $Y = g(X)$

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



The continuous case

- **Two-step procedure:**
  - Get CDF of Y: $F_Y(y) = \mathbf{P}(Y \le y)$
  - Differentiate to get

$$f_Y(y) = \frac{dF_Y}{dy}(y)$$

Example

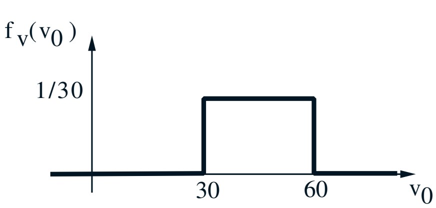
- X: uniform on [0,2]
- Find PDF of $Y = X^3$
- **Solution:**

$$\begin{aligned} F_Y(y) &= \mathbf{P}(Y \le y) = \mathbf{P}(X^3 \le y) \\ &= \mathbf{P}(X \le y^{1/3}) = \frac{1}{2} y^{1/3} \end{aligned}$$

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

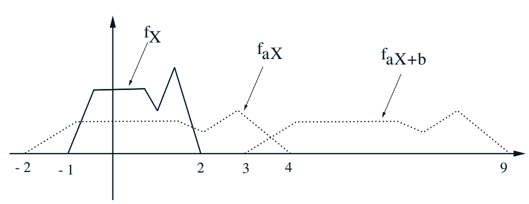
Example

- Joan is driving from Boston to New York. Her speed is uniformly distributed between 30 and 60 mph. What is the distribution of the duration of the trip?
- Let $T(V) = \dfrac{200}{V}$.
- Find $f_T(t)$



The pdf of Y=aX+b

$Y = 2X + 5$:



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

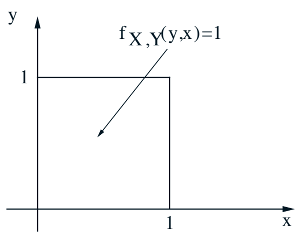
- Use this to check that if X is normal, then $Y = aX + b$ is also normal.

LECTURE 11

Derived distributions; convolution; covariance and correlation

- **Readings:**
  Finish Section 4.1;
  Section 4.2

Example



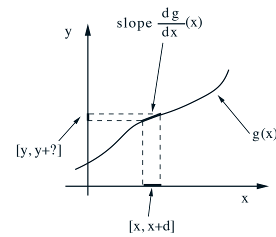
Find the PDF of $Z = g(X,Y) = Y/X$

$F_Z(z) =$ $\qquad\qquad\qquad z \le 1$

$F_Z(z) =$ $\qquad\qquad\qquad z \ge 1$

A general formula

- Let $Y = g(X)$
  g strictly monotonic.



- Event $x \le X \le x + \delta$ is the same as
  $g(x) \le Y \le g(x + \delta)$
  or (approximately)
  $g(x) \le Y \le g(x) + \delta |(dg/dx)(x)|$

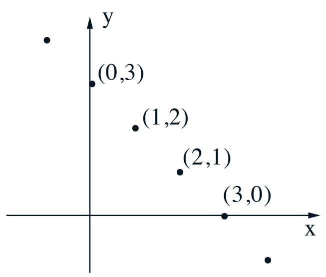
- Hence,

$$\delta f_X(x) = \delta f_Y(y) \left| \frac{dg}{dx}(x) \right|$$

where $y = g(x)$

The distribution of $X + Y$

- $W = X + Y$; X, Y independent

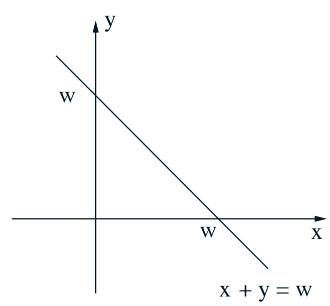


$$\begin{aligned} p_W(w) &= \mathbf{P}(X + Y = w) \\ &= \sum_x \mathbf{P}(X = x)\mathbf{P}(Y = w - x) \\ &= \sum_x p_X(x) p_Y(w - x) \end{aligned}$$

- Mechanics:
  - Put the pmf's on top of each other
  - Flip the pmf of Y
  - Shift the flipped pmf by w
    (to the right if $w > 0$)
  - Cross-multiply and add

The continuous case

- $W = X + Y$; X, Y independent



- $f_{W|X}(w \mid x) = f_Y(w - x)$

- $f_{W,X}(w, x) = f_X(x) f_{W|X}(w \mid x)$
  $\qquad\qquad = f_X(x) f_Y(w - x)$

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

Two independent normal r.v.s

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

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

- PDF is constant on the ellipse where

$$\frac{(x-\mu_x)^2}{2\sigma_x^2} + \frac{(y-\mu_y)^2}{2\sigma_y^2}$$

  is constant

- Ellipse is a circle when $\sigma_x = \sigma_y$

The sum of independent normal r.v.'s

- $X \sim N(0, \sigma_x^2)$, $Y \sim N(0, \sigma_y^2)$, independent

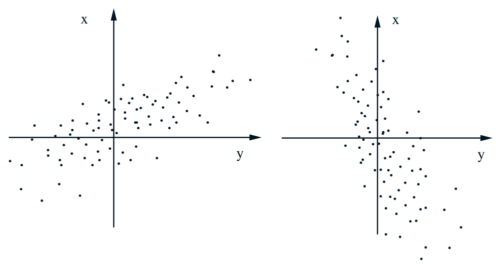
- Let $W = X + Y$

$$\begin{aligned} f_W(w) &= \int_{-\infty}^{\infty} f_X(x) f_Y(w-x)\, dx \\ &= \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} e^{-x^2/2\sigma_x^2} e^{-(w-x)^2/2\sigma_y^2}\, dx \\ \text{(algebra)} &= c e^{-\gamma w^2} \end{aligned}$$

- Conclusion: W is normal
  - mean=0, variance=$\sigma_x^2 + \sigma_y^2$
  - same argument for nonzero mean case

Covariance

- $\text{cov}(X,Y) = \mathbf{E}\Big[(X - \mathbf{E}[X]) \cdot (Y - \mathbf{E}[Y])\Big]$
- Zero-mean case: $\text{cov}(X,Y) = \mathbf{E}[XY]$



- $\text{cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$
- $\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}(X_i) + 2 \sum_{(i,j):i \neq j} \text{cov}(X_i, X_j)$
- independent $\Rightarrow$ $\text{cov}(X,Y) = 0$
  (converse is not true)

Correlation coefficient

- Dimensionless version of covariance:

$$\begin{aligned} \rho &= \mathbf{E}\left[\frac{(X - \mathbf{E}[X])}{\sigma_X} \cdot \frac{(Y - \mathbf{E}[Y])}{\sigma_Y}\right] \\ &= \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \end{aligned}$$

- $-1 \leq \rho \leq 1$
- $|\rho| = 1 \Leftrightarrow (X - \mathbf{E}[X]) = c(Y - \mathbf{E}[Y])$
  (linearly related)
- Independent $\Rightarrow$ $\rho = 0$
  (converse is not true)

LECTURE 12

- **Readings:** Section 4.3; parts of Section 4.5 (mean and variance only; no transforms)

Lecture outline

- Conditional expectation
  - Law of iterated expectations
  - Law of total variance
- Sum of a random number of independent r.v.'s
  - mean, variance

Conditional expectations

- Given the value y of a r.v. Y:

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

(integral in continuous case)

- Stick example: stick of length ℓ
  break at uniformly chosen point Y
  break again at uniformly chosen point X
- $\mathbf{E}[X \mid Y = y] = \frac{y}{2}$ (number)

$$\mathbf{E}[X \mid Y] = \frac{Y}{2} \quad \text{(r.v.)}$$

- **Law of iterated expectations:**

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

- In stick example:
  $\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X \mid Y]] = \mathbf{E}[Y/2] = \ell/4$

$\text{var}(X \mid Y)$ and its expectation

- $\text{var}(X \mid Y = y) = \mathbf{E}\left[(X - \mathbf{E}[X \mid Y = y])^2 \mid Y = y\right]$
- $\text{var}(X \mid Y)$: a r.v.
  with value $\text{var}(X \mid Y = y)$ when $Y = y$
- **Law of total variance:**

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

Proof:

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

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

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

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

Sum of right-hand sides of (c), (d):
$\mathbf{E}[X^2] - (\mathbf{E}[X])^2 = \text{var}(X)$

Section means and variances

Two sections:
$y = 1$ (10 students); $y = 2$ (20 students)

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

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

$\mathbf{E}[X \mid Y = 1] = 90, \quad \mathbf{E}[X \mid Y = 2] = 60$

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

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

$$\begin{aligned} \text{var}(\mathbf{E}[X \mid Y]) &= \frac{1}{3}(90 - 70)^2 + \frac{2}{3}(60 - 70)^2 \\ &= \frac{600}{3} = 200 \end{aligned}$$

Section means and variances (ctd.)

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

$$\text{var}(X \mid Y = 1) = 10 \qquad \text{var}(X \mid Y = 2) = 20$$

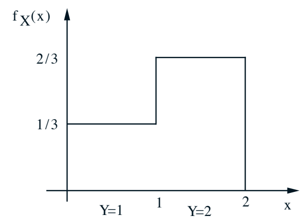
$$\text{var}(X \mid Y) = \begin{cases} 10, & \text{w.p. } 1/3 \\ 20, & \text{w.p. } 2/3 \end{cases}$$

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

$$\begin{aligned} \text{var}(X) &= \mathbf{E}[\text{var}(X \mid Y)] + \text{var}(\mathbf{E}[X \mid Y]) \\ &= \frac{50}{3} + 200 \\ &= (\text{average variability } \textbf{within} \text{ sections}) \\ &\quad + (\text{variability } \textbf{between} \text{ sections}) \end{aligned}$$

Example

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



$\mathbf{E}[X \mid Y = 1] =$ $\qquad$ $\mathbf{E}[X \mid Y = 2] =$

$\text{var}(X \mid Y = 1) =$ $\qquad$ $\text{var}(X \mid Y = 2) =$

$\mathbf{E}[X] =$

$\text{var}(\mathbf{E}[X \mid Y]) =$

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

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

$$\begin{aligned} \mathbf{E}[Y \mid N = n] &= \mathbf{E}[X_1 + X_2 + \cdots + X_n \mid N = n] \\ &= \mathbf{E}[X_1 + X_2 + \cdots + X_n] \\ &= \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n] \\ &= n\,\mathbf{E}[X] \end{aligned}$$

- $\mathbf{E}[Y \mid N] = N\,\mathbf{E}[X]$

$$\begin{aligned} \mathbf{E}[Y] &= \mathbf{E}[\mathbf{E}[Y \mid N]] \\ &= \mathbf{E}[N\,\mathbf{E}[X]] \\ &= \mathbf{E}[N]\,\mathbf{E}[X] \end{aligned}$$

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

- $\text{var}(Y) = \mathbf{E}[\text{var}(Y \mid N)] + \text{var}(\mathbf{E}[Y \mid N])$
- $\mathbf{E}[Y \mid N] = N\,\mathbf{E}[X]$
  $\text{var}(\mathbf{E}[Y \mid N]) = (\mathbf{E}[X])^2\,\text{var}(N)$
- $\text{var}(Y \mid N = n) = n\,\text{var}(X)$
  $\text{var}(Y \mid N) = N\,\text{var}(X)$
  $\mathbf{E}[\text{var}(Y \mid N)] = \mathbf{E}[N]\,\text{var}(X)$

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

LECTURE 13

The Bernoulli process

- **Readings:** Section 6.1

Lecture outline

- Definition of Bernoulli process
- Random processes
- Basic properties of Bernoulli process
- Distribution of interarrival times
- The time of the kth success
- Merging and splitting

The Bernoulli process

- A sequence of independent Bernoulli trials
- At each trial, i:
  - $\mathbf{P}(\text{success}) = \mathbf{P}(X_i = 1) = p$
  - $\mathbf{P}(\text{failure}) = \mathbf{P}(X_i = 0) = 1 - p$
- Examples:
  - Sequence of lottery wins/losses
  - Sequence of ups and downs of the Dow Jones
  - Arrivals (each second) to a bank
  - Arrivals (at each time slot) to server

Random processes

- First view: sequence of random variables $X_1, X_2, \ldots$
- $\mathbf{E}[X_t] =$
- $\text{Var}(X_t) =$
- Second view: what is the right sample space?
- $\mathbf{P}(X_t = 1 \text{ for all } t) =$
- Random processes we will study:
  - Bernoulli process (memoryless, discrete time)
  - Poisson process (memoryless, continuous time)
  - Markov chains (with memory/dependence across time)

Number of successes S in n time slots

- $\mathbf{P}(S = k) =$
- $\mathbf{E}[S] =$
- $\text{Var}(S) =$

Interarrival times

- T_1: number of trials until first success
  - $\mathbf{P}(T_1 = t) =$
  - Memoryless property
  - $\mathbf{E}[T_1] =$
  - Var$(T_1) =$

- If you buy a lottery ticket every day, what is the distribution of the length of the first string of losing days?

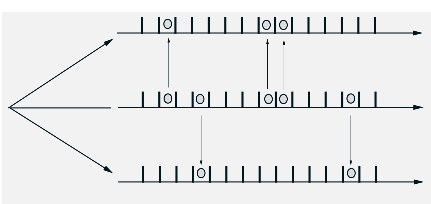
Time of the kth arrival

- Given that first arrival was at time t i.e., $T_1 = t$: additional time, T_2, until next arrival
  - has the same (geometric) distribution
  - independent of T_1

- Y_k: number of trials to kth success
  - $\mathbf{E}[Y_k] =$
  - Var$(Y_k) =$
  - $\mathbf{P}(Y_k = t) =$

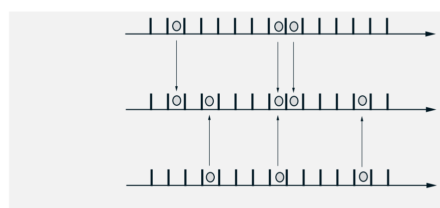
Splitting of a Bernoulli Process

(using independent coin flips)



yields Bernoulli processes

Merging of Indep. Bernoulli Processes



yields a Bernoulli process
(collisions are counted as one arrival)

LECTURE 14

The Poisson process

- **Readings:** Start Section 6.2.

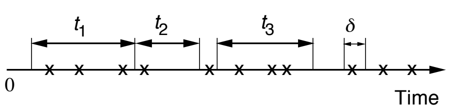
Lecture outline

- Review of Bernoulli process
- Definition of Poisson process
- Distribution of number of arrivals
- Distribution of interarrival times
- Other properties of the Poisson process

Bernoulli review

- Discrete time; success probability p
- Number of arrivals in n time slots: binomial pmf
- Interarrival times: geometric pmf
- Time to k arrivals: Pascal pmf
- Memorylessness

Definition of the Poisson process

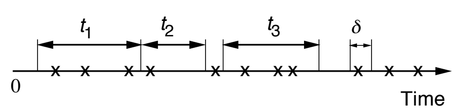


- **Time homogeneity:**
  $P(k, \tau)$ = Prob. of k arrivals in interval of duration τ
- Numbers of arrivals in disjoint time intervals are **independent**
- **Small interval probabilities:**
  For VERY small δ:

$$P(k,\delta) \approx \begin{cases} 1-\lambda\delta, & \text{if } k=0; \\ \lambda\delta, & \text{if } k=1; \\ 0, & \text{if } k>1. \end{cases}$$

  – λ: "arrival rate"

PMF of Number of Arrivals N



- Finely discretize $[0, t]$: approximately Bernoulli
- N_t (of discrete approximation): binomial
- Taking $\delta \to 0$ (or $n \to \infty$) gives:

$$P(k,\tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \qquad k = 0, 1, \ldots$$

- $\mathbf{E}[N_t] = \lambda t$, $\qquad$ $\text{var}(N_t) = \lambda t$

Example

- You get email according to a Poisson process at a rate of $\lambda = 5$ messages per hour. You check your email every thirty minutes.

- Prob(no new messages) =

- Prob(one new message) =

Interarrival Times

- Y_k time of kth arrival

- **Erlang** distribution:

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \qquad y \geq 0$$

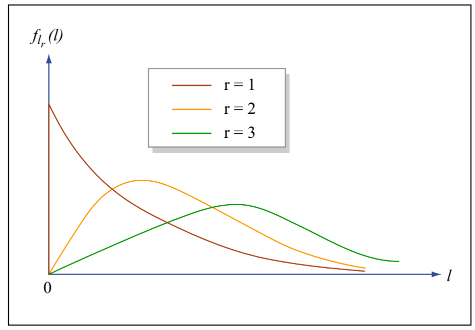
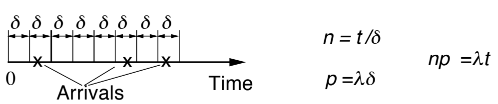


Image by MIT OpenCourseWare.

- Time of first arrival ($k = 1$): **exponential**: $f_{Y_1}(y) = \lambda e^{-\lambda y}, \quad y \geq 0$
  - Memoryless property: The time to the next arrival is independent of the past

Bernoulli/Poisson Relation

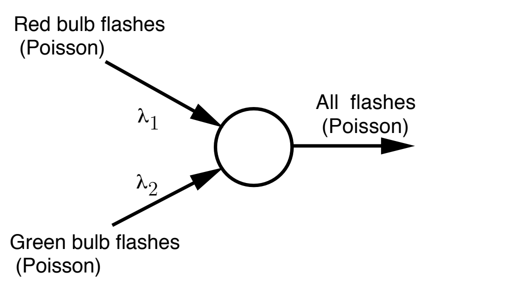


	POISSON	BERNOULLI
Times of Arrival	Continuous	Discrete
Arrival Rate	λ/unit time	p/per trial
PMF of # of Arrivals	Poisson	Binomial
Interarrival Time Distr.	Exponential	Geometric
Time to k-th arrival	Erlang	Pascal

Merging Poisson Processes

- Sum of independent Poisson **random variables** is Poisson

- Merging of independent Poisson **processes** is Poisson



  - What is the probability that the next arrival comes from the first process?

LECTURE 15

Poisson process — II

- **Readings:** Finish Section 6.2.

- Review of Poisson process

- Merging and splitting

- Examples

- Random incidence

Review

- Defining characteristics
  - **Time homogeneity**: $P(k, \tau)$
  - **Independence**
  - **Small interval probabilities** (small δ):

$$P(k,\delta) \approx \begin{cases} 1-\lambda\delta, & \text{if } k=0, \\ \lambda\delta, & \text{if } k=1, \\ 0, & \text{if } k>1. \end{cases}$$

- N_τ is a Poisson r.v., with parameter $\lambda\tau$:

$$P(k,\tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \qquad k = 0, 1, \dots$$

$\mathbf{E}[N_\tau] = \text{var}(N_\tau) = \lambda\tau$

- Interarrival times ($k = 1$): exponential:

$$f_{T_1}(t) = \lambda e^{-\lambda t}, \quad t \geq 0, \qquad \mathbf{E}[T_1] = 1/\lambda$$

- Time Y_k to kth arrival: Erlang(k):

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \qquad y \geq 0$$

Poisson fishing

- Assume: Poisson, $\lambda = 0.6$/hour.
  - Fish for two hours.
  - if no catch, continue until first catch.

a) $\mathbf{P}$(fish for more than two hours)=

b) $\mathbf{P}$(fish for more than two and less than five hours)=

c) $\mathbf{P}$(catch at least two fish)=

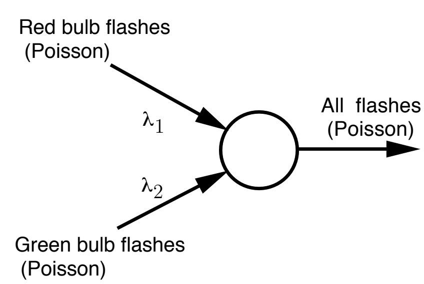
d) $\mathbf{E}$[number of fish]=

e) $\mathbf{E}$[future fishing time | fished for four hours]=

f) $\mathbf{E}$[total fishing time]=

Merging Poisson Processes (again)

- Merging of independent Poisson **processes** is Poisson



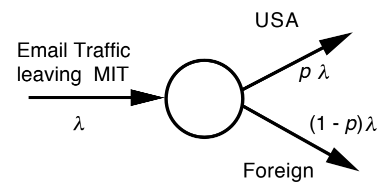
  - What is the probability that the next arrival comes from the first process?

Light bulb example

- Each light bulb has independent, exponential(λ) lifetime
- Install three light bulbs.
  Find expected time until last light bulb dies out.

Splitting of Poisson processes

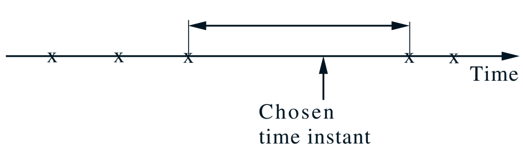
- Assume that email traffic through a server is a Poisson process.
  Destinations of different messages are independent.



- Each output stream is Poisson.

Random incidence for Poisson

- Poisson process that has been running forever
- Show up at some "random time" (really means "arbitrary time")



- What is the distribution of the length of the chosen interarrival interval?

Random incidence in "renewal processes"

- Series of successive arrivals
  - i.i.d. interarrival times
    (but not necessarily exponential)
- **Example:**
  Bus interarrival times are equally likely to be 5 or 10 minutes
- If you arrive at a "random time":
  - what is the probability that you selected a 5 minute interarrival interval?
  - what is the expected time to next arrival?

LECTURE 16

Markov Processes – I

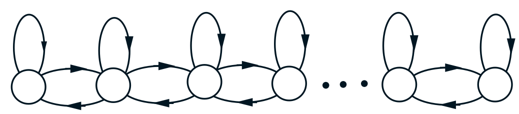
- **Readings:** Sections 7.1–7.2

Lecture outline

- Checkout counter example
- Markov process definition
- n-step transition probabilities
- Classification of states

Checkout counter model

- Discrete time $n = 0, 1, \ldots$
- Customer arrivals: Bernoulli(p)
  - geometric interarrival times
- Customer service times: geometric(q)
- "State" X_n: number of customers at time n



Finite state Markov chains

- X_n: state after n transitions
  - belongs to a finite set, e.g., $\{1, \ldots, m\}$
  - X_0 is either given or random
- **Markov property/assumption:** (given current state, the past does not matter)

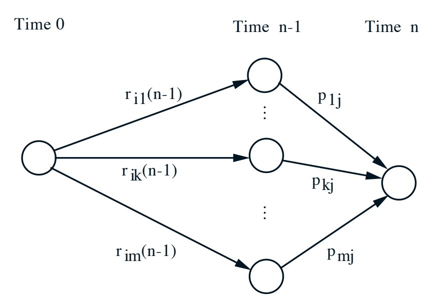
$$\begin{aligned} p_{ij} &= \mathbf{P}(X_{n+1} = j \mid X_n = i) \\ &= \mathbf{P}(X_{n+1} = j \mid X_n = i,\ X_{n-1}, \ldots, X_0) \end{aligned}$$

- Model specification:
  - identify the possible states
  - identify the possible transitions
  - identify the transition probabilities

n-step transition probabilities

- State occupancy probabilities, given initial state i:

$$r_{ij}(n) = \mathbf{P}(X_n = j \mid X_0 = i)$$



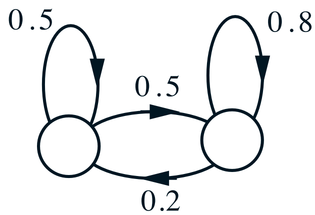
- Key recursion:

$$r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1) p_{kj}$$

- With random initial state:

$$\mathbf{P}(X_n = j) = \sum_{i=1}^{m} \mathbf{P}(X_0 = i) r_{ij}(n)$$

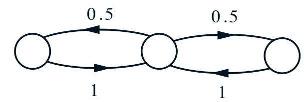
Example



	$n=0$	$n=1$	$n=2$	$n=100$	$n=101$
$r_{11}(n)$					
$r_{12}(n)$					
$r_{21}(n)$					
$r_{22}(n)$					

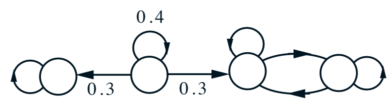
Generic convergence questions:

- Does $r_{ij}(n)$ converge to something?



n odd: $r_{22}(n)=$ n even: $r_{22}(n)=$

- Does the limit depend on initial state?



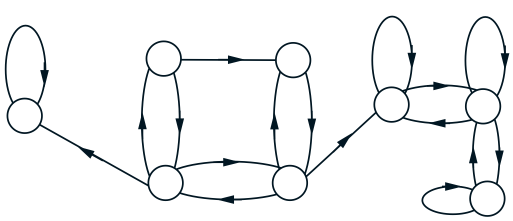
$r_{11}(n)=$

$r_{31}(n)=$

$r_{21}(n)=$

Recurrent and transient states

- State i is **recurrent** if:
  starting from i,
  and from wherever you can go,
  there is a way of returning to i
- If not recurrent, called **transient**



  - i transient:
    $\mathbf{P}(X_n = i) \to 0$,
    i visited finite number of times

- **Recurrent class**:
  collection of recurrent states that
  "communicate" with each other
  and with no other state

LECTURE 17

Markov Processes – II

- **Readings:** Section 7.3

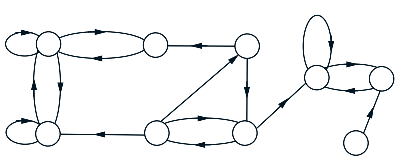
Lecture outline

- Review
- Steady-State behavior
  - Steady-state convergence theorem
  - Balance equations
- Birth-death processes

Review

- Discrete state, discrete time, time-homogeneous
  - Transition probabilities p_{ij}
  - Markov property
- $r_{ij}(n) = \mathbf{P}(X_n = j \mid X_0 = i)$
- Key recursion:
  $r_{ij}(n) = \sum_k r_{ik}(n-1)p_{kj}$

Warmup



$\mathbf{P}(X_1 = 2, X_2 = 6, X_3 = 7 \mid X_0 = 1) =$

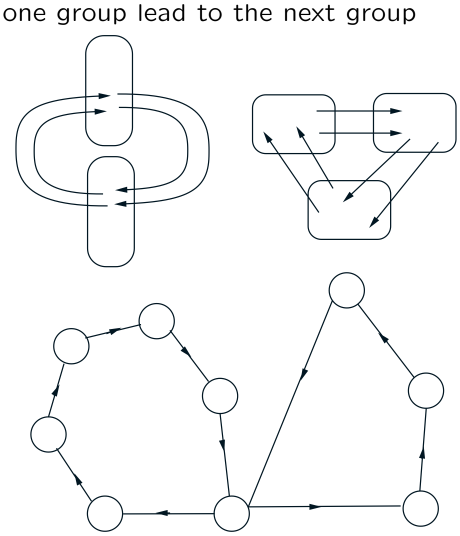
$\mathbf{P}(X_4 = 7 \mid X_0 = 2) =$

Recurrent and transient states

- State i is **recurrent** if:
  starting from i,
  and from wherever you can go,
  there is a way of returning to i
- If not recurrent, called **transient**
- **Recurrent class**:
  collection of recurrent states that
  "communicate" to each other
  and to no other state

Periodic states

- The states in a recurrent class are **periodic** if they can be grouped into $d > 1$ groups so that all transitions from one group lead to the next group



Steady-State Probabilities

- Do the $r_{ij}(n)$ converge to some π_j? (independent of the initial state i)
- Yes, if:
  - recurrent states are all in a single class, and
  - single recurrent class is not periodic
- Assuming "yes," start from key recursion

$$r_{ij}(n) = \sum_k r_{ik}(n-1)p_{kj}$$

  - take the limit as $n \to \infty$

$$\pi_j = \sum_k \pi_k p_{kj}, \qquad \text{for all } j$$

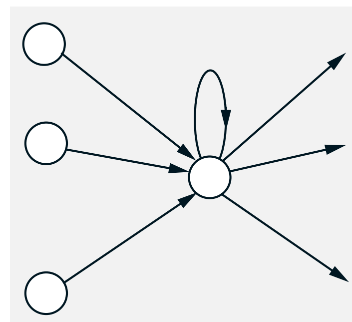
  - Additional equation:

$$\sum_j \pi_j = 1$$

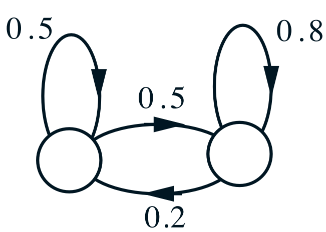
Visit frequency interpretation

$$\pi_j = \sum_k \pi_k p_{kj}$$

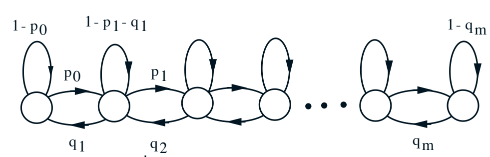
- (Long run) frequency of being in j: π_j
- Frequency of transitions $k \to j$: $\pi_k p_{kj}$
- Frequency of transitions into j: $\sum_k \pi_k p_{kj}$

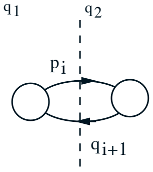


Example



Birth-death processes





$$\pi_i p_i = \pi_{i+1} q_{i+1}$$

- Special case: $p_i = p$ and $q_i = q$ for all i
  $\rho = p/q$ =load factor

$$\pi_{i+1} = \pi_i \frac{p}{q} = \pi_i \rho$$

$$\pi_i = \pi_0 \rho^i, \qquad i = 0, 1, \ldots, m$$

- Assume $p < q$ and $m \approx \infty$

$\pi_0 = 1 - \rho$

$\mathbf{E}[X_n] = \dfrac{\rho}{1-\rho}$ (in steady-state)

LECTURE 18

Markov Processes – III

Readings: Section 7.4

Lecture outline

- Review of steady-state behavior
- Probability of blocked phone calls
- Calculating absorption probabilities
- Calculating expected time to absorption

Review

- Assume a single class of recurrent states, aperiodic;
  plus transient states. Then,

$$\lim_{n\to\infty} r_{ij}(n) = \pi_j$$

where π_j does not depend on the initial conditions:

$$\lim_{n\to\infty} \mathbf{P}(X_n = j \mid X_0 = i) = \pi_j$$

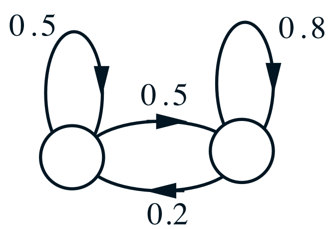
- $\pi_1, \ldots, \pi_m$ can be found as the unique solution to the balance equations

$$\pi_j = \sum_k \pi_k p_{kj}, \qquad j = 1, \ldots, m,$$

together with

$$\sum_j \pi_j = 1$$

Example

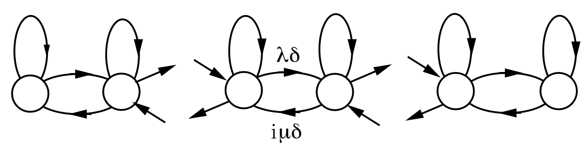


$\pi_1 = 2/7$, $\pi_2 = 5/7$

- Assume process starts at state 1.
- $\mathbf{P}(X_1 = 1,$ and $X_{100} = 1)$=
- $\mathbf{P}(X_{100} = 1$ and $X_{101} = 2)$

The phone company problem

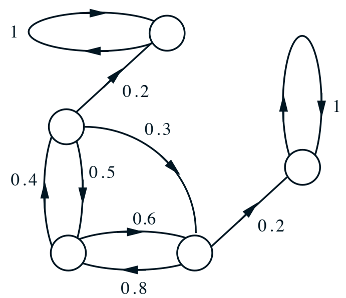
- Calls originate as a Poisson process, rate λ
  - Each call duration is exponentially distributed (parameter μ)
  - B lines available
- Discrete time intervals of (small) length δ



- Balance equations: $\lambda\pi_{i-1} = i\mu\pi_i$
- $\pi_i = \pi_0 \dfrac{\lambda^i}{\mu^i i!}$ $\qquad \pi_0 = 1 / \sum_{i=0}^{B} \dfrac{\lambda^i}{\mu^i i!}$

Calculating absorption probabilities

- What is the probability a_i that: process eventually settles in state 4, given that the initial state is i?

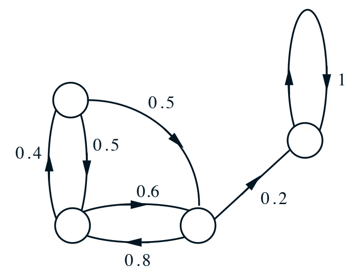


For $i = 4$, $a_i =$
For $i = 5$, $a_i =$

$$a_i = \sum_j p_{ij} a_j, \quad \text{for all other } i$$

– unique solution

Expected time to absorption



- Find expected number of transitions μ_i, until reaching the absorbing state, given that the initial state is i?

$\mu_i = 0$ for $i =$

For all other i: $\mu_i = 1 + \sum_j p_{ij} \mu_j$

– unique solution

Mean first passage and recurrence times

- Chain with one recurrent class; fix s recurrent
- **Mean first passage time from i to s:**

  $t_i = \mathbf{E}[\min\{n \geq 0 \text{ such that } X_n = s\} \mid X_0 = i]$

- $t_1, t_2, \ldots, t_m$ are the unique solution to

$$\begin{aligned} t_s &= 0, \\ t_i &= 1 + \sum_j p_{ij} t_j, \qquad \text{for all } i \neq s \end{aligned}$$

- **Mean recurrence time of s:**

  $t_s^* = \mathbf{E}[\min\{n \geq 1 \text{ such that } X_n = s\} \mid X_0 = s]$

- $t_s^* = 1 + \sum_j p_{sj} t_j$

LECTURE 19
Limit theorems – I

- **Readings:** Sections 5.1-5.3; start Section 5.4

- $X_1, \ldots, X_n$ i.i.d.
$$M_n = \frac{X_1 + \cdots + X_n}{n}$$
What happens as $n \to \infty$?

- Why bother?

- A tool: Chebyshev's inequality

- Convergence "in probability"

- Convergence of M_n (weak law of large numbers)

Chebyshev's inequality

- Random variable X (with finite mean μ and variance σ^2)

$$\sigma^2 = \int (x-\mu)^2 f_X(x)\,dx$$
$$\geq \int_{-\infty}^{-c} (x-\mu)^2 f_X(x)\,dx + \int_c^{\infty} (x-\mu)^2 f_X(x)\,dx$$
$$\geq c^2 \cdot \mathbf{P}(|X-\mu| \geq c)$$

$$\mathbf{P}(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

$$\mathbf{P}(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Deterministic limits

- Sequence a_n
Number a

- a_n converges to a
$$\lim_{n\to\infty} a_n = a$$
"a_n eventually gets and stays (arbitrarily) close to a"

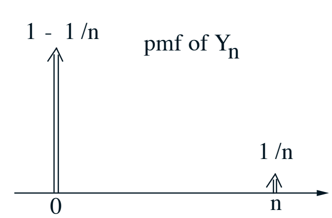
- For every $\epsilon > 0$,
there exists n_0,
such that for every $n \geq n_0$,
we have $|a_n - a| \leq \epsilon$.

Convergence "in probability"

- Sequence of random variables Y_n

- converges in probability to a number a:
"(almost all) of the PMF/PDF of Y_n , eventually gets concentrated (arbitrarily) close to a"

- For every $\epsilon > 0$,
$$\lim_{n\to\infty} \mathbf{P}(|Y_n - a| \geq \epsilon) = 0$$



Does Y_n converge?

Convergence of the sample mean
(Weak law of large numbers)

- $X_1, X_2, \ldots$ i.i.d.
  finite mean μ and variance σ^2

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

- $\mathbf{E}[M_n] =$

- $\mathrm{Var}(M_n) =$

$$\mathbf{P}(|M_n - \mu| \geq \epsilon) \;\leq\; \frac{\mathrm{Var}(M_n)}{\epsilon^2} \;=\; \frac{\sigma^2}{n\epsilon^2}$$

- M_n converges in probability to μ

The pollster's problem

- f: fraction of population that "..."
- ith (randomly selected) person polled:

$$X_i = \begin{cases} 1, & \text{if yes,} \\ 0, & \text{if no.} \end{cases}$$

- $M_n = (X_1 + \cdots + X_n)/n$
  fraction of "yes" in our sample
- Goal: 95% confidence of ≤1% error

$$\mathbf{P}(|M_n - f| \geq .01) \leq .05$$

- Use Chebyshev's inequality:

$$\mathbf{P}(|M_n - f| \geq .01) \;\leq\; \frac{\sigma^2_{M_n}}{(0.01)^2}$$
$$= \frac{\sigma^2_x}{n(0.01)^2} \leq \frac{1}{4n(0.01)^2}$$

- If $n = 50,000$,
  then $\mathbf{P}(|M_n - f| \geq .01) \leq .05$
  (conservative)

Different scalings of M_n

- $X_1, \ldots, X_n$ i.i.d.
  finite variance σ^2
- Look at three variants of their sum:
- $S_n = X_1 + \cdots + X_n$ variance $n\sigma^2$
- $M_n = \dfrac{S_n}{n}$ variance σ^2/n
  converges "in probability" to $\mathbf{E}[X]$ (WLLN)
- $\dfrac{S_n}{\sqrt{n}}$ constant variance σ^2
  - Asymptotic shape?

The central limit theorem

- "Standardized" $S_n = X_1 + \cdots + X_n$:

$$Z_n = \frac{S_n - \mathbf{E}[S_n]}{\sigma_{S_n}} = \frac{S_n - n\mathbf{E}[X]}{\sqrt{n}\,\sigma}$$

  - zero mean
  - unit variance
- Let Z be a standard normal r.v.
  (zero mean, unit variance)
- **Theorem:** For every c:

$$\mathbf{P}(Z_n \leq c) \to \mathbf{P}(Z \leq c)$$

- $\mathbf{P}(Z \leq c)$ is the standard normal CDF, $\Phi(c)$, available from the normal tables

LECTURE 20
THE CENTRAL LIMIT THEOREM

- Readings: Section 5.4
- $X_1, \ldots, X_n$ i.i.d., finite variance σ^2
- "Standardized" $S_n = X_1 + \cdots + X_n$:

$$Z_n = \frac{S_n - \mathbf{E}[S_n]}{\sigma_{S_n}} = \frac{S_n - n\mathbf{E}[X]}{\sqrt{n}\sigma}$$

  – $\mathbf{E}[Z_n] = 0$, $\quad \text{var}(Z_n) = 1$

- Let Z be a standard normal r.v. (zero mean, unit variance)
- **Theorem:** For every c:

$$\mathbf{P}(Z_n \le c) \to \mathbf{P}(Z \le c)$$

- $\mathbf{P}(Z \le c)$ is the standard normal CDF, $\Phi(c)$, available from the normal tables

Usefulness

- universal; only means, variances matter
- accurate computational shortcut
- justification of normal models

What exactly does it say?

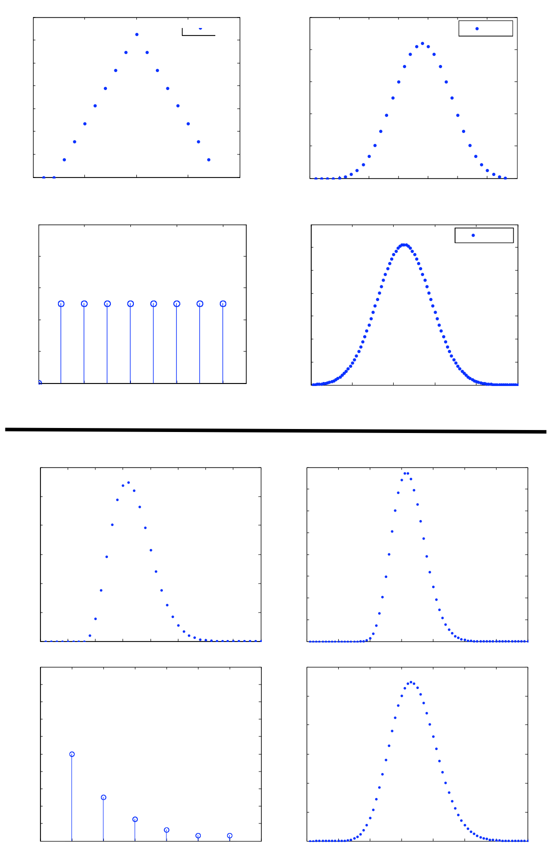
- CDF of Z_n converges to normal CDF
  – not a statement about convergence of PDFs or PMFs

Normal approximation

- Treat Z_n as if normal
  – also treat S_n as if normal

Can we use it when n is "moderate"?

- Yes, but no nice theorems to this effect
- Symmetry helps a lot



The pollster's problem using the CLT

- f: fraction of population that "..."
- ith (randomly selected) person polled:

$$X_i = \begin{cases} 1, & \text{if yes,} \\ 0, & \text{if no.} \end{cases}$$

- $M_n = (X_1 + \cdots + X_n)/n$
- Suppose we want:

$$\mathbf{P}(|M_n - f| \ge .01) \le .05$$

- Event of interest: $|M_n - f| \ge .01$

$$\left| \frac{X_1 + \cdots + X_n - nf}{n} \right| \ge .01$$

$$\left| \frac{X_1 + \cdots + X_n - nf}{\sqrt{n}\sigma} \right| \ge \frac{.01\sqrt{n}}{\sigma}$$

$$\begin{aligned} \mathbf{P}(|M_n - f| \ge .01) &\approx \mathbf{P}(|Z| \ge .01\sqrt{n}/\sigma) \\ &\le \mathbf{P}(|Z| \ge .02\sqrt{n}) \end{aligned}$$

Apply to binomial

- Fix p, where $0 < p < 1$
- X_i: Bernoulli(p)
- $S_n = X_1 + \cdots + X_n$: Binomial(n, p)
  - mean np, variance $np(1-p)$
- CDF of $\dfrac{S_n - np}{\sqrt{np(1-p)}} \longrightarrow$ standard normal

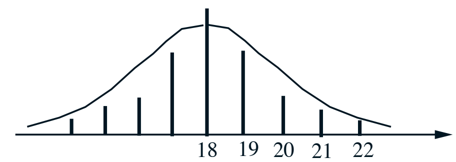
Example

- $n = 36$, $p = 0.5$; find $\mathbf{P}(S_n \le 21)$
- Exact answer:

$$\sum_{k=0}^{21} \binom{36}{k} \left(\frac{1}{2}\right)^{36} = 0.8785$$

The 1/2 correction for binomial approximation

- $\mathbf{P}(S_n \le 21) = \mathbf{P}(S_n < 22)$, because S_n is integer
- Compromise: consider $\mathbf{P}(S_n \le 21.5)$



De Moivre–Laplace CLT (for binomial)

- When the 1/2 correction is used, CLT can also approximate the binomial p.m.f. (not just the binomial CDF)

$$\mathbf{P}(S_n = 19) = \mathbf{P}(18.5 \le S_n \le 19.5)$$

$$18.5 \le S_n \le 19.5 \iff$$

$$\frac{18.5 - 18}{3} \le \frac{S_n - 18}{3} \le \frac{19.5 - 18}{3} \iff$$

$$0.17 \le Z_n \le 0.5$$

$$\begin{aligned} \mathbf{P}(S_n = 19) &\approx \mathbf{P}(0.17 \le Z \le 0.5) \\ &= \mathbf{P}(Z \le 0.5) - \mathbf{P}(Z \le 0.17) \\ &= 0.6915 - 0.5675 \\ &= 0.124 \end{aligned}$$

- Exact answer:

$$\binom{36}{19} \left(\frac{1}{2}\right)^{36} = 0.1251$$

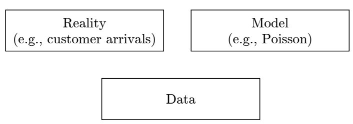
Poisson vs. normal approximations of the binomial

- Poisson arrivals during unit interval equals: sum of n (independent) Poisson arrivals during n intervals of length $1/n$
  - Let $n \to \infty$, apply CLT (??)
  - Poisson=normal (????)
- Binomial(n, p)
  - p fixed, $n \to \infty$: normal
  - np fixed, $n \to \infty$, $p \to 0$: Poisson
- $p = 1/100$, $n = 100$: Poisson
- $p = 1/10$, $n = 500$: normal

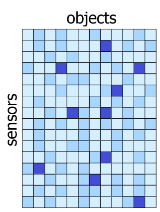
LECTURE 21

- **Readings:** Sections 8.1-8.2

"It is the mark of truly educated people to be deeply moved by **statistics**." (Oscar Wilde)



- Design & interpretation of experiments
  - polling, medical/pharmaceutical trials...
- Netflix competition
- Finance



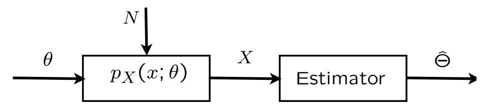
Graph of S&P 500 index removed due to copyright restrictions.

- Signal processing
  - Tracking, detection, speaker identification,...

Types of Inference models/approaches

- Model building versus inferring unknown variables. E.g., assume $X = aS + W$
  - Model building: know "signal" S, observe X, infer a
  - Estimation in the presence of noise: know a, observe X, estimate S.
- **Hypothesis testing:** unknown takes one of few possible values; aim at small probability of incorrect decision
- **Estimation:** aim at a small estimation error

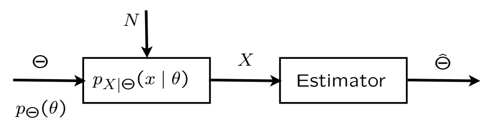
- **Classical statistics:**



θ: unknown parameter (not a r.v.)

- E.g., θ = mass of electron

- **Bayesian:** Use priors & Bayes rule



Bayesian inference: Use Bayes rule

- **Hypothesis testing**
  - discrete data

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, p_{X|\Theta}(x \mid \theta)}{p_X(x)}$$

  - continuous data

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

- **Estimation;** continuous data

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$\begin{aligned} Z_t &= \Theta_0 + t\Theta_1 + t^2\Theta_2 \\ X_t &= Z_t + W_t, \qquad t = 1, 2, \ldots, n \end{aligned}$$

Bayes rule gives:

$$f_{\Theta_0,\Theta_1,\Theta_2|X_1,\ldots,X_n}(\theta_0, \theta_1, \theta_2 \mid x_1, \ldots, x_n)$$

Estimation with discrete data

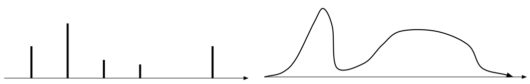
$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, p_{X|\Theta}(x \mid \theta)}{p_X(x)}$$

$$p_X(x) = \int f_\Theta(\theta) p_{X|\Theta}(x \mid \theta)\, d\theta$$

- **Example:**
  - Coin with unknown parameter θ
  - Observe X heads in n tosses
- What is the Bayesian approach?
  - Want to find $f_{\Theta|X}(\theta \mid x)$
  - Assume a prior on Θ (e.g., uniform)

Output of Bayesian Inference

- Posterior distribution:
  - pmf $p_{\Theta|X}(\cdot \mid x)$ or pdf $f_{\Theta|X}(\cdot \mid x)$



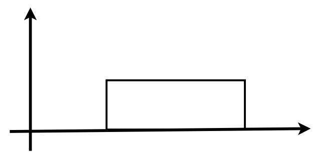
- If interested in a single answer:
  - Maximum a posteriori probability (MAP):
    - $p_{\Theta|X}(\theta^* \mid x) = \max_\theta p_{\Theta|X}(\theta \mid x)$
      minimizes probability of error;
      often used in hypothesis testing
    - $f_{\Theta|X}(\theta^* \mid x) = \max_\theta f_{\Theta|X}(\theta \mid x)$
  - Conditional expectation:

$$\mathbf{E}[\Theta \mid X = y] = \int \theta f_{\Theta|X}(\theta \mid x)\, d\theta$$

  - Single answers can be misleading!

Least Mean Squares Estimation

- Estimation in the absence of information



- find estimate c, to:

$$\text{minimize } \mathbf{E}\left[(\Theta - c)^2\right]$$

- Optimal estimate: $c = \mathbf{E}[\Theta]$
- Optimal mean squared error:

$$\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta])^2\right] = \text{Var}(\Theta)$$

LMS Estimation of Θ based on X

- Two r.v.'s Θ, X
- we observe that $X = x$
  - new universe: condition on $X = x$
- $\mathbf{E}\left[(\Theta - c)^2 \mid X = x\right]$ is minimized by
  $c =$
- $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X = x])^2 \mid X = x\right]$
  $\leq \mathbf{E}[(\Theta - g(x))^2 \mid X = x]$

◦ $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X])^2 \mid X\right] \leq \mathbf{E}\left[(\Theta - g(X))^2 \mid X\right]$

◦ $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X])^2\right] \leq \mathbf{E}\left[(\Theta - g(X))^2\right]$

> $\mathbf{E}[\Theta \mid X]$ minimizes $\mathbf{E}\left[(\Theta - g(X))^2\right]$
> over all estimators $g(\cdot)$

LMS Estimation w. several measurements

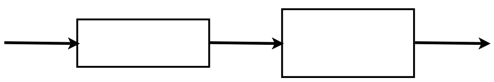
- Unknown r.v. Θ
- Observe values of r.v.'s $X_1, \ldots, X_n$
- Best estimator: $\mathbf{E}[\Theta \mid X_1, \ldots, X_n]$
- Can be hard to compute/implement
  - involves multi-dimensional integrals, etc.

LECTURE 22

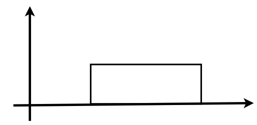
- **Readings:** pp. 225-226; Sections 8.3-8.4

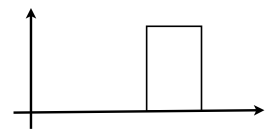
Topics

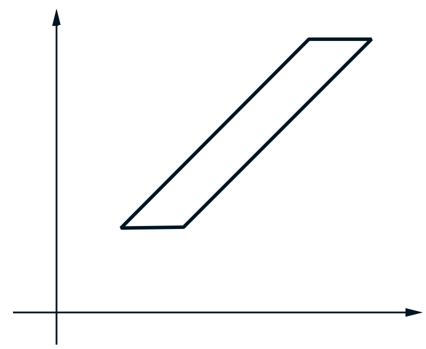
- (Bayesian) Least means squares (LMS) estimation
- (Bayesian) Linear LMS estimation



- MAP estimate: $\hat{\theta}_{\text{MAP}}$ maximizes $f_{\Theta|X}(\theta \mid x)$
- LMS estimation:
  - $\hat{\Theta} = \mathbf{E}[\Theta \mid X]$ minimizes $\mathbf{E}\big[(\Theta - g(X))^2\big]$ over all estimators $g(\cdot)$
  - for any x, $\hat{\theta} = \mathbf{E}[\Theta \mid X = x]$ minimizes $\mathbf{E}\big[(\Theta - \hat{\theta})^2 \mid X = x\big]$ over all estimates $\hat{\theta}$

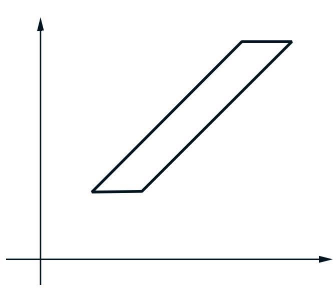


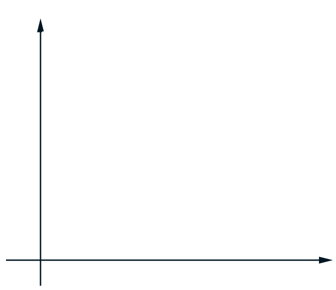




Conditional mean squared error

- $\mathbf{E}[(\Theta - \mathbf{E}[\Theta \mid X])^2 \mid X = x]$
  - same as Var$(\Theta \mid X = x)$: variance of the conditional distribution of Θ





Some properties of LMS estimation

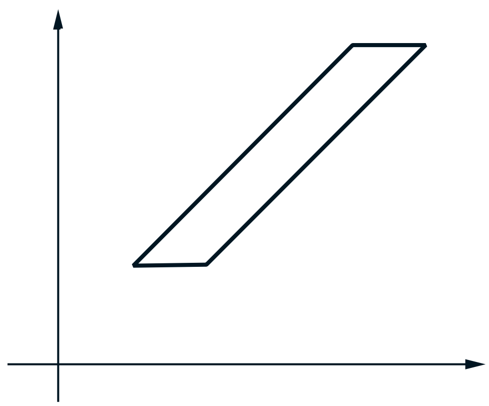
- Estimator: $\hat{\Theta} = \mathbf{E}[\Theta \mid X]$
- Estimation error: $\tilde{\Theta} = \hat{\Theta} - \Theta$

- $\mathbf{E}[\tilde{\Theta}] = 0$ $\qquad$ $\mathbf{E}[\tilde{\Theta} \mid X = x] = 0$
- $\mathbf{E}[\tilde{\Theta} h(X)] = 0$, for any function h
- cov$(\tilde{\Theta}, \hat{\Theta}) = 0$
- Since $\Theta = \hat{\Theta} - \tilde{\Theta}$:
  var$(\Theta) = $ var$(\hat{\Theta}) +$ var$(\tilde{\Theta})$

Linear LMS

- Consider estimators of Θ,
  of the form $\hat{\Theta} = aX + b$
- Minimize $\mathbf{E}\left[(\Theta - aX - b)^2\right]$
- Best choice of a,b; best linear estimator:

$$\hat{\Theta}_L = \mathbf{E}[\Theta] + \frac{\text{Cov}(X, \Theta)}{\text{var}(X)}(X - \mathbf{E}[X])$$



Linear LMS properties

$$\hat{\Theta}_L = \mathbf{E}[\Theta] + \frac{\text{Cov}(X, \Theta)}{\text{var}(X)}(X - \mathbf{E}[X])$$

$$\mathbf{E}[(\hat{\Theta}_L - \Theta)^2] = (1 - \rho^2)\sigma_\Theta^2$$

Linear LMS with multiple data

- Consider estimators of the form:

$$\hat{\Theta} = a_1 X_1 + \cdots + a_n X_n + b$$

- Find best choices of $a_1, \ldots, a_n, b$
- Minimize:

$$\mathbf{E}[(a_1 X_1 + \cdots + a_n X_n + b - \Theta)^2]$$

- Set derivatives to zero
  linear system in b and the a_i
- Only means, variances, covariances matter

The cleanest linear LMS example

$X_i = \Theta + W_i, \qquad \Theta, W_1, \ldots, W_n$ independent
$\Theta \ \sim\ \mu,\ \sigma_0^2 \qquad W_i \ \sim\ 0,\ \sigma_i^2$

$$\hat{\Theta}_L = \frac{\mu/\sigma_0^2 + \sum_{i=1}^{n} X_i/\sigma_i^2}{\sum_{i=0}^{n} 1/\sigma_i^2}$$

(weighted average of $\mu, X_1, \ldots, X_n$)

- If all normal, $\hat{\Theta}_L = \mathbf{E}[\Theta \mid X_1, \ldots, X_n]$

Choosing X_i in linear LMS

- $\mathbf{E}[\Theta \mid X]$ is the same as $\mathbf{E}[\Theta \mid X^3]$
- Linear LMS is different:
  - $\hat{\Theta} = aX + b$ versus $\hat{\Theta} = aX^3 + b$
  - Also consider $\hat{\Theta} = a_1 X + a_2 X^2 + a_3 X^3 + b$

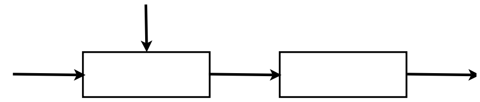
Big picture

- **Standard examples:**
  - X_i uniform on $[0, \theta]$;
    uniform prior on θ
  - X_i Bernoulli(p);
    uniform (or Beta) prior on p
  - X_i normal with mean θ, known variance σ^2;
    normal prior on θ;
    $X_i = \Theta + W_i$
- **Estimation methods:**
  - MAP
  - MSE
  - Linear MSE

LECTURE 23

- **Readings:** Section 9.1 (not responsible for t-based confidence intervals, in pp. 471-473)
- **Outline**
  - Classical statistics
  - Maximum likelihood (ML) estimation
  - Estimating a sample mean
  - Confidence intervals (CIs)
  - CIs using an estimated variance

Classical statistics



- also for vectors X and θ: $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)$
- These are NOT conditional probabilities; θ is NOT random
  - mathematically: many models, one for each possible value of θ
- **Problem types:**
  - 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

Maximum Likelihood Estimation

- Model, with unknown parameter(s): $X \sim p_X(x;\theta)$
- Pick θ that "makes data most likely"

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta} p_X(x;\theta)$$

- Compare to Bayesian MAP estimation:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p_{\Theta|X}(\theta \mid x)$$

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \frac{p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{p_X(x)}$$

- **Example:** $X_1,\ldots,X_n$: i.i.d., exponential(θ)

$$\max_{\theta} \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

$$\max_{\theta} \left(n\log\theta - \theta\sum_{i=1}^{n} x_i\right)$$

$$\hat{\theta}_{\text{ML}} = \frac{n}{x_1+\cdots+x_n} \qquad \hat{\Theta}_n = \frac{n}{X_1+\cdots+X_n}$$

Desirable properties of estimators (should hold FOR ALL θ !!!)

- **Unbiased:** $\mathbf{E}[\hat{\Theta}_n] = \theta$
  - exponential example, with $n=1$: $\mathbf{E}[1/X_1] = \infty \neq \theta$ (biased)
- **Consistent:** $\hat{\Theta}_n \to \theta$ (in probability)
  - exponential example: $(X_1+\cdots+X_n)/n \to \mathbf{E}[X] = 1/\theta$
  - can use this to show that: $\hat{\Theta}_n = n/(X_1+\cdots+X_n) \to 1/\mathbf{E}[X] = \theta$
- **"Small" mean squared error (MSE)**

$$\mathbf{E}[(\hat{\Theta}-\theta)^2] = \text{var}(\hat{\Theta}-\theta) + (\mathbf{E}[\hat{\Theta}-\theta])^2 = \text{var}(\hat{\Theta}) + (\text{bias})^2$$

Estimate a mean

- $X_1, \ldots, X_n$: i.i.d., mean θ, variance σ^2

$X_i = \theta + W_i$

W_i: i.i.d., mean, 0, variance σ^2

$$\hat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

Properties:

- $\mathbf{E}[\hat{\Theta}_n] = \theta$ (unbiased)
- WLLN: $\hat{\Theta}_n \to \theta$ (consistency)
- MSE: σ^2/n

- Sample mean often turns out to also be the ML estimate.
  E.g., if $X_i \sim N(\theta, \sigma^2)$, i.i.d.

Confidence intervals (CIs)

- An estimate $\hat{\Theta}_n$ may not be informative enough
- An $1-\alpha$ **confidence interval** is a (random) interval $[\hat{\Theta}_n^-, \hat{\Theta}_n^+]$,

  s.t. $\mathbf{P}(\hat{\Theta}_n^- \le \theta \le \hat{\Theta}_n^+) \ge 1-\alpha, \quad \forall\ \theta$

  - often $\alpha = 0.05$, or 0.25, or 0.01
  - interpretation is subtle

- CI in estimation of the mean
  $\hat{\Theta}_n = (X_1 + \cdots + X_n)/n$
  - normal tables: $\Phi(1.96) = 1 - 0.05/2$

$$\mathbf{P}\left(\frac{|\hat{\Theta}_n - \theta|}{\sigma/\sqrt{n}} \le 1.96\right) \approx 0.95 \quad \text{(CLT)}$$

$$\mathbf{P}\left(\hat{\Theta}_n - \frac{1.96\,\sigma}{\sqrt{n}} \le \theta \le \hat{\Theta}_n + \frac{1.96\,\sigma}{\sqrt{n}}\right) \approx 0.95$$

More generally: let z be s.t. $\Phi(z) = 1 - \alpha/2$

$$\mathbf{P}\left(\hat{\Theta}_n - \frac{z\sigma}{\sqrt{n}} \le \theta \le \hat{\Theta}_n + \frac{z\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

The case of unknown σ

- Option 1: use upper bound on σ
  - if X_i Bernoulli: $\sigma \le 1/2$
- Option 2: use ad hoc estimate of σ
  - if X_i Bernoulli(θ): $\hat{\sigma} = \sqrt{\hat{\Theta}(1-\hat{\Theta})}$
- Option 3: Use generic estimate of the variance
  - Start from $\sigma^2 = \mathbf{E}[(X_i - \theta)^2]$

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 \ \to\ \sigma^2$$

(but do not know θ)

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\Theta}_n)^2 \ \to\ \sigma^2$$

(unbiased: $\mathbf{E}[\hat{S}_n^2] = \sigma^2$)

LECTURE 24

- **Reference:** Section 9.3
- Course Evaluations (until 12/16)
  http://web.mit.edu/subjectevaluation

Outline

- Review
  - Maximum likelihood estimation
  - Confidence intervals
- Linear regression
- Binary hypothesis testing
  - Types of error
  - Likelihood ratio test (LRT)

Review

- **Maximum likelihood estimation**
  - Have model with unknown parameters: $X \sim p_X(x;\theta)$
  - Pick θ that "makes data most likely"

$$\max_\theta p_X(x;\theta)$$

  - Compare to Bayesian MAP estimation:

$$\max_\theta p_{\Theta|X}(\theta \mid x) \quad \text{or} \quad \max_\theta \frac{p_{X|\Theta}(x|\theta)p_\Theta(\theta)}{p_Y(y)}$$

- **Sample mean estimate of** $\theta = \mathbf{E}[X]$

$$\hat{\Theta}_n = (X_1 + \cdots + X_n)/n$$

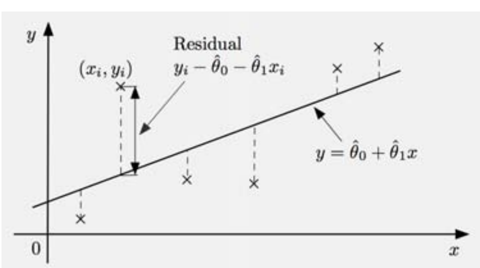
- $1-\alpha$ **confidence interval**

$$\mathbf{P}(\hat{\Theta}_n^- \le \theta \le \hat{\Theta}_n^+) \ge 1-\alpha, \quad \forall\ \theta$$

- confidence interval for sample mean
  - let z be s.t. $\Phi(z) = 1-\alpha/2$

$$\mathbf{P}\left(\hat{\Theta}_n - \frac{z\sigma}{\sqrt{n}} \le \theta \le \hat{\Theta}_n + \frac{z\sigma}{\sqrt{n}}\right) \approx 1-\alpha$$

Regression



- Data: $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$
- Model: $y \approx \theta_0 + \theta_1 x$

$$\min_{\theta_0,\theta_1} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2 \qquad (*)$$

- One interpretation:
  $Y_i = \theta_0 + \theta_1 x_i + W_i, \quad W_i \sim N(0,\sigma^2)$, i.i.d.
  - Likelihood function $f_{X,Y|\theta}(x,y;\theta)$ is:

$$c \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2\right\}$$

  - Take logs, same as (*)
  - Least sq. $\leftrightarrow$ pretend W_i i.i.d. normal

Linear regression

- **Model** $y \approx \theta_0 + \theta_1 x$

$$\min_{\theta_0,\theta_1} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2$$

- **Solution** (set derivatives to zero):

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}, \qquad \overline{y} = \frac{y_1 + \cdots + y_n}{n}$$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$$\hat{\theta}_0 = \overline{y} - \hat{\theta}_1 \overline{x}$$

- **Interpretation** of the form of the solution
  - Assume a model $Y = \theta_0 + \theta_1 X + W$
    W independent of X, with zero mean
  - Check that

$$\theta_1 = \frac{\text{cov}(X,Y)}{\text{var}(X)} = \frac{\mathbf{E}\Big[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\Big]}{\mathbf{E}\Big[(X - \mathbf{E}[X])^2\Big]}$$

  - Solution formula for $\hat{\theta}_1$ uses natural estimates of the variance and covariance

The world of linear regression

- **Multiple linear regression:**
  - **data:** (x_i, x'_i, x''_i, y_i), $i = 1, \ldots, n$
  - **model:** $y \approx \theta_0 + \theta x + \theta' x' + \theta'' x''$
  - **formulation:**

$$\min_{\theta,\theta',\theta''} \sum_{i=1}^{n} (y_i - \theta_0 - \theta x_i - \theta' x'_i - \theta'' x''_i)^2$$

- **Choosing the right variables**
  - model $y \approx \theta_0 + \theta_1 h(x)$
    e.g., $y \approx \theta_0 + \theta_1 x^2$
  - work with data points $(y_i, h(x))$
  - formulation:

$$\min_{\theta} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 h_1(x_i))^2$$

The world of regression (ctd.)

- **In practice,** one also reports
  - Confidence intervals for the θ_i
  - "Standard error" (estimate of σ)
  - R^2, a measure of "explanatory power"

- **Some common concerns**
  - Heteroskedasticity
  - Multicollinearity
  - Sometimes misused to conclude causal relations
  - etc.

Binary hypothesis testing

- Binary θ; new terminology:
  - **null hypothesis** H_0:
    $X \sim p_X(x; H_0)$ [or $f_X(x; H_0)$]
  - **alternative hypothesis** H_1:
    $X \sim p_X(x; H_1)$ [or $f_X(x; H_1)$]

- Partition the space of possible data vectors
  Rejection region R:
  reject H_0 iff data $\in R$

- Types of errors:
  - **Type I** (**false rejection**, false alarm):
    H_0 true, but rejected

$$\alpha(R) = \mathbf{P}(X \in R\,; H_0)$$

  - **Type II** (**false acceptance**, missed detection):
    H_0 false, but accepted

$$\beta(R) = \mathbf{P}(X \notin R\,; H_1)$$

Likelihood ratio test (LRT)

- Bayesian case (MAP rule): choose H_1 if:
  $\mathbf{P}(H_1 \mid X = x) > \mathbf{P}(H_0 \mid X = x)$
  or

$$\frac{\mathbf{P}(X = x \mid H_1)\mathbf{P}(H_1)}{\mathbf{P}(X = x)} > \frac{\mathbf{P}(X = x \mid H_0)\mathbf{P}(H_0)}{\mathbf{P}(X = x)}$$

  or

$$\frac{\mathbf{P}(X = x \mid H_1)}{\mathbf{P}(X = x \mid H_0)} > \frac{\mathbf{P}(H_0)}{\mathbf{P}(H_1)}$$

  (likelihood ratio test)

- Nonbayesian version: choose H_1 if

$$\frac{\mathbf{P}(X = x; H_1)}{\mathbf{P}(X = x; H_0)} > \xi \quad \text{(discrete case)}$$

$$\frac{f_X(x; H_1)}{f_X(x; H_0)} > \xi \qquad \text{(continuous case)}$$

- threshold ξ trades off the two types of error
  - choose ξ so that $\mathbf{P}(\text{reject } H_0; H_0) = \alpha$ (e.g., $\alpha = 0.05$)

LECTURE 25
Outline

- **Reference:** Section 9.4
- Course Evaluations (until 12/16)
  http://web.mit.edu/subjectevaluation
- Review of simple binary hypothesis tests
  - examples
- Testing composite hypotheses
  - is my coin fair?
  - is my die fair?
  - goodness of fit tests

Simple binary hypothesis testing

- **null hypothesis** H_0:
  $X \sim p_X(x; H_0)$ [or $f_X(x; H_0)$]
- **alternative hypothesis** H_1:
  $X \sim p_X(x; H_1)$ [or $f_X(x; H_1)$]
- Choose a **rejection region** R;
  reject H_0 iff data $\in R$

- Likelihood ratio test: reject H_0 if

$$\frac{p_X(x; H_1)}{p_X(x; H_0)} > \xi \quad \text{or} \quad \frac{f_X(x; H_1)}{f_X(x; H_0)} > \xi$$

  - fix false rejection probability α (e.g., $\alpha = 0.05$)
  - choose ξ so that $\mathbf{P}(\text{reject } H_0; H_0) = \alpha$

Example (test on normal mean)

- n data points, i.i.d.
  H_0: $X_i \sim N(0, 1)$
  H_1: $X_i \sim N(1, 1)$
- Likelihood ratio test; rejection region:

$$\frac{(1/\sqrt{2\pi})^n \exp\{-\sum_i (X_i - 1)^2/2\}}{(1/\sqrt{2\pi})^n \exp\{-\sum_i X_i^2/2\}} > \xi$$

  - algebra: reject H_0 if: $\sum_i X_i > \xi'$
- Find ξ' such that

$$\mathbf{P}\Big(\sum_{i=1}^{n} X_i > \xi'; H_0\Big) = \alpha$$

  - use normal tables

Example (test on normal variance)

- n data points, i.i.d.
  H_0: $X_i \sim N(0, 1)$
  H_1: $X_i \sim N(0, 4)$
- Likelihood ratio test; rejection region:

$$\frac{(1/2\sqrt{2\pi})^n \exp\{-\sum_i X_i^2/(2 \cdot 4)\}}{(1/\sqrt{2\pi})^n \exp\{-\sum_i X_i^2/2\}} > \xi$$

  - algebra: reject H_0 if $\sum_i X_i^2 > \xi'$
- Find ξ' such that

$$\mathbf{P}\Big(\sum_{i=1}^{n} X_i^2 > \xi'; H_0\Big) = \alpha$$

  - the distribution of $\sum_i X_i^2$ is known (derived distribution problem)
  - "chi-square" distribution; tables are available

Composite hypotheses

- Got $S = 472$ heads in $n = 1000$ tosses; is the coin fair?
  - $H_0 : p = 1/2$ versus $H_1 : p \neq 1/2$
- Pick a **"statistic"** (e.g., S)
- Pick shape of **rejection region** (e.g., $|S - n/2| > \xi$)
- Choose **significance level** (e.g., $\alpha = 0.05$)
- Pick **critical value** ξ so that:

$$\mathbf{P}(\text{reject } H_0; H_0) = \alpha$$

  Using the CLT:

$$\mathbf{P}(|S - 500| \leq 31; H_0) \approx 0.95; \quad \xi = 31$$

- In our example: $|S - 500| = 28 < \xi$
  H_0 **not rejected** (at the 5% level)

Is my die fair?

- Hypothesis H_0:
  $\mathbf{P}(X = i) = p_i = 1/6, \ i = 1, \ldots, 6$
- Observed occurrences of i: N_i
- Choose form of rejection region; chi-square test:

$$\text{reject } H_0 \text{ if } T = \sum_i \frac{(N_i - np_i)^2}{np_i} > \xi$$

- Choose ξ so that:

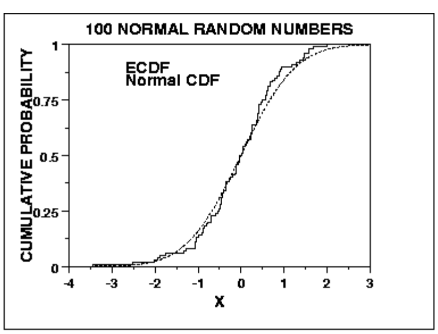
$$\mathbf{P}(\text{reject } H_0; H_0) = 0.05$$

$$\mathbf{P}(T > \xi; H_0) = 0.05$$

- Need the distribution of T:
  (CLT + derived distribution problem)
  - for large n, T has approximately a chi-square distribution
  - available in tables

Do I have the correct pdf?

- Partition the range into bins
  - np_i: expected incidence of bin i (from the pdf)
  - N_i: observed incidence of bin i
  - Use chi-square test (as in die problem)
- Kolmogorov-Smirnov test:
  form **empirical CDF**, $\hat{F}_X$, from data



(http://www.itl.nist.gov/div898/handbook/)

- $D_n = \max_x |F_X(x) - \hat{F}_X(x)|$
- $\mathbf{P}(\sqrt{n}D_n \geq 1.36) \approx 0.05$

What else is there?

- Systematic methods for coming up with shape of rejection regions
- Methods to estimate an unknown PDF (e.g., form a histogram and "smooth" it out)
- Efficient and recursive signal processing
- Methods to select between less or more complex models
  - (e.g., identify relevant "explanatory variables" in regression models)
- Methods tailored to high-dimensional unknown parameter vectors and huge number of data points (data mining)
- etc. etc. . . .