What is Probability?

- Measured relative frequency of occurrence of an event.
  Example: toss a coin 100 times, measure frequency of heads or compute probability of raining on a particular day and month (using past years' data)
- Or subjective belief about how "likely" an event is (when do not have data to estimate frequency).
  Example: any one-time event in history or "how likely is it that a new experimental drug will work?"
  This may either be a subjective belief or derived from the physics, for e.g. if I flip a symmetric coin (equal weight on both sides), I will get a head with probability $1/2$.
- For probabilistic reasoning, **two** types of problems need to be solved

1. Specify the probability “model” or learn it (covered in a statistics class).
2. Use the “model” to compute probability of different events (covered here).

- We will assume the model is given and will focus on problem 2. in this course.

Set Theory Basics

- Set: any collection of objects (elements of a set).
- Discrete sets
  - Finite number of elements, e.g. numbers of a die
  - Or infinite but countable number of elements, e.g. set of integers
- Continuous sets
  - Cannot count the number of elements, e.g. all real numbers between 0 and 1.
- "Universe" (denoted Ω): consists of all possible elements that could be of interest. In case of random experiments, it is the set of all possible outcomes. Example: for coin tosses, $\Omega = \{H, T\}$.
- Empty set (denoted ϕ): a set with no elements

- Subset: $A \subseteq B$: if every element of A also belongs to B.
- Strict subset: $A \subset B$: if every element of A also belongs to B and B has more elements than A.
- Belongs: $\in$, Does not belong: $\notin$
- Complement: A' or A^c, Union: $A \cup B$, Intersection: $A \cap B$
  - $A' \triangleq \{x \in \Omega | x \notin A\}$
  - $A \cup B \triangleq \{x | x \in A, \ or \ x \in B\}$, $x \in \Omega$ is assumed.
  - $A \cap B \triangleq \{x | x \in A, \ and \ x \in B\}$
  - Visualize using Venn diagrams (see book)
- **Disjoint sets: A and B are disjoint if $A \cap B = \phi$ (empty), i.e. they have no common elements.**

- DeMorgan's Laws

$$(A \cup B)' = A' \cap B' \quad (1)$$

$$(A \cap B)' = A' \cup B' \quad (2)$$

  - Proofs: Need to show that every element of LHS (left hand side) is also an element of RHS (right hand side), i.e. LHS $\subseteq$ RHS and show vice versa, i.e. RHS $\subseteq$ LHS.
  - We show the proof of the first property
    * If $x \in (A \cup B)'$, it means that x does not belong to A or B. In other words x does not belong to A and x does not B either. This means x belongs to the complement of A and to the complement of B, i.e. $x \in A' \cap B'$.
    * Just showing this much does not complete the proof, need to show the other side also.
    * If $x \in A' \cap B'$, it means that x does not belong to A and it does not

belong to B, i.e. it belongs to neither A nor B, i.e. $x \in (A \cup B)'$

  * This completes the argument

– Please read the section on Algebra of Sets, pg 5

Probabilistic models

- There is an underlying process called **experiment** that produces exactly ONE **outcome**.
- A probabilistic model: consists of a sample space and a probability law
  - Sample space (denoted Ω): set of all possible outcomes of an experiment
  - Event: any subset of the sample space
  - Probability Law: assigns a probability to every set A of possible outcomes (event)
  - Choice of sample space (or universe): every element should be distinct and mutually exclusive (disjoint); and the space should be "collectively exhaustive" (every possible outcome of an experiment should be included).

- **Probability Axioms:**
  1. **Nonnegativity.** $P(A) \geq 0$ for every event A.
  2. **Additivity.** If A and B are two **disjoint** events, then
     $P(A \cup B) = P(A) + P(B)$
     (also extends to any countable number of disjoint events).
  3. **Normalization.** Probability of the entire sample space, $P(\Omega) = 1$.
- Probability of the empty set, $P(\phi) = 0$ (follows from Axioms 2 & 3).
- Sequential models, e.g. three coin tosses or two sequential rolls of a die. Tree-based description: see Fig. 1.3
- Discrete probability law: sample space consists of a finite number of possible outcomes, law specified by probability of single element events.
  - Example: for a fair coin toss, $\Omega = \{H, T\}$, $P(H) = P(T) = 1/2$
  - Discrete uniform law for any event A:

$$P(A) = \frac{\text{number of elements in A}}{n}$$

- Continuous probability law: e.g. $\Omega = [0, 1]$: probability of any single element event is zero, need to talk of probability of a subinterval, $[a, b]$ of $[0, 1]$.
  See Example 1.4, 1.5 (This is slightly more difficult. We will cover continuous probability and examples later).

- Properties of probability laws
  1. If $A \subseteq B$, then $P(A) \leq P(B)$
  2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  3. $P(A \cup B) \leq P(A) + P(B)$
  4. $P(A \cup B \cup C) = P(A) + P(A' \cap B) + P(A' \cap B' \cap C)$
  5. Note: book uses A^c for A' (complement of set A).
  6. Proofs: Will be covered in next class. Visualize: Venn diagrams.

Conditional Probability

- Given that we know that an event B has occurred, what is the probability that event A occurred? Denoted by $P(A|B)$. Example: Roll of a 6-sided die. Given that the outcome is even, what is the probability of a 6? Answer: 1/3

- When number of outcomes is finite and all are equally likely,

$$P(A|B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B} \tag{3}$$

- In general,

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)} \tag{4}$$

- $P(A|B)$ is a probability law (satisfies axioms) on the universe B. Exercise: show this.

- Examples/applications
  - Example 1.7, 1.8, 1.11
  - Construct sequential models: $P(A \cap B) = P(B)P(A|B)$. Example: Radar detection (Example 1.9). What is the probability of the aircraft not present and radar registers it (false alarm)?
  - See Fig. 1.9: Tree based sequential description

Total Probability and Bayes Rule

- Total Probability Theorem: Let $A_1, \ldots A_n$ be disjoint events which form a partition of the sample space ($\cup_{i=1}^{n} A_i = \Omega$). Then for any event B,

$$\begin{aligned} P(B) &= P(A_1 \cap B) + \ldots P(A_n \cap B) \\ &= P(A_1)P(B|A_1) + \ldots P(A_n)P(B|A_n) \end{aligned} \quad (5)$$

  Visualization and proof: see Fig. 1.13

- Example 1.13, 1.15

- Bayes rule: Let $A_1, \ldots A_n$ be disjoint events which form a partition of the sample space. Then for any event B, s.t. $P(B) > 0$, we have

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + \ldots P(A_n)P(B|A_n)} \quad (6)$$

- Inference using Bayes rule
  - There are multiple "causes" $A_1, A_2, ..A_n$ that result in a certain "effect" B. Given that we observe the effect B, what is the probability that the cause was A_i? Answer: use Bayes rule. See Fig. 1.14
  - Radar detection: what is the probability of the aircraft being present given that the radar registers it? Example 1.16
  - False positive puzzle, Example 1.18: very interesting!

Independence

- $P(A|B) = P(A)$ and so $P(A \cap B) = P(B)P(A)$: the fact that B has occurred gives no information about the probability of occurrence of A. Example: A= head in first coin toss, B = head in second coin toss.
- **"Independence": DIFFERENT from "mutually exclusive" (disjoint)**
  - Events A and B are disjoint if $P(A \cap B) = 0$: cannot be independent if $P(A) > 0$ and $P(B) > 0$.
    Example: A = head in a coin toss, B = tail in a coin toss
  - Independence: a concept for events in a sequence. Independent events with $P(A) > 0, P(B) > 0$ cannot be disjoint
- Conditional independence **
- Independence of a collection of events

- $P(\cap_{i \in S} A_i) = \Pi_{i \in S} P(A_i)$ for every subset S of $\{1, 2, ..n\}$

- Reliability analysis of complex systems: independence assumption often simplifies calculations
  - Analyze Fig. 1.15: what is $P(\text{system fails})$ of the system $A \rightarrow B$?
    * Let p_i = probability of success of component i.
    * m components in series: $P(\text{system fails}) = 1 - p_1 p_2 \ldots p_m$ (succeeds if all components succeed).
    * m components in parallel: $P(\text{system fails}) = (1 - p_1) \ldots (1 - p_m)$ (fails if all the components fail).

- Independent Bernoulli trials and Binomial probabilities
  - A Bernoulli trial: a coin toss (or any experiment with two possible outcomes, e.g. it rains or does not rain, bit values)
  - Independent Bernoulli trials: sequence of independent coin tosses

- Binomial: Given n independent coin tosses,what is the probability of k heads (denoted $p(k)$)?
  - probability of any one sequence with k heads is $p^k(1-p)^{n-k}$
  - number of such sequences (from counting arguments): $\left(\begin{array}{c} n \\ k \end{array} \right)$
  - $p(k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k}$, where $\left(\begin{array}{c} n \\ k \end{array} \right) \triangleq \frac{n!}{(n-k)!k!}$
- Application: what is the probability that more than c customers need an internet connection at a given time? We know that at a given time, the probability that any one customer needs connection is p.

  Answer: $\displaystyle\sum_{k=c+1}^{n} p(k)$

Counting

- Needed in many situations. Two examples are:
  1. Sample space has a finite number of equally likely outcomes (discrete uniform), compute probability of any event A.
  2. Or compute the probability of an event A which consists of a finite number of equally likely outcomes each with probability p, e.g. probability of k heads in n coin tosses.
- Counting principle (See Fig. 1.17): Consider a process consisting of r stages. If at stage 1, there are n_1 possibilities, at stage 2, n_2 possibilities and so on, then the total number of possibilities = $n_1 n_2 \ldots n_r$.
  - Example 1.26 (number of possible telephone numbers)
  - Counting principle applies even when second stage depends on the first stage and so on, Ex. 1.28 (no. of words with 4 distinct letters)

- Applications: k-permutations.
  - n distinct objects, how many different ways can we pick k objects and arrange them in a sequence?
    * Use counting principle: choose first object in n possible ways, second one in $n-1$ ways and so on. Total no. of ways: $n(n-1)\ldots(n-k+1) = \frac{n!}{(n-k)!}$
    * If $k = n$, then total no. of ways = $n!$
    * Example 1.28, 1.29
- Applications: k-combinations.
  - Choice of k elements out of an n-element set without regard to order.
  - Most common example: There are n people, how many different ways can we form a committee of k people? Here order of choosing the k members is not important. Denote answer by $\binom{n}{k}$
  - Note that selecting a k-permutation is the same as first selecting a

k-combination and then ordering the elements (in $k!$) different ways, i.e. $\frac{n!}{(n-k)!} = \binom{n}{k} k!$

- Thus $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.
- How will you relate this to the binomial coefficient (number of ways to get k heads out of n tosses)?

  Toss number j = person j, a head in a toss = the person (toss number) is in committee

- Applications: k-partitions. **
  - A combination is a partition of a set into two parts
  - Partition: given an n-element set, consider its partition into r subsets of size $n_1, n_2, \ldots, n_r$ where $n_1 + n_2 + \ldots n_r = n$.
    - $*$ Use counting principle and k-combinations result.
    - $*$ Form the first subset. Choose n_1 elements out of n: $\binom{n}{n_1}$ ways.
    - $*$ Form second subset. Choose n_2 elements out of $n - n_1$ available

elements: $\begin{pmatrix} n - n_1 \\ n_2 \end{pmatrix}$ and so on.

* Total number of ways to form the partition:

$$\begin{pmatrix} n \\ n_1 \end{pmatrix} \begin{pmatrix} n - n_1 \\ n_2 \end{pmatrix} \ldots \begin{pmatrix} (n - n_1 - n_2 \ldots n_{r-1}) \\ n_r \end{pmatrix} = \frac{n!}{n_1! n_2! \ldots n_r!}$$

Note: Handouts DO NOT replace the book. In most cases, they only provide a guideline on topics and an intuitive feel.

1 What is a random variable (r.v.)?

- A real valued function of the outcome of an experiment
- Example: Coin tosses. r.v. $X = 1$ if heads and $X = 0$ if tails (Bernoulli r.v.).
- A function of a r.v. defines another r.v.
- Discrete r.v.: X takes values from the set of integers

2 Discrete Random Variables & Probability Mass Function (PMF)

- **Probability Mass Function (PMF):** Probability that the r.v. X takes a value x is PMF of X computed at $X = x$. Denoted by $p_X(x)$. Thus

$$p_X(x) = P(\{X = x\}) = P(\text{all possible outcomes that result in the event } \{X = x\}) \quad (1)$$

- Everything that we learnt in Chap 1 for events applies. Let Ω is the sample space (space of all possible values of X in an experiment). Applying the axioms,
  - $p_X(x) \geq 0$
  - $P(\{X \in S\}) = \sum_{x \in S} p_X(x)$ (follows from Additivity since different events $\{X = x\}$ are disjoint)
  - $\sum_{x \in \Omega} p_X(x) = 1$ (follows from Additivity and Normalization).
  - Example: X = number of heads in 2 fair coin tosses ($p = 1/2$). $P(X > 0) = \sum_{x=1}^{2} p_X(x) = 0.75$.
- Can also define a binary r.v. for any event A as: $X = 1$ if A occurs and $X = 0$ otherwise. Then X is a Bernoulli r.v. with $p = P(A)$.
- Bernoulli ($X = 1$ (heads) or $X = 0$ (tails)) r.v. with probability of heads p

$$Bernoulli(p): \ p_X(x) = p^x (1-p)^{1-x}, \ x = 0, \ or \ x = 1 \quad (2)$$

- Binomial ($X = x$ heads out of n independent tosses, probability of heads p)

$$Binomial(n, p): \ p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \ x = 0, 1, \ldots n \quad (3)$$

- Geometric r.v., X, with probability of heads p (X= number of coin tosses needed for a head to come up for the first time or number of independent trials needed to achieve the first "success").

  - Example: I keep taking a test until I pass it. Probability of passing the test in the x^{th} try is $p_X(x)$.
  - Easy to see that

$$Geometric(p): \; p_X(x) = (1-p)^{x-1}p, \quad x = 0, 1, 2, \ldots \infty \tag{4}$$

- Poisson r.v. X with expected number of arrivals Λ (e.g. if X = number of arrivals in time τ with arrival rate λ, then $\Lambda = \lambda\tau$)

$$Poisson(\Lambda): \; p_X(x) = \frac{e^{-\Lambda}(\Lambda)^x}{x!}, \; x = 0, 1, \ldots \infty \tag{5}$$

- Uniform(a,b):

$$p_X(x) = \begin{cases} 1/(b-a+1), & if \; x = a, a+1, \ldots b \\ 0, & otherwise \end{cases} \tag{6}$$

- pmf of $Y = g(X)$
  - $p_Y(y) = P(\{Y = y\}) = \sum_{x|g(x)=y} p_X(x)$
    Example $Y = |X|$. Then $p_Y(y) = p_X(y) + p_X(-y), \; if \; y > 0$ and $p_Y(0) = p_X(0)$. Exercise: $X \sim Uniform(-4, 4)$ and $Y = |X|$, find $p_Y(y)$.

- Expectation, mean, variance
  - Motivating example: Read pg 81
  - Expected value of X (or mean of X): $E[X] \triangleq \sum_{x \in \Omega} x p_X(x)$
  - Interpret mean as center of gravity of a bar with weights $p_X(x)$ placed at location x (Fig. 2.7)
  - Expected value of $Y = g(X)$: $E[Y] = E[g(X)] = \sum_{x \in \Omega} g(x) p_X(x)$. Exercise: show this.
  - n^{th} moment of X: $E[X^n]$. n^{th} central moment: $E[(X - E[X])^n]$.
  - Variance of X: $var[X] \triangleq E[(X - E[X])^2]$ (2nd central moment)
  - $Y = aX + b$ (linear fn): $E[Y] = aE[X] + b$, $var[Y] = a^2 var[X]$
  - Poisson: $E[X] = \Lambda$, $var[X] = \Lambda$ (show this)
  - Bernoulli: $E[X] = p$, $var[X] = p(1-p)$ (show this)
  - Uniform(a,b): $E[X] = (a+b)/2$, $var[X] = \frac{(b-a+1)^2 - 1}{12}$ (show this)

- Application: Computing average time. Example 2.4

- Application: Decision making using expected values. Example 2.8 (Quiz game, compute expected reward with two different strategies to decide which is a better strategy).

- $Binomial(n, p)$ becomes $Poisson(np)$ if time interval between two coin tosses becomes very small (so that n becomes very large and p becomes very small, but $\Lambda = np$ is finite). **

3 Multiple Discrete Random Variables: Topics

- Joint PMF, Marginal PMF of 2 and or more than 2 r.v.'s
- PMF of a function of 2 r.v.'s
- Expected value of functions of 2 r.v's
- Expectation is a linear operator. Expectation of sums of n r.v.'s
- Conditioning on an event and on another r.v.
- Bayes rule
- Independence

4 Joint & Marginal PMF, PMF of function of r.v.s, Expectation

- For everything in this handout, you can think in terms of events $\{X = x\}$ and $\{Y = y\}$ and apply what you have learnt in Chapter 1.
- The **joint PMF** of two random variables X and Y is defined as

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

  where $P(X = x, Y = y)$ is the same as $P(\{X = x\} \cap \{Y = y\})$.

  - Let A be the set of all values of x, y that satisfy a certain property, then $P((X,Y) \in A) = \sum_{(x,y)\in A} p_{X,Y}(x,y)$
  - e.g. X = outcome of first die toss, Y is outcome of second die toss, A = sum of outcomes of the two tosses is even.

- **Marginal PMF** is another term for the PMF of a single r.v. obtained by **"marginalizing"** the joint PMF over the other r.v., i.e. the marginal PMF of X, $p_X(x)$ can be computed as follows:
  Apply Total Probability Theorem to $p_{X,Y}(x,y)$, i.e. sum over $\{Y = y\}$ for different values y (these are a set of disjoint events whose union is the sample space):

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

  Similarly the marginal PMF of Y, $p_Y(y)$ can be computed by "marginalizing" over X

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

- **PMF of a function of r.v.'s:** If $Z = g(X,Y)$,

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

  - Read the above as $p_Z(z) = P(Z = z) = P(\text{all values of } (X,Y) \text{ for which } g(X,Y) = z)$

- **Expected value of functions of multiple r.v.'s**
  If $Z = g(X, Y)$,
$$E[Z] = \sum_{(x,y)} g(x, y) p_{X,Y}(x, y)$$

- See Example 2.9

- **More than 2 r.v.s.**
  - Joint PMF of n r.v.'s: $p_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots x_n)$
  - We can **marginalize** over one or more than one r.v.,
    e.g. $p_{X_1,X_2,\ldots X_{n-1}}(x_1, x_2, \ldots x_{n-1}) = \sum_{x_n} p_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots x_n)$
    e.g. $p_{X_1,X_2}(x_1, x_2) = \sum_{x_3,x_4,\ldots x_n} p_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots x_n)$
    e.g. $p_{X_1}(x_1) = \sum_{x_2,x_3,\ldots x_n} p_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots x_n)$
    See book, Page 96, for special case of 3 r.v.'s

- **Expectation is a linear operator.** *Exercise: show this*
$$E[a_1 X_1 + a_2 X_2 + \ldots a_n X_n] = a_1 E[X_1] + a_2 E[X_2] + \ldots a_n E[X_n]$$
  - Application: Binomial(n, p) is the sum of n Bernoulli r.v.'s. with success probability p, so its expected value is np (See Example 2.10)
  - See Example 2.11

5 Conditioning and Bayes rule

- **PMF of r.v. X conditioned on an event A with $P(A) > 0$**
$$p_{X|A}(x) \triangleq P(\{X = x\}|A) = \frac{P(\{X = x\} \cap A)}{P(A)}$$
  - $p_{X|A}(x)$ is a legitimate PMF, i.e. $\sum_x p_{X|A}(x) = 1$. *Exercise: Show this*
  - Example 2.12, 2.13

- **PMF of r.v. X conditioned on r.v. Y.** Replace A by $\{Y = y\}$
$$p_{X|Y}(x|y) \triangleq P(\{X = x\}|\{Y = y\}) = \frac{P(\{X = x\} \cap \{Y = y\})}{P(\{Y = y\})} = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$
  The above holds for all y for which $p_y(y) > 0$. The above is equivalent to
$$p_{X,Y}(x, y) = p_{X|Y}(x|y) p_Y(y)$$
$$p_{X,Y}(x, y) = p_{Y|X}(y|x) p_X(x)$$
  - $p_{X|Y}(x|y)$ (with $p_Y(y) > 0$) is a legitimate PMF, i.e. $\sum_x p_{X|Y}(x|y) = 1$.
  - Similarly, $p_{Y|X}(y|x)$ is also a legitimate PMF, i.e. $\sum_y p_{Y|X}(y|x) = 1$. *Show this.*
  - Example 2.14 (I did a modification in class), 2.15

- **Bayes rule.** How to compute $p_{X|Y}(x|y)$ using $p_X(x)$ and $p_{Y|X}(y|x)$,

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

- **Conditional Expectation given event A**

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

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

- **Conditional Expectation given r.v. $Y = y$.** Replace A by $\{Y = y\}$

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

Note this is a function of $Y = y$.

- **Total Expectation Theorem**

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

Proof on page 105.

- **Total Expectation Theorem for disjoint events $A_1, A_2, \ldots A_n$ which form a partition of sample space.**

$$E[X] = \sum_{i=1}^{n} P(A_i) E[X|A_i]$$

Note A_i's are disjoint and $\cup_{i=1}^{n} A_i = \Omega$

  - Application: Expectation of a geometric r.v., Example 2.16, 2.17

6 Independence

- **Independence of a r.v. & an event A.** r.v. X is independent of A with $P(A) > 0$, iff

$$p_{X|A}(x) = p_X(x), \quad \text{for all } x$$

.

  - This also implies: $P(\{X = x\} \cap A) = p_X(x)P(A)$.
  - See Example 2.19

- **Independence of 2 r.v.'s.** R.v.'s X and Y are independent iff

$$p_{X|Y}(x|y) = p_X(x), \text{ for all } x \text{ and for all } y \text{ for which } p_Y(y) > 0$$

  This is equivalent to the following two things(*show this*)

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

$$p_{Y|X}(y|x) = p_Y(y), \text{ for all } y \text{ and for all } x \text{ for which } p_X(x) > 0$$

- **Conditional Independence of r.v.s X and Y given event A with $P(A) > 0$ ****
  $p_{X|Y,A}(x|y) = p_{X|A}(x)$ for all x and for all y for which $p_{Y|A}(y) > 0$ or that
  $p_{X,Y|A}(x,y) = p_{X|A}(x)p_{Y|A}(y)$

- **Expectation of product of independent r.v.s.**
  - If X and Y are independent, $E[XY] = E[X]E[Y]$.

$$\begin{aligned} E[XY] &= \sum_y \sum_x xy p_{X,Y}(x,y) \\ &= \sum_y \sum_x xy p_X(x) p_Y(y) \\ &= \sum_y y p_Y(y) \sum_x x p_X(x) \\ &= E[X]E[Y] \end{aligned}$$

  - If X and Y are independent, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$. (Show).

- If $X_1, X_2, \ldots X_n$ are independent,

$$p_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots x_n) = p_{X_1}(x_1)p_{X_2}(x_2)\ldots p_{X_n}(x_n)$$

- **Variance of sum of 2 independent r.v.'s.**
  Let X, Y are independent, then $Var[X+Y] = Var[X] + Var[Y]$.
  See book page 112 for the proof

- **Variance of sum of n independent r.v.'s.**
  If $X_1, X_2, \ldots X_n$ are independent,

$$Var[X_1 + X_2 + \ldots X_n] = Var[X_1] + Var[X_2] + \ldots Var[X_n]$$

  .

  - **Application: Variance of a Binomial**, See Example 2.20
    Binomial r.v. is a sum of n independent Bernoulli r.v.'s. So its variance is $np(1-p)$
  - **Application: Mean and Variance of Sample Mean**, Example 2.21
    Let $X_1, X_2, \ldots X_n$ be independent and *identically distributed*, i.e. $p_{X_i}(x) = p_{X_1}(x)$ for all i. Thus all have the same mean (denote by a) and same variance (denote by v).
    Sample mean is defined as $S_n = \frac{X_1+X_2+\ldots X_n}{n}$.
    Since $E[.]$ is a linear operator, $E[S_n] = \sum_{i=1}^n \frac{1}{n}E[X_i] = \frac{na}{n} = a$.
    Since the X_i's are independent, $Var[S_n] = \sum_{i=1}^n \frac{1}{n^2}Var[X_i] = \frac{nv}{n^2} = \frac{v}{n}$
  - **Application:** Estimating Probabilities by Simulation, See Example 2.22

Note: Handouts DO NOT replace the book. In most cases, they only provide a guideline on topics.
For exams/quizzes, you are not expected to know items with ** (these are provided as extra information).

1 Continuous R.V. & Probability Density Function (PDF)

- Example: velocity of a car
- A r.v. X is called **continuous** if there is a function $f_X(x)$ with $f_X(x) \geq 0$, called **probability density function (PDF)**, s.t. $P(X \in B) = \int_B f_X(x)dx$ for all subsets B of the real line.
- Specifically, for $B = [a, b]$,

$$P(a \leq X \leq b) = \int_{x=a}^{b} f_X(x)dx \quad (1)$$

  and can be interpreted as the area under the graph of the PDF $f_X(x)$.
- For any single value a, $P(\{X = a\}) = \int_{x=a}^{a} f_X(x)dx = 0$.
- Thus $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$
- Sample space $\Omega = (-\infty, \infty)$
- Normalization: $P(\Omega) = P(-\infty < X < \infty) = 1$. Thus $\int_{x=-\infty}^{\infty} f_X(x)dx = 1$
- Interpreting the PDF: For an interval $[x, x + \delta]$ with very small δ,

$$P([x, x + \delta]) = \int_{t=x}^{x+\delta} f_X(t)dt \approx f_X(x)\delta \quad (2)$$

  Thus $f_X(x)$= probability mass per unit length near x. See Fig. 3.2.
- Continuous uniform PDF, Example 3.1
- Piecewise constant PDF, Example 3.2
- Connection with a PMF (explained after CDF is explained) **
- Expected value: $E[X] = \int_{x=-\infty}^{\infty} x f_X(x)dx$. Similarly define $E[g(X)]$ and $var[X]$
- Mean and variance of uniform, Example 3.4
- Exponential r.v.

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & if\ x \geq 0 \\ 0, & otherwise \end{cases} \quad (3)$$

  - Show it is a legitimate PDF.
  - $E[X] = 1/\lambda$, $var[X] = 1/\lambda^2$ (show).
  - Example: X= amount of time until an equipment breaks down or a bulb burns out.
  - Example 3.5 (Note: you need to use the correct time unit in the problem, here days).

2 Cumulative Distribution Function (CDF)

- Cumulative Distribution Function (CDF), $F_X(x) \triangleq P(X \leq x)$ (probability of event $\{X \leq x\}$).
- Defined for discrete and continuous r.v.'s

$$\text{Discrete:} \quad F_X(x) = \sum_{k \leq x} p_X(k) \tag{4}$$

$$\text{Continuous:} \quad F_X(x) = \int_{-\infty}^{x} f_X(t)dt \tag{5}$$

- Note the PDF $f_X(x)$ is NOT a probability of any event, it can be > 1.
- But $F_X(x)$ is the probability of the event $\{X \leq x\}$ for both continuous and discrete r.v.'s.
- Properties
  - $F_X(x)$ is monotonically nondecreasing in x.
  - $F_X(x) \to 0$ as $x \to -\infty$ and $F_X(x) \to 1$ as $x \to \infty$
  - $F_X(x)$ is continuous for continuous r.v.'s and it is piecewise constant for discrete r.v.'s
- Relation to PMF, PDF

$$\text{Discrete:} \quad p_X(k) = F_X(k) - F_X(k-1) \tag{6}$$

$$\text{Continuous:} \quad f_X(x) = \frac{dF_X}{dx}(x) \tag{7}$$

- Using CDF to compute PMF.
  - Example 3.6: Compute PMF of maximum of 3 r.v.'s: What is the PMF of the maximum score of 3 test scores, when each test score is independent of others and each score takes any value between 1 and 10 with probability 1/10?
    Answer: Compute $F_X(k) = P(X \leq k) = P(\{X_1 \leq k\}, \, and \, \{X_2 \leq k\}, \, and \, \{X_3 \leq k\}) = P(\{X_1 \leq k\})P(\{X_2 \leq k\})P\{X_3 \leq k\})$ (follows from independence of the 3 events) and then compute the PMF using (6).
  - For continuous r.v.'s, in almost all cases, the correct way to compute the CDF of a function of a continuous r.v. (or of a set of continuous r.v.'s) is to compute the CDF first and then take its derivative to get the PDF. We will learn this later.
- Connection of a PDF with a PMF **
  - You learnt the Dirac delta function in EE 224. We use it to define a PDF for discrete r.v.
  - The PDF of a discrete r.v. X, $f_X(x) \triangleq \sum_{j=-\infty}^{\infty} p_X(j)\delta(x-j)$.
  - If I integrate this, I get $F_X(x) = \int_{t \leq x} f_X(t)dt = \sum_{j \leq x} p_X(j)$ which is the same as the CDF definition given in (4)

- Geometric and exponential CDF **
  - Let $X_{geo,p}$ be the number of trials required for the first success (geometric) with probability of success $= p$. Then we can show that the probability of $\{X_{geo,p} \le k\}$ is equal to the probability of an exponential r.v. $\{X_{expo,\lambda} \le k\delta\}$ with parameter λ, if δ satisfies $1 - p = e^{-\lambda\delta}$ or $\delta = -\ln(1-p)/\lambda$
    Proof: Equate $F_{X_{geo,p}}(k) = 1 - (1-p)^k$ to $F_{X_{expo,\lambda}}(k\delta) = 1 - e^{-\lambda k\delta}$
  - Implication: When δ (time interval between two Bernoulli trials (coin tosses)) is small, then $F_{X_{geo,p}}(k) \approx F_{X_{expo,\lambda}}(k\delta)$ with $p = \lambda\delta$ (follows because $e^{-\lambda\delta} \approx 1 - \lambda\delta$ for δ small).
- $Binomial(n,p)$ becomes $Poisson(np)$ for small time interval, δ, between coin tosses (Details in Chap 5) **
  Proof idea:
  - Consider a sequence of n independent coin tosses with probability of heads p in any toss (number of heads $\sim$ $Binomial(n,p)$).
  - Assume the time interval between two tosses is δ.
  - Then expected value of X in one toss (in time δ) is p.
  - When δ small, expected value of X per unit time is $\lambda = p/\delta$.
  - The total time duration is $\tau = n\delta$.
  - When $\delta \to 0$, but λ and τ are finite, $n \to \infty$ and $p \to 0$.
  - When δ small, can show that the PMF of a $Binomial(n,p)$ r.v. is approximately equal to the PMF of $Poisson(\lambda\tau)$ r.v. with $\lambda\tau = np$
- The Poisson process is a continuous time analog of a Bernoulli process (Details in Chap 5) **

3 Normal (Gaussian) Random Variable

- The most commonly used r.v. in Communications and Signal Processing
- X is normal or Gaussian if it has a PDF of the form
  $$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$
  where one can show that $\mu = E[X]$ and $\sigma^2 = var[X]$.
- Standard normal: Normal r.v. with $\mu = 0$, $\sigma^2 = 1$.
- Cdf of a standard normal Y, denoted $\Phi(y)$
  $$\Phi(y) \triangleq P(Y \le y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2} dt$$
  It is recorded as a table (See pg 155).
- Let X is a normal r.v. with mean μ, variance σ^2. Then can show that $Y = \frac{X-\mu}{\sigma}$ is a standard normal r.v.

- Computing CDF of any normal r.v. X using the table for Φ: $F_X(x) = \Phi(\frac{x-\mu}{\sigma})$. See Example 3.7.
- Signal detection example (computing probability of error): Example 3.8. See Fig. 3.11. A binary message is tx as a signal S which is either -1 or +1. The channel corrupts the tx with additive Gaussian noise, N, with mean zero and variance σ^2. The received signal, $Y = S + N$. The receiver concludes that a -1 (or +1) was tx'ed if $Y < 0$ ($Y \geq 0$). What is the probability of error? Answer: It is given by $P(N \geq 1) = 1 - \Phi(1/\sigma)$. How we get the answer will be discussed in class.
- Normal r.v. models the additive effect of many independent factors well **
  - This is formally stated as the central limit theorem (see Chap 7) : sum of a large number of independent and identically distributed (not necessarily normal) r.v.'s has an approximately normal CDF.

4 Multiple Continuous Random Variables: Topics

- Conditioning on an event
- Joint and Marginal PDF
- Expectation, Independence, Joint CDF, Bayes rule
- Derived distributions
  - Function of a Single random variable: $Y = g(X)$ for any function g
  - Function of a Single random variable: $Y = g(X)$ for linear function g
  - Function of a Single random variable: $Y = g(X)$ for strictly monotonic g
  - Function of Two random variables: $Z = g(X, Y)$ for any function g

5 Conditioning on an event.

$$f_{X|A}(x) := \begin{cases} \frac{f_X(x)}{P(A)} & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Consider the special case when $A := \{X \in R\}$, e.g. the region R can be the interval $[a, b]$. In this case, we should be writing $f_{X|\{X \in R\}}$. But to keep things simple, we misuse notation to also write

$$\begin{aligned} f_{X|R}(x) &:= \begin{cases} \frac{f_X(x)}{P(X \in R)} & \text{if } x \in R \\ 0 & \text{otherwise} \end{cases} \\ &:= \begin{cases} \frac{f_X(x)}{\int_{t \in R} f_X(t) dt} & \text{if } x \in R \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

6 Joint and Marginal PDF

- Two r.v.s X and Y are **jointly continuous** iff there is a function $f_{X,Y}(x,y)$ with $f_{X,Y}(x,y) \geq 0$, called the **joint PDF**, s.t. $P((X,Y) \in B) = \int_B f_{X,Y}(x,y)dxdy$ for all subsets B of the 2D plane.
- Specifically, for $B = [a,b] \times [c,d] \triangleq \{(x,y) : a \leq x \leq b, c \leq y \leq d\}$,

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_{y=c}^{d} \int_{x=a}^{b} f_{X,Y}(x,y)dxdy$$

- **Interpreting the joint PDF:** For small positive numbers δ_1, δ_2,

$$P(a \leq X \leq a+\delta_1, c \leq Y \leq c+\delta_2) = \int_{y=c}^{c+\delta_2} \int_{x=a}^{a+\delta_1} f_{X,Y}(x,y)dxdy \approx f_{X,Y}(a,c)\delta_1\delta_2$$

  Thus $f_{X,Y}(a,c)$ is the probability mass per unit area near (a,c).

- **Marginal PDF:** The PDF obtained by integrating the joint PDF over the entire range of one r.v. (in general, integrating over a set of r.v.'s)

$$\begin{aligned} P(a \leq X \leq b) &= P(a \leq X \leq b, -\infty \leq Y \leq \infty) = \int_{x=a}^{b} \int_{y=-\infty}^{\infty} f_{X,Y}(x,y)dydx \\ \Longrightarrow f_X(x) &= \int_{y=-\infty}^{\infty} f_{X,Y}(x,y)dy \end{aligned}$$

- Example 3.12, 3.13

7 Conditional PDF

- Conditional PDF of X given that $Y = y$ is defined as

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

- For any y, $f_{X|Y}(x|y)$ is a legitimate PDF: integrates to 1.
- Example 3.15
- **Interpretation:** For small positive numbers δ_1, δ_2, consider the probability that X belongs to a small interval $[x, x+\delta_1]$ given that Y belongs to a small interval $[y, y+\delta_2]$

$$\begin{aligned} P(x \leq X \leq x+\delta_1 | y \leq Y \leq y+\delta_2) &= \frac{P(x \leq X \leq x+\delta_1, y \leq Y \leq y+\delta_2)}{P(y \leq Y \leq y+\delta_2)} \\ &\approx \frac{f_{X,Y}(x,y)\delta_1\delta_2}{f_Y(y)\delta_2} \\ &= f_{X|Y}(x|y)\delta_1 \end{aligned}$$

- **Since $f_{X|Y}(x|y)\delta_1$ does not depend on δ_2, we can think of the limiting case when $\delta_2 \to 0$ and so we get**

$$P(x \leq X \leq x+\delta_1 | Y = y) = \lim_{\delta_2 \to 0} P(x \leq X \leq x+\delta_1 | y \leq Y \leq y+\delta_2) \approx f_{X|Y}(x|y)\delta_1 \quad \delta_1 \text{ small}$$

- In general, for any region A, we have that

$$P(X \in A|Y=y) = \lim_{\delta \to 0} P(X \in A|y \le Y \le y+\delta) = \int_{x \in A} f_{X|Y}(x|y)dx$$

8 Expectation, Independence, Joint & Conditional CDF, Bayes

- **Expectation:** See page 172 for $E[g(X)|Y=y]$, $E[g(X,Y)|Y=y]$ and total expectation theorem for $E[g(X)]$ and for $E[g(X,Y)]$.

- **Independence:** X and Y are independent iff $f_{X|Y} = f_X$ (or iff $f_{X,Y} = f_X f_Y$, or iff $f_{Y|X} = f_Y$)

- If X and Y independent, any two events $\{X \in A\}$ and $\{Y \in B\}$ are independent.

- If X and Y independent, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ and $Var[X+Y] = Var[X]+Var[Y]$ Exercise: show this.

- **Joint CDF:**

$$F_{X,Y}(x,y) := P(X \le x, Y \le y) = \int_{t=-\infty}^{y} \int_{s=-\infty}^{x} f_{X,Y}(s,t)dsdt$$

- Obtain joint PDF from joint CDF:

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

- **Conditional CDF:**

$$F_{X|Y}(x|y) := P(X \le x|Y=y) = \lim_{\delta \to 0} P(X \le x|y \le Y \le y+\delta) = \int_{t=-\infty}^{x} f_{X|Y}(t|y)dt$$

- **Bayes rule when unobserved phenomenon is continuous.** Pg 175 and Example 3.18. Recall that $f_{X|Y}(x|y)$ is, by definition, such that, for δ small,

$$P(X \in [x, x+\delta]|Y=y) = f_{X|Y}(x|y)\delta$$

  Also, for δ, δ_2 small,

$$P(X \in [x, x+\delta], Y \in [y, y+\delta_2]) = f_{X,Y}(x,y)\delta\delta_2$$

  Using Bayes rule for events,

$$P(X \in [x, x+\delta]|Y \in [y, y+\delta_2]) = \frac{P(X \in [x, x+\delta], Y \in [y, y+\delta_2])}{P(Y \in [y, y+\delta_2])} = \frac{f_{X,Y}(x,y)\delta\delta_2}{f_Y(y)\delta_2} = \frac{f_{X,Y}(x,y)\delta}{f_Y(y)}$$

  Notice that the right hand side does not depend on δ_2. Taking the limit $\delta_2 \to 0$, we get

$$P(X \in [x, x+\delta]|Y=y) = \lim_{\delta_2 \to 0} P(X \in [x, x+\delta]|Y \in [y, y+\delta_2]) = \frac{f_{X,Y}(x,y)\delta}{f_Y(y)}$$

  Thus,

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

- **Bayes rule when unobserved phenomenon is discrete.** Pg 176 and Example 3.19. For e.g., say discrete r.v. N is the unobserved phenomenon. Then for δ small,

$$\begin{aligned} P(N=i|X\in[x,x+\delta]) &= P(N=i|X\in[x,x+\delta]) \\ &= \frac{P(N=i)P(X\in[x,x+\delta]|N=i)}{P(X\in[x,x+\delta])} \\ &= \frac{p_N(i)f_{X|N}(x|i)\delta}{\sum_j p_N(j)f_{X|N}(x|j)\delta} \\ &= \frac{p_N(i)f_{X|N}(x|i)}{\sum_j p_N(j)f_{X|N}(x|j)} \end{aligned}$$

  Notice that the right hand side is independent of δ. Thus we can take $\lim_{\delta\to 0}$ on both sides and the right side will not change. Thus we get

$$p_{N|X}(i|x) = P(N=i|X=x) = \lim_{\delta\to 0} P(N=i|X\in[x,x+\delta]) = \frac{p_N(i)f_{X|N=i}(x)}{\sum_j p_N(j)f_{X|N=j}(x)}$$

- **Bayes rule with conditioning on events.** The derivation is analogous to the above conditioning on discrete r.v.'s case.
  Suppose that events $A_1, A_2, \ldots A_n$ form *a partition*, i.e. they are disjoint and their union is the entire sample space. The simplest example is $n=2$, $A_1 = A$, $A_2 = A^c$.
  Then

$$P(A_i|X=x) = \frac{P(A_i)f_{X|A_i}(x)}{\sum_j P(A_j)f_{X|A_j}(x)}$$

- More than 2 random variables (Pg 178, 179) **

9 Derived distributions: PDF of $g(X)$ and of $g(X,Y)$

- **Obtaining PDF of $Y = g(X)$.** ALWAYS use the following 2 step procedure:
  - Compute CDF first. $F_Y(y) = P(g(X) \le y) = \int_{x|g(x)\le y} f_X(x)dx$
  - Obtain PDF by differentiating F_Y, i.e. $f_Y(y) = \frac{\partial F_Y}{\partial y}(y)$
- Example 3.20, 3.21, 3.22
- **Special Case 1: Linear Case: $Y = aX + b$.** Can show that

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

  Proof: see Pg 183.
- Example 3.23, 3.24
- **Special Case 2: Strictly Monotonic Case.**
  - Consider $Y = g(X)$ with g being a **strictly monotonic** function of X.
  - Thus g is a one to one function.

- Thus there exists a function h s.t. $y = g(x)$ iff $x = h(y)$ (i.e. h is the inverse function of g, often denotes as $h \triangleq g^{-1}$).
- Then can show that

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

- Proof for strictly monotonically increasing g:
  $F_Y(y) = P(g(X) \leq Y) = P(X \leq h(Y)) = F_X(h(y))$.
  Differentiate both sides w.r.t y (apply chain rule on the right side) to get:

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

  For strictly monotonically decreasing g, using a similar procedure, we get $f_Y(y) = -f_X(h(y))\frac{dh}{dy}(y)$
- See Figure 3.22, 3.23 for intuition

- Example 3.21 (page 186)

- **Functions of two random variables.** Two possible ways to solve this depending on which is easier. Try the first method first: if easy to find the region to integrate over then just do that. Else use the second method.

  1. Do the following
     (a) Compute CDF of $Z = g(X, Y)$, i.e compute $F_Z(z)$. In general, this computed as: $F_Z(z) = P(g(X, Y) \leq z) = \int_{x,y:g(x,y)\leq z} f_{X,Y}(x, y)dydx$.
     (b) Differentiate w.r.t. z to get the PDF, i.e. compute $f_Z(z) = \frac{\partial F_Z(z)}{\partial z}$.
  2. Use a three step procedure
     (a) Compute conditional CDF, $F_{Z|X}(z|x) := P(Z \leq z|X = x)$
     (b) Differentiate w.r.t. z to get conditional PDF, $f_{Z|X}(z|x) = \frac{\partial F_{Z|X}(z|x)}{\partial z}$
     (c) Compute $f_Z(z) = \int f_{Z,X}(z, x)dx = \int f_{Z|X}(z|x)f_X(x)dx$

- Example 3.26, 3.27, 3.28: first method works.

- Special case: PDF of $Z = X + Y$ when X, Y are independent: convolution of PDFs of X and Y. Here need to use the second method.