COMPSCI 240: Reasoning Under Uncertainty

Andrew Lan and Nic Herndon

University of Massachusetts at Amherst

Spring 2019, Section 01

Lecture 32: Review for Final Exam

Topics

- Basic counting problems
- Probability
- Discrete random variables
- Midterm Exam #1
- Continuous random variables
- Central limit theorem
- Probabilistic reasoning
- Game theory
- Midterm Exam #2
- Markov chains
- Bayesian network
- Final Exam

Part I Overview

- Basic counting problems
  - Set theory: size of, subset, disjoint sets, partitions, power set, universal set, operations (complement, union, intersection)
  - Counting: permutations, k-permutations, combinations, partitions
- Probability
  - Probability axioms
  - Conditional probability (sequential model)
  - Multiplication rule
  - Total probability theorem
  - Bayes' rule
  - Independence
  - Conditional independence
- Discrete random variables
  - Probability mass function (PMF)
  - Common discrete RVs: uniform, Bernoulli, binomial, geometric, Poisson
  - Expectation and Variance + their properties (e.g., functions of RVs)
  - Multiple RVs (joint, marginal, conditional PMF; functions of two RVs, expectation and variance)

Part II Overview

- Continuous random variables
  - Probability *density* function (PDF), cumulative density function (CDF), and probability mass
  - Expectation and Variance + their properties
  - Common continuous RVs: uniform, exponential, (standard) normal/Gaussian
  - Multiple RVs (joint, marginal, conditional PDFs), covariance, correlation
- Limit theorems
  - Markov bound
  - Chebyshev bound
  - Weak law of large numbers, and convergence in probability
  - Strong law of large numbers
  - Central limit theorem
- Game theory
  - Strategies: pure, IESDS, mixed
  - Nash equilibrium
  - Zero-sum games

Problems from MIT OCW:
Probabilistic Systems Analysis and Applied Probability

Problem 1 (Tutorial 4 Problem 1)

Let X and Y be Gaussian random variables, with $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(1,4)$.

(a) Find $P(X \leq 1.5)$ and $P(X \leq -1)$.

(b) What is the distribution of $\frac{Y-1}{2}$?

(c) Find $P(-1 \leq Y \leq 1)$.

Solution:

(a) $P(X \leq 1.5) = \Phi(1.5) = 0.9332$ and
$P(X \leq -1) = 1 - P(X \leq 1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$.

(b) $E\left[\frac{Y-1}{2}\right] = \frac{1}{2}(E[Y] - 1) = 0$ and $var\left(\frac{Y-1}{2}\right) = var\left(\frac{Y}{2}\right) = \frac{1}{4}var(Y) = 1$
Thus, the distribution of $\frac{Y-1}{2} \sim \mathcal{N}(0,1)$.

(c) $P(-1 \leq Y \leq 1) = P\left(\frac{-1-1}{2} \leq \frac{Y-1}{2} \leq \frac{1-1}{2}\right) = \Phi(0) - \Phi(-1) =$
$\Phi(0) - [1 - \Phi(1)] = 0.3413$.

Problem 2 (Tutorial 4 Problem 2)

Ben throws a dart at a circular target of radius r. We assume that he always hits the target, and that all points of impact (x, y) are equally likely. Define the joint PDF $f_{X,Y}(x, y)$ of the random variables X and Y and the conditional PDF $f_{X|Y}(x|y)$.

Solution:

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

$$= \begin{cases} \frac{1}{\pi r^2} & \text{, if } x^2 + y^2 \leq r^2 \\ 0 & \text{, otherwise} \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \frac{1}{\pi r^2} \int_{-\sqrt{r^2-y^2}}^{\sqrt{r^2-y^2}} dx = \frac{2\sqrt{r^2-y^2}}{\pi r^2}, \text{ if } |y| \leq r$$

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{1}{\pi r^2}}{\frac{2\sqrt{r^2-y^2}}{\pi r^2}} = \frac{1}{2\sqrt{r^2-y^2}}, \text{ if } x^2 + y^2 \leq r^2$$

Problem 3 (Recitation 8 Problem 3)

Let λ be a positive number. The continuous random variable X is called exponential with parameter λ when its probability density function is

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

(a) Find the cumulative distribution function (CDF) of X.

(b) Find the mean of X.

(c) Find the variance of X.

Solution:

(a) For $x \geq 0$, $F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_{0}^{x} e^{-\lambda t}dt = 1 - e^{-\lambda x}$.
For $x < 0$, $F_X(x) = 0$.

(b) $E[X] = \frac{1}{\lambda}$.

(c) $var(X) = \frac{1}{\lambda^2}$.

Problem 4 (Recitation 21 Problem 1)

Let $X_1, \ldots, X_{10}$ be independent random variables, uniformly distributed over the unit interval $[0, 1]$.

(a) Estimate $P(X_1, \ldots, X_{10} \geq 7)$ using the Markov inequality.

(b) Repeat part (a) using the Chebyshev inequality.

(c) Repeat part (a) using the central limit theorem.

Solution:

(a) Let $X = \sum_{i=1}^{10} X_i$. Then $E[X] = 10E[X_i] = 5$.

$$P(X \geq 7) \leq \frac{5}{7}$$

(b) $2P(X - 5 \geq 2) = P(|X - 5| \geq 2) \leq \frac{var(X)}{4} = \frac{10/12}{4}$

$$P(X - 5 \geq 2) \leq \frac{5}{48} = 0.1042$$

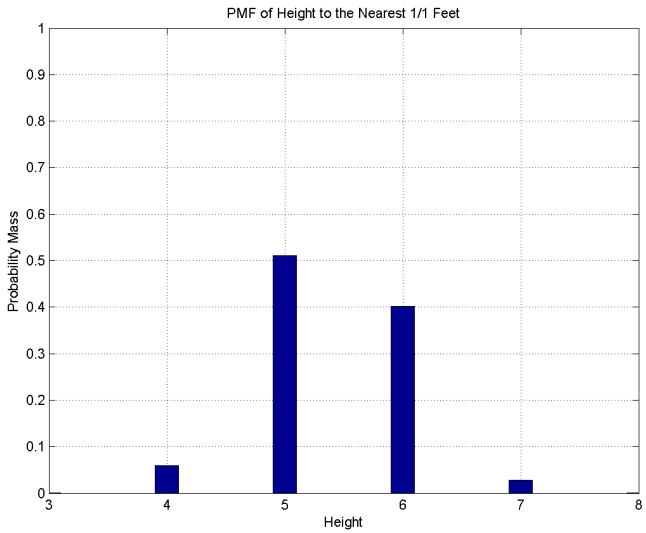
(c) $P\left(\sum_{i=1}^{10} X_i \geq 7\right) = 1 - P\left(\sum_{i=1}^{10} X_i \leq 7\right) = 1 - P\left(\frac{\sum_{i=1}^{10} X_i - 5}{\sqrt{10/12}} \leq \frac{7-5}{\sqrt{10/12}}\right) \approx 1 - \Phi(2.19)$

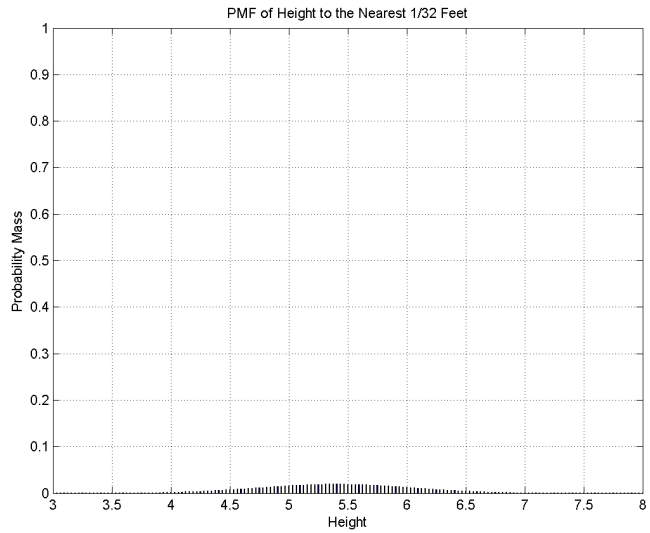
Continuous Random Variables

Continuous Random Variables

- There are many random variables that are much more naturally thought of as taking continuous values than a finite or countable number of values (ex: height, weight, distance, time, speed, cost, etc...).
- **Example:** Suppose we measure the height of a randomly selected person to the nearest foot. This gives a discrete random variable with a probability mass function.
- **Question:** What happens to the probability mass function of height if we measure it in smaller and smaller units?

Example: Height - Nearest Foot to Nearest 1/32 Foot



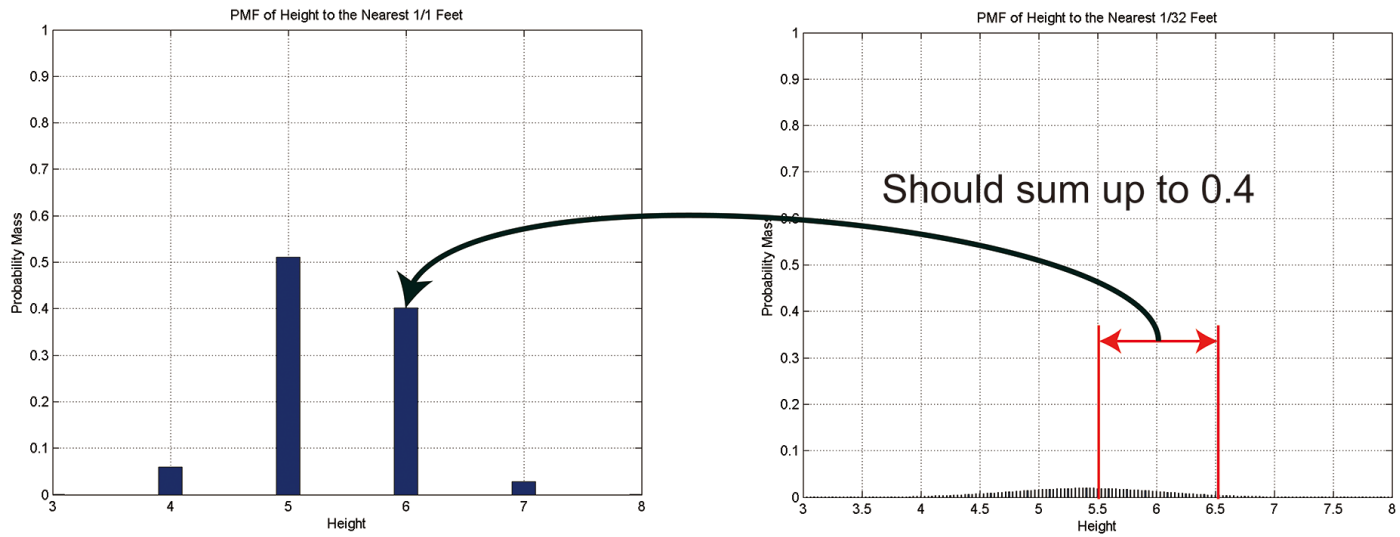


Example: Height

- **Question:** What happens to the probability mass function if we measure the height of people in smaller and smaller units?
- **Answer:** As we measure the height of people in smaller and smaller units, the number of values in the range of the random variable becomes uncountably infinite and the probability mass assigned to any particular value goes to zero!
- However, the probability mass $P(a < X < b)$ associated with an **interval** $[a, b]$ of $\mathcal{X}$ of length greater than zero can be non-zero.

Probability of Intervals

- **Example:** In the height example, the probability that the height of a person was between 5.5 and 6.5 feet was about 0.4. In other words $P(5.5 < X < 6.5) = 0.4$.



- In general, if $A \subseteq \mathcal{X}$ is any subset of the range of X that has non-zero length, then $P(X \in A)$ can be non-zero.

Continuous Random Variable

- We see that continuous random variable can provide more fine-grained probability profiles of a random variable.
- Is that it? In the height example, in the real world, we don't need to have high resolutions of $1/32$ feet.
- Most of real-world applications have discrete measurements (e.g., height, speed, etc.)
- Continuous random variables allow the use of powerful tools from calculus and often admit an insightful analysis that would not be possible under a discrete model.

Probability Density

- The standard way to construct probability laws for continuous random variables is using a **probability density function**.
- The only restrictions on the density function are that
  - **Non-negativity**: $f_X(x) \geq 0$ for all $X \subseteq \mathcal{X}$
  - **Normalization**: $\int_{\mathcal{X}} f_X(x)dx = \int_{-\infty}^{\infty} f_X(x)dx = 1$
- If we have a random variable X with probability density function $f_X(x)$ then the probability of any set $A \subseteq \mathcal{X}$ is given by the integral of the density over A:

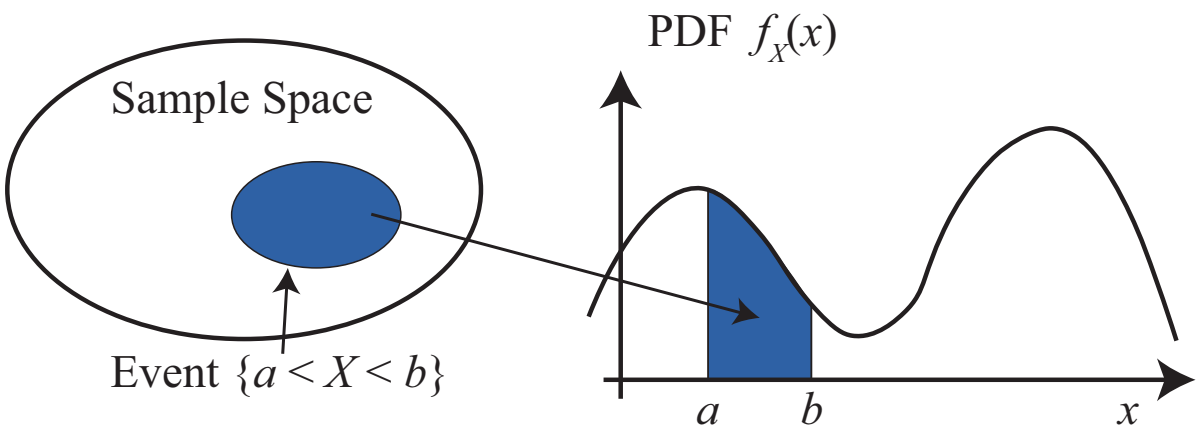
$$P(X \in A) = \int_A f_X(x)dx$$

Probability Density

- In the simplest case $A = [a, b]$ is a single interval and this definition reduces to a definite integral:

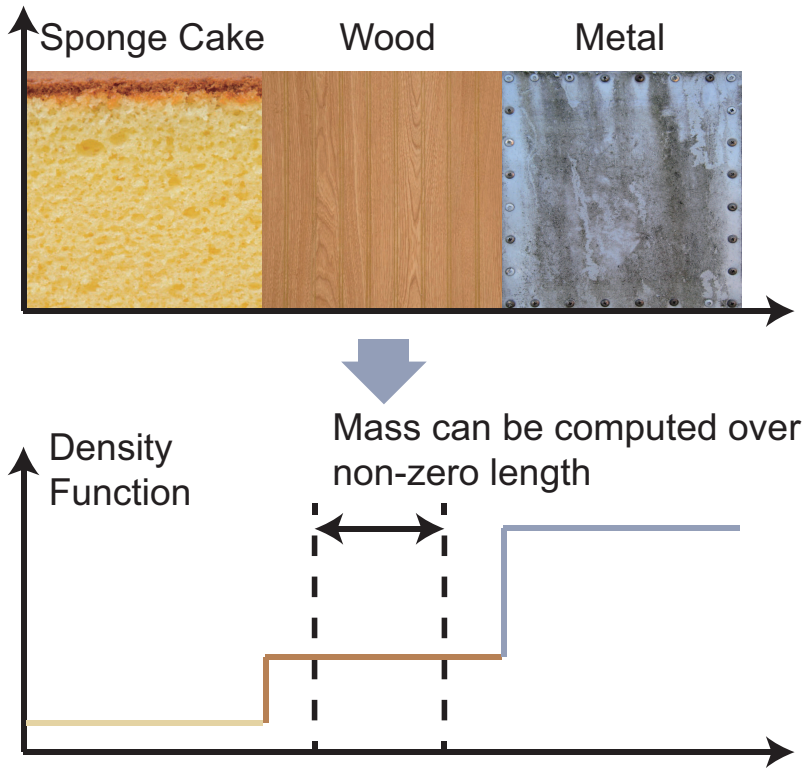
$$P(a < X < b) = \int_a^b f_X(x)dx$$

- Intuitively, the **probability mass** of an interval $[a, b]$ is $P(a < X < b)$.



Probability Density - Conceptual Visualization

- Image you have a thin panels of 1) sponge cake, 2) wood, and 3) metal.



- This is why we name it PDF and PMF.

Probability Density - Additivity

- When these two requirements (non-negativity and normalization) are met, defining probabilities of sets by integrating the density function over the set will automatically satisfy the additivity axiom.
- For example, if $[a, c]$ is any interval and $a < b < c$ then:

$$\begin{aligned} P(a < X < c) &= \int_a^c f_X(x)dx \\ &= \int_a^b f_X(x)dx + \int_b^c f_X(x)dx \\ &= P(a < X < b) + P(b < X < c) \end{aligned}$$

- However, if $[a, d]$ is any interval and $a < b < c < d$ then:

$$\begin{aligned} P(a < X < d) &= \int_a^d f_X(x)dx \\ &\neq \int_a^c f_X(x)dx + \int_b^d f_X(x)dx \\ &\neq P(a < X < c) + P(b < X < d) \end{aligned}$$

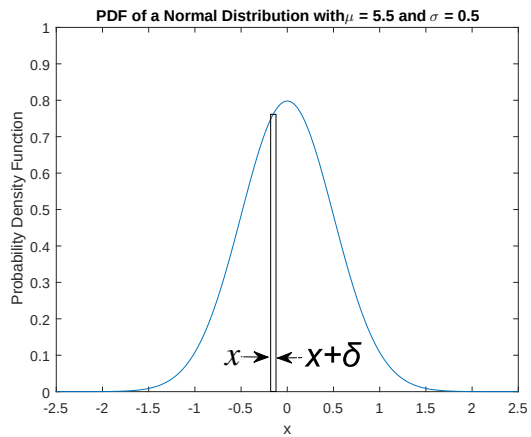
Probability Density

- For a single value a, we have

$$P(X = a) = \int_a^a f_X(x)dx = 0$$

- More precisely,

$$\begin{aligned} P(X = a) &= P(a < X < a + \delta) \text{ where } \delta \approx 0 \\ &= \int_x^{x+\delta} f_X(t)dt \approx f_X(x) \cdot \delta \\ &= 0 \end{aligned}$$



- For this reason,

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

Expectation and Variance

- The expected value or mean of a continuous random variable X can be computed in a similar manner as a discrete random variable:

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

- The expected value of a function of a random variable $g(X)$ can be computed as

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

- The variance of X can be computed as

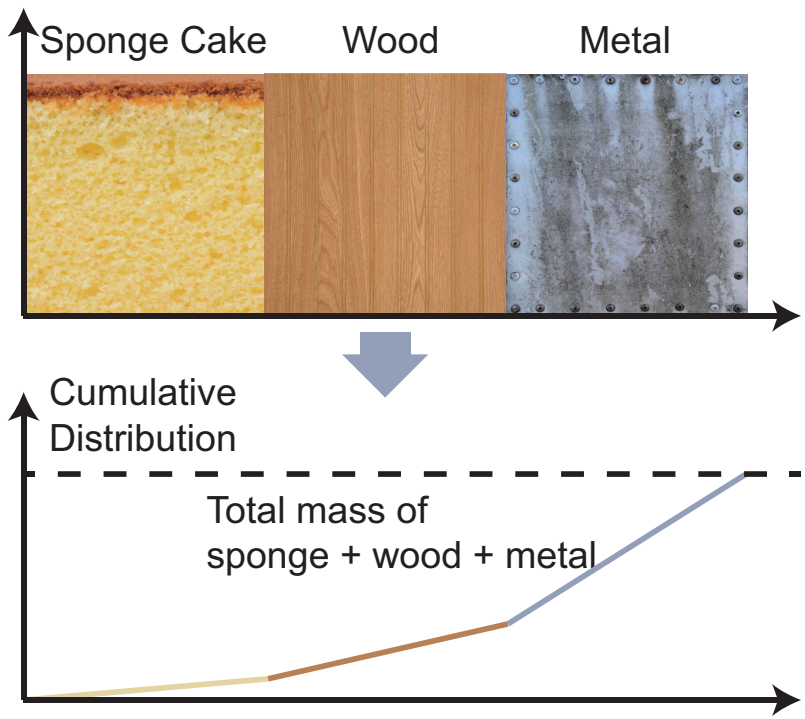
$$\begin{aligned} var(X) &= E\left[(x - E(x))^2\right] \\ &= \int_{-\infty}^{\infty} [x - E(x)]^2 f_X(x) dx \\ &= E(X^2) - E(X)^2 \end{aligned}$$

Cumulative Distribution Functions

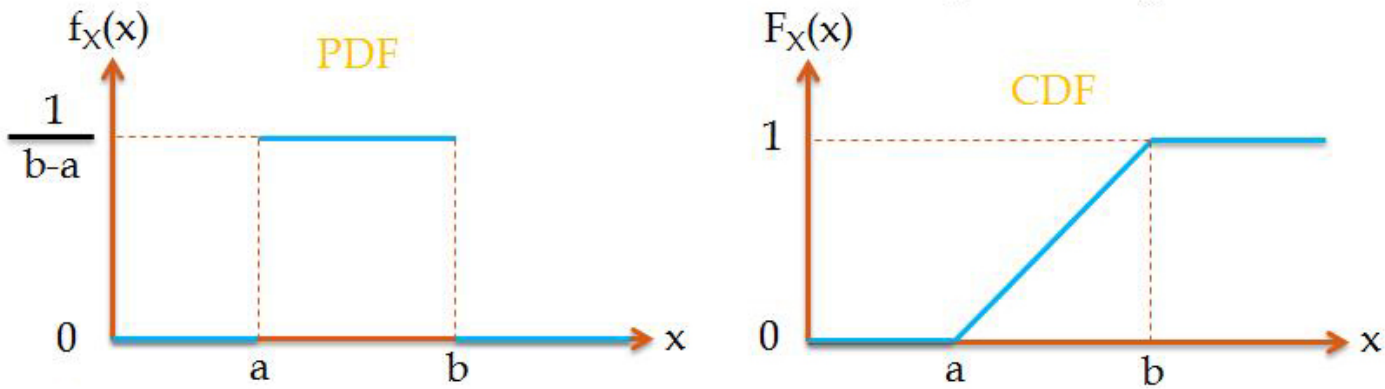
- The cumulative distribution function (CDF) for a continuous random variable X is defined as

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

- Intuitively, the CDF accumulates probability upto the value of x.



CDF - Example



- The PDF of the above graph can be defined as

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise,} \end{cases}$$

- **Question**: What is its CDF?
- **Answer**:

$$F_X(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & x > b \end{cases}$$

Cumulative Distribution Functions

- CDF
  - is a continuous function of x, if X is a continuous RV.
  - is monotonically non-decreasing:

$$\text{if } x \leq y, \text{ then } F_X(x) \leq F_X(y).$$

  - approaches 0 as $x \to -\infty$, and 1 as $x \to \infty$.
- If CDF is known, its PDF can be similarly derived as

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

Cumulative Distribution Functions

- **Question:** Well, if we have PDF, why do we need CDF?
- **Answer**: If we have CDF, we do not need to integrate every time when we compute $P(a \leq X \leq b)$.

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

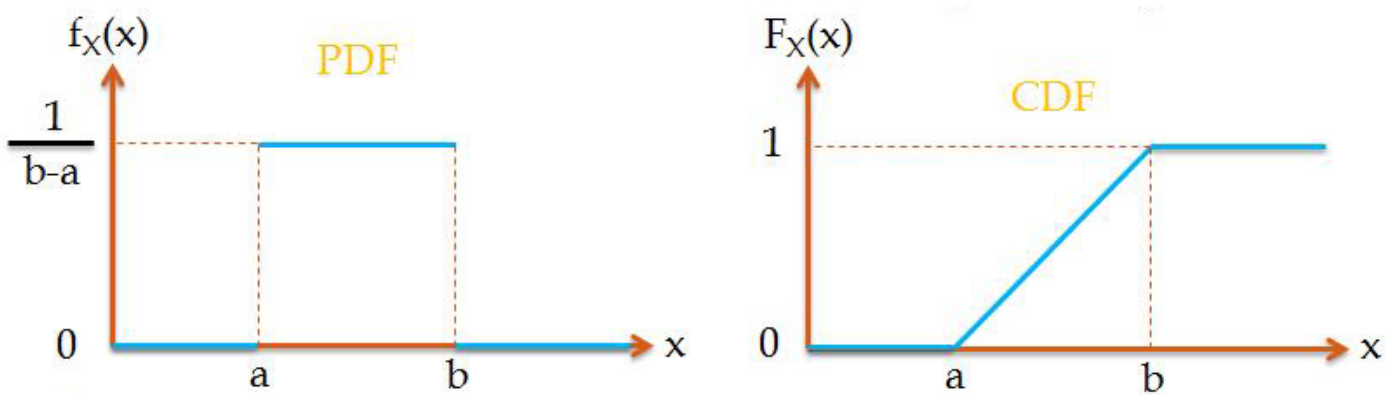
Common Continuous Random Variables

Uniform Continuous Random Variables

- Consider a RV that takes continuous values in an interval $[a, b]$.
- Uniform continuous RV has a uniform probability density in $[a, b]$.
- In other words, it has the same probability for two sub-intervals of the same length.
- Do not confuse with the discrete random variable!
- Its PDF can be defined as

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise,} \end{cases}$$

- We have looked at this distribution already.



Uniform Random Variables

- Its CDF can be defined as

$$F_X(x) = \begin{cases} 0, & \text{x <a} \\ \dfrac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & \text{x >b} \end{cases}$$

- When $b = 2$ and $a = 0$, what is $P(0.5 < X < 1.5)$?
- **Answer**: $F_X(1.5) - F_X(0.5) = \dfrac{1.5}{2} - \dfrac{0.5}{2} = \dfrac{1}{2}$.

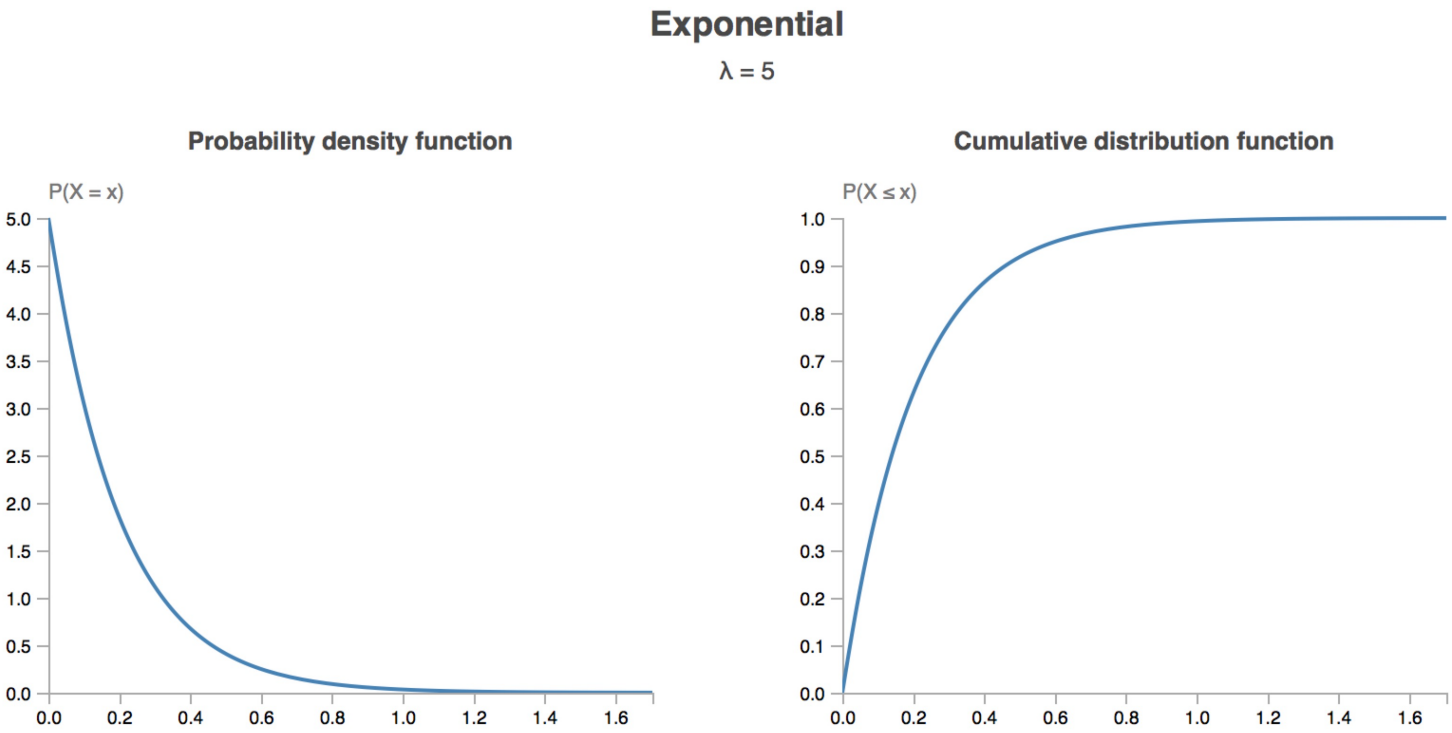
Exponential Random Variables

- An **exponential random variable** X is a continuous random variable with PDF:

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

  where λ must be strictly greater than 0.
- Exponential random variables are often used to model waiting times (eg: the length of time between calls at a call center, the length of time between people entering a store, the length of time between hits on a website, etc...).
- Closely connected to the geometric (discrete) random variable, which also relates to the discrete time that will elapse until an incident of interest occurs.

Exponential Random Variables: $\lambda = 5$



- The **probability density** can be greater than 1 at some points.

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

CDF of Exponential RV

- Its PDF is:

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

- We know that

$$\int_{-\infty}^{\infty} e^{ax} = \frac{1}{a} e^{ax}.$$

  or

$$\frac{d}{dx} e^{ax} = a e^{ax}.$$

- By definition of CDF,

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(t)dt = \int_{0}^{x} \lambda e^{-\lambda t} dt \\ &= -e^{-\lambda x}\Big|_0^x \\ &= 1 - e^{-\lambda x}, \text{ if } x \geq 0 \end{aligned}$$

- Thus,

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}.$$

Probability Mass

- Using these substitutions we can find the value of the probability mass for an interval $[a, b]$ as follows:

$$\begin{aligned} P(a < X < b) &= \int_a^b \lambda e^{-\lambda x} dx \\ &= \lambda \int_a^b e^{-\lambda x} dx \\ &= -e^{-\lambda x}\Big|_a^b \\ &= -(e^{-\lambda b}) - (-e^{-\lambda a}) \\ &= e^{-\lambda a} - e^{-\lambda b}. \end{aligned}$$

- Or Similarly

$$\begin{aligned} P(a < X < b) &= F_X(b) - F_X(a) \\ &= \left(1 - e^{-\lambda b}\right) - \left(1 - e^{-\lambda a}\right) \\ &= e^{-\lambda a} - e^{-\lambda b}. \end{aligned}$$

Mean and Variance of Exponential RV

- The mean and the variance can be calculated as

$$E[X] = \frac{1}{\lambda} \text{ and}$$

$$var(X) = \frac{1}{\lambda^2}$$

- Show this by using the following:
  - Integration by parts: $\int udv = uv - \int vdu$
  - $\int_{-\infty}^{\infty} e^{ax} dx = \frac{1}{a} e^{ax}$ and/or $\frac{d}{dx} e^{ax} = ae^{ax}$.

Example

- **Question**: Let the number of miles traveled by a car before its engine fails to function be governed by the exponential distribution with a mean of $100,000$ miles. What is the probability that a car's engine will fail during its first $50,000$ miles of operation?
- **Solution**: Since $E(X) = \frac{1}{\lambda}$ for an exponential random variable X. Thus $\lambda = 1/100000$. Then,

$$\begin{aligned} P(X < 50,000) = F_X(50,000) &= 1 - e^{-\lambda 50,000} \\ &= 1 - e^{-\frac{50,000}{100,000}} \\ &= 1 - e^{-\frac{1}{2}} \\ &= 0.3934 \end{aligned}$$

Normal (Gaussian) Random Variables

- A **normal random variable** X is a continuous random variable with probability density function:

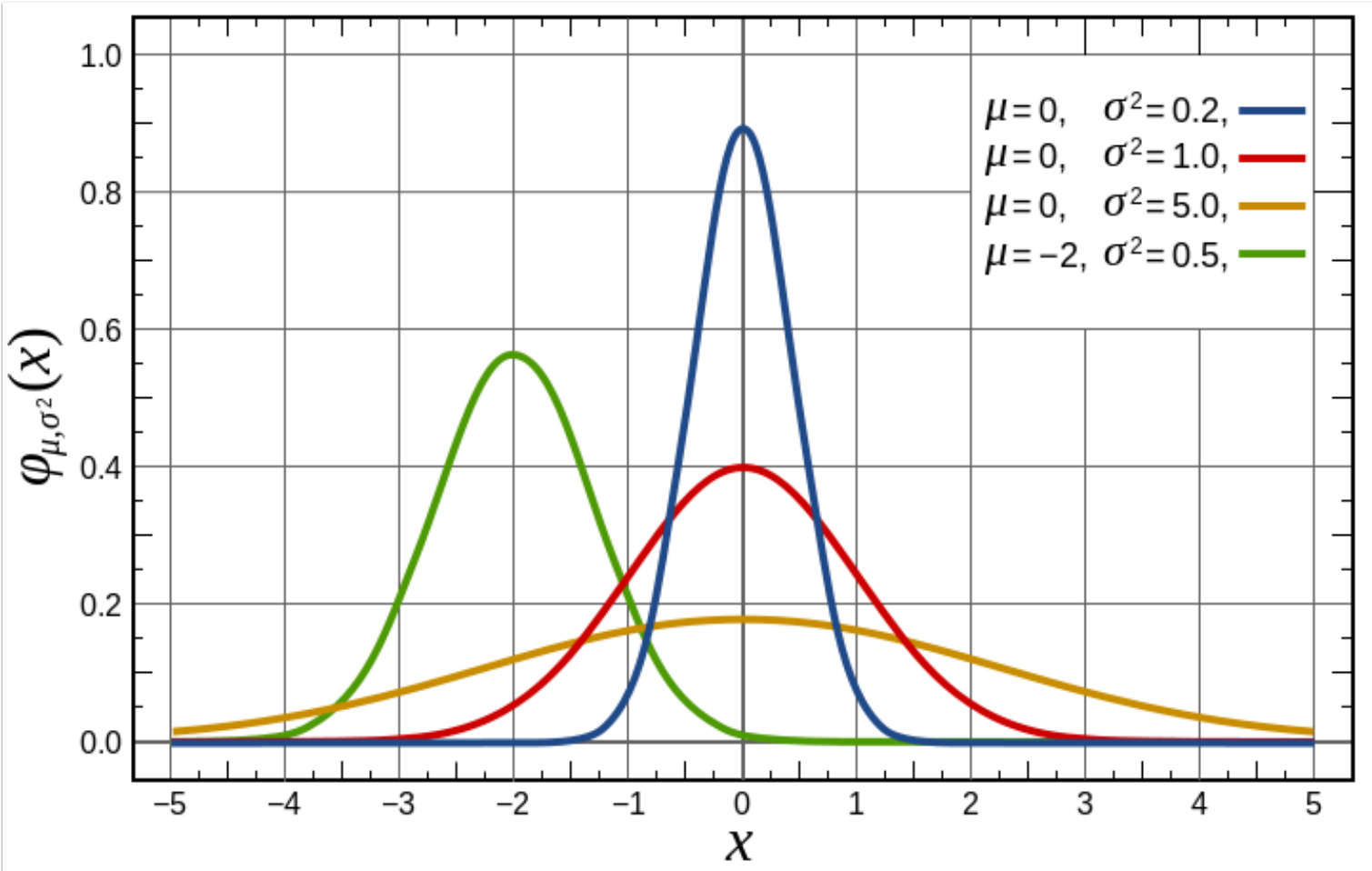
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- The range of this random variable is $\mathcal{X} = (-\infty, \infty)$.
- The parameter σ must be strictly greater than 0.
- The parameter μ can be any real value.
- Often also abbreviated as $X \sim \mathcal{N}(\mu, \sigma^2)$.
- Normal random variables have extremely important theoretical properties and are a default choice for modeling problems involving continuous measurements.

Why Study Normal RV?

- Normal RV plays an important role in a broad disciplines, including but not limited to computer science, engineering, physical, and statistical context.
- Few examples related to computer science include linear regression and Gaussian Process.
- The normal random variable is a convenient tool to approximate various types of phenomena (or observations), which allows us to derive mathematically tractable solutions.
- **The key fact is that the sum of a large number of independent and identically distributed (not necessarily normal) random variables has an approximately normal behavior (Central Limit Theorem).**

Normal (Gaussian) Random Variables



Mean, Variance, and CDF

- The mean and variance can be calculated to be

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

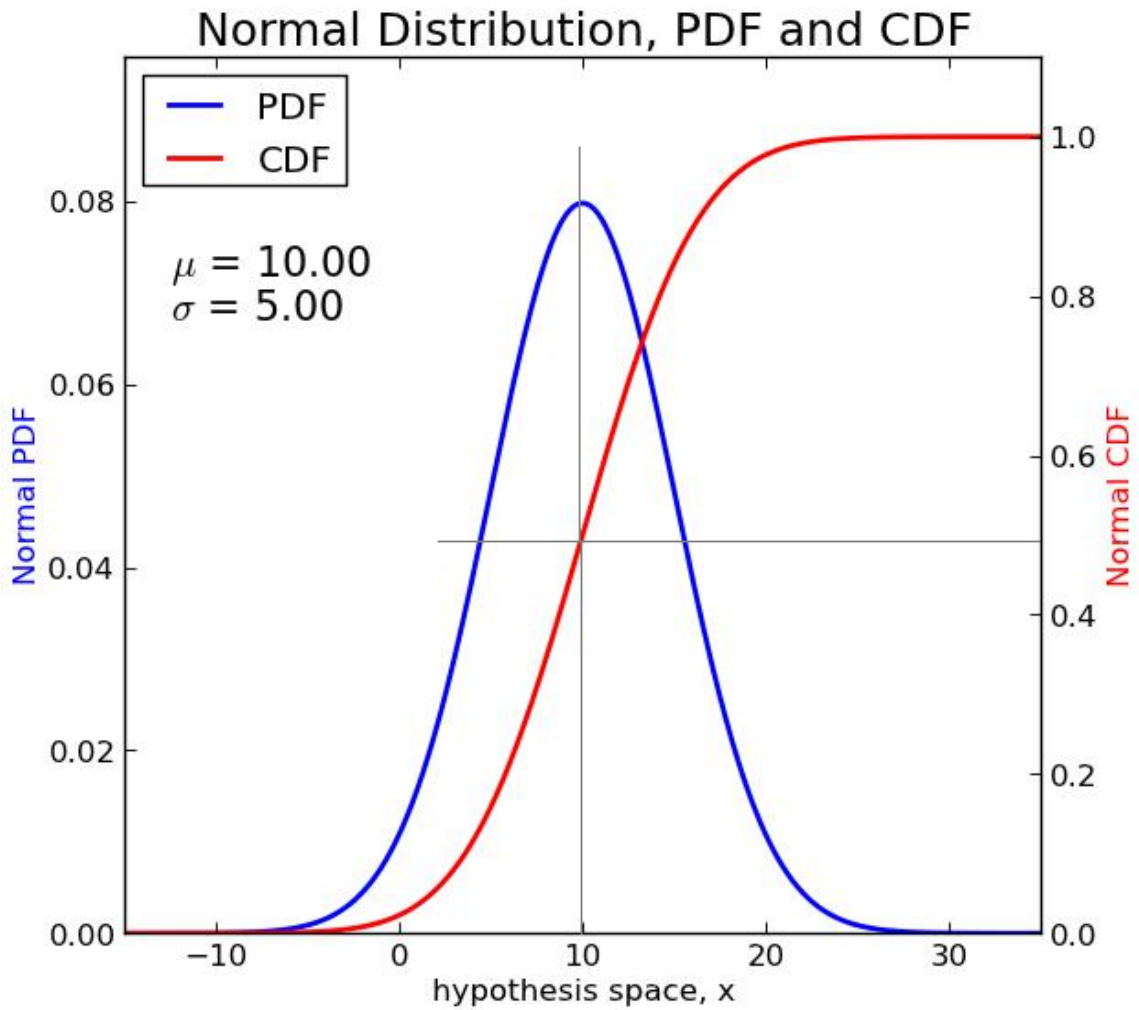
- Its CDF is defined as

$$F_X(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

- The **probability mass** of an interval $[a, b]$ is the definite integral:

$$\begin{aligned} P(a < X < b) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{a}^{b} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= F_X(b) - F_X(a) \end{aligned}$$

Normal (Gaussian) Random Variables



Preserving the Normality (or Gaussianity)

- Let X be a normal random variable with mean μ and variance σ^2, and if $a \neq 0$ and b are scalars, then a random variable

$$Y = aX + b,$$

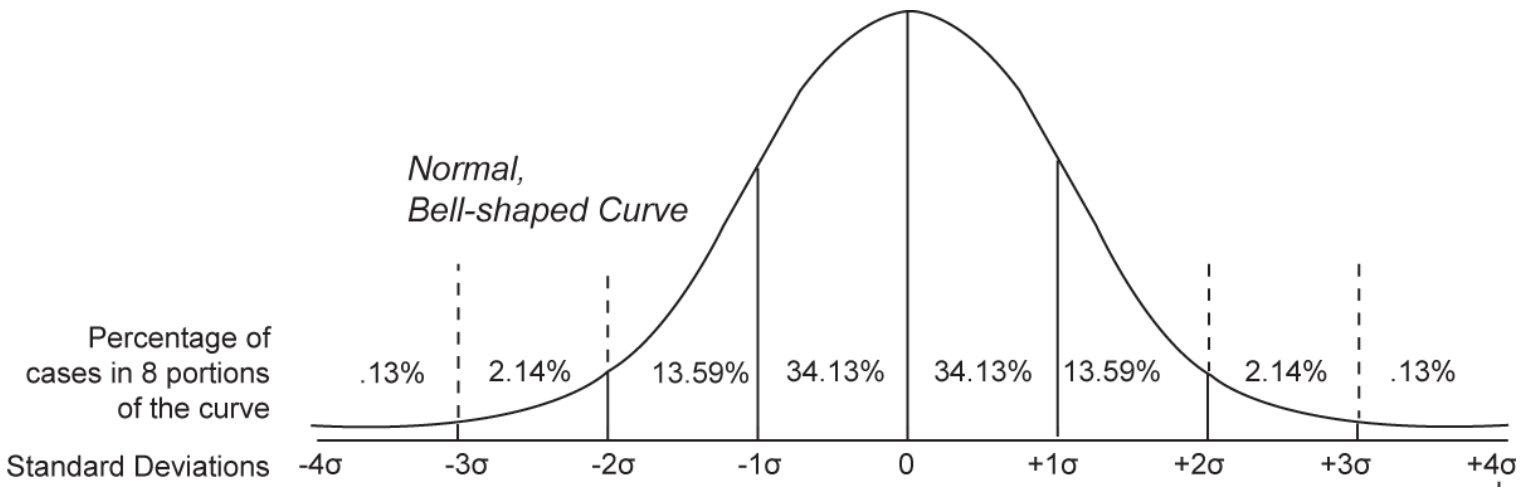
is also a normal random variable with

$$E[Y] = a\mu + b \text{ and } var(Y) = a^2\sigma^2.$$

The Standard Normal Random Variable

- A normal random variable X with $E[X] = 0$ and $var(X) = 1$ is said to be a **standard normal random variable**.
- Its PDF can be simplified as

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



The Standard Normal Random Variable

- Its CDF can be defined as

$$\Phi(x) = P(X \leq x) = P(X < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$$

The Standard Normal Table

$\Phi(x)$	0	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	.5	.50399	.50798	.51197	.51595	.51994	.52392	.5279	.53188	.53586
.1	.53983	.5438	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
.3	.61791	.62172	.62552	.6293	.63307	.63683	.64058	.64431	.64803	.65173
.4	.65542	.6591	.66276	.6664	.67003	.67364	.67724	.68082	.68439	.68793
.5	.69146	.69497	.69847	.70194	.7054	.70884	.71226	.71566	.71904	.7224
.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.7549
.7	.75804	.76115	.76424	.7673	.77035	.77337	.77637	.77935	.7823	.78524
.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.8665	.86864	.87076	.87286	.87493	.87698	.879	.881	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.9032	.9049	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.9222	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.9452	.9463	.94738	.94845	.9495	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.9608	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.9732	.97381	.97441	.975	.97558	.97615	.9767
2	.97725	.97778	.97831	.97882	.97932	.97982	.9803	.98077	.98124	.98169
2.1	.98214	.98257	.983	.98341	.98382	.98422	.98461	.985	.98537	.98574
2.2	.9861	.98645	.98679	.98713	.98745	.98778	.98809	.9884	.9887	.98899
2.3	.98928	.98956	.98983	.9901	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.9918	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.9943	.99446	.99461	.99477	.99492	.99506	.9952
2.6	.99534	.99547	.9956	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.9972	.99728	.99736
2.8	.99744	.99752	.9976	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861

Standardizing a Normal Variable

- For a given normal random variable X with mean μ and variance σ^2, you can standardize it by defining a new random variable Y given by

$$Y = \frac{X - \mu}{\sigma}$$

- Since Y is a form of $aX + b$, where $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$, we know that the normality of Y is preserved.
- The mean and variance of Y can be computed as

$$E[Y] = \frac{E[X] - \mu}{\sigma} = 0 \text{ and}$$

$$var(Y) = \frac{var(X)}{\sigma^2} = 1.$$

Example

- **Question**: The midterm for our CS240 class can be modeled as a normal random variable with a mean of $\mu = 75\%$ and standard deviation of $\sigma = 10\%$. What is the probability that a randomly chosen student has less than or equal to 80% of score?
- **Answer:** Let X be the score. Then, we first normalize X to have zero mean and unit variance.

$$Y = \frac{X - 75}{10}$$

- Then,

$$\begin{aligned} P(X < 80) &= P\left(Y < \frac{80 - 75}{10}\right) = P\left(Y < \frac{1}{2}\right) \\ &= \Phi\left(\frac{1}{2}\right) \\ &= 0.69146. \end{aligned}$$

Example

- **Question**: What is the probability that a randomly chosen student has greater than or equal to 80% of score?
- **Answer:**

$$P(X \geq 80) = 1 - P(X < 80) = 1 - \Phi\left(\frac{1}{2}\right) = 0.30854$$

Joint PDFs

A Joint PDF of Multiple RVs

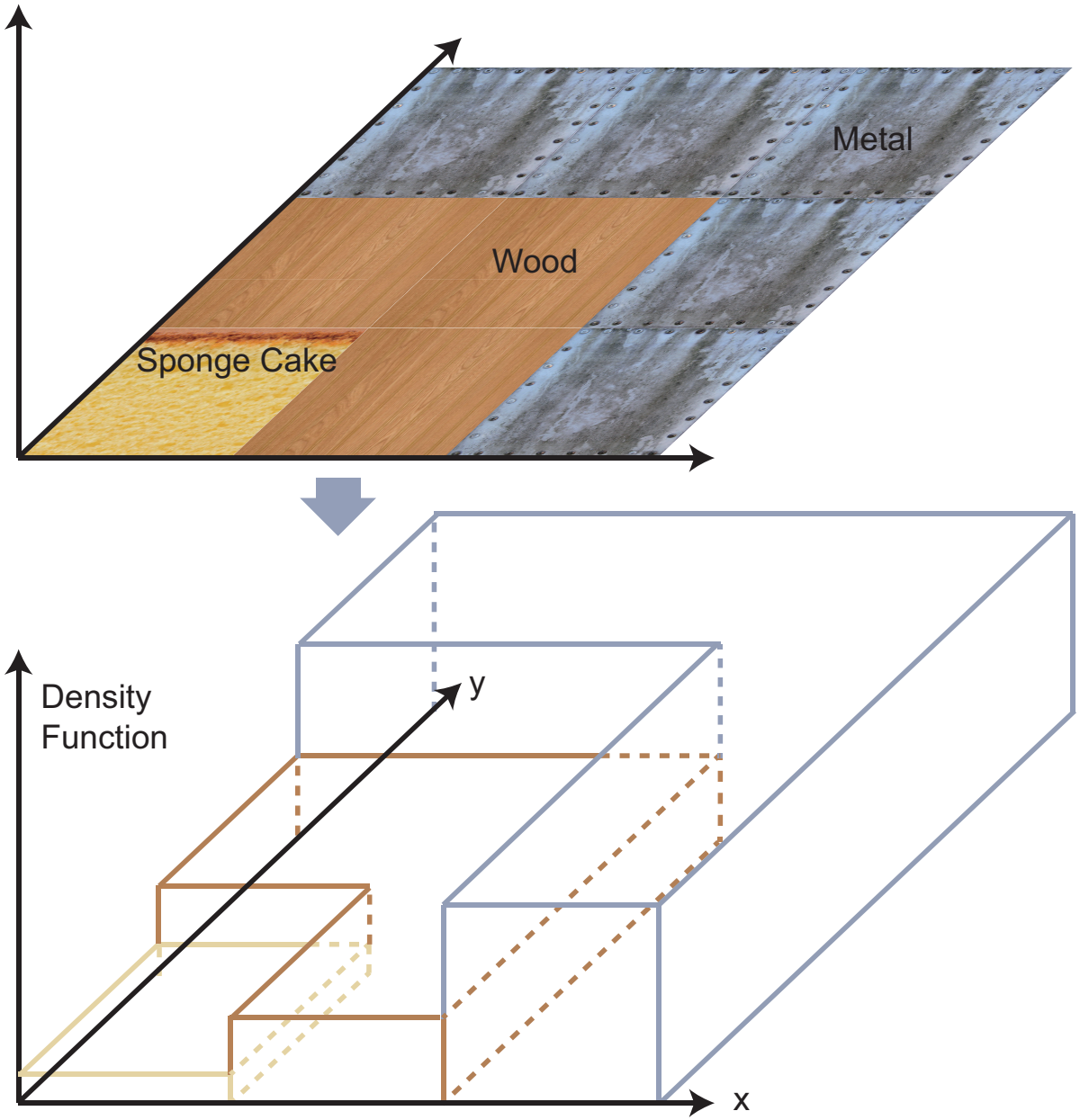
- We now consider a joint PDF of multiple random variables.
- We say that two continuous random variables associated with the same experiment are jointly continuous and have a joint PDF $f_{X,Y}$.

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

- If B is defined such that $B = \{(x,y)|a \leq x \leq b, c \leq y \leq d\}$, then

$$\begin{aligned} P(a \leq x \leq b, c \leq y \leq d) &= \int_c^d \int_a^b f_{X,Y}(x,y)dxdy \\ &= \int_a^b \int_c^d f_{X,Y}(x,y)dydx. \end{aligned}$$

Joint Normal Random Variables



A Joint PDF of Multiple RVs

- A joint PDF should satisfy:
  - **Non-negative**: $f_{X,Y}(x, y) \geq 0$ for all $(X, Y) \subseteq \mathcal{X}^2$
  - **Normalization**: $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$.
- We can compute **marginal PDFs** f_X and f_Y as

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

and

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

Example

- Let $f_{X,Y}(x,y)$ be a two-dimensional uniform PDF within $-1 \leq x \leq 1$ and $2 \leq y \leq 6$.

$$f_{X,Y}(x,y) = \begin{cases} c, & \text{if } -1 \leq x \leq 1 \text{ and } 2 \leq y \leq 6 \\ 0, & \text{otherwise,} \end{cases}$$

  Then, what is $P(0 \leq x \leq 1, 2 \leq y \leq 3)$?
- **Solution**: We know that

$$\int_2^6 \int_{-1}^1 c dx dy = 1.$$

- Then, we know that $c = \frac{1}{8}$
- Then,

$$\begin{aligned} P(0 \leq 1 \leq b, 2 \leq y \leq 3) &= \int_2^3 \int_0^1 \frac{1}{8} dx dy \\ &= \frac{1}{8} \end{aligned}$$

Joint CDF

- We define a joint CDF of two RVs X and Y as

$$\begin{aligned} F_{X,Y}(x,y) &= P(X \leq x, Y \leq y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t) dt ds \end{aligned}$$

- Conversely, the joint PDF can be derived from the joint CDF as

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

Expectation

- If X and Y are random variables, then $Z = g(X, Y)$ is also a random variable.
- The expected value of Z can be computed as

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

- Note that when $Z = X$, then we can compute the expected value of X.
- If $g(X, Y)$ is a linear function of X and Y, e.g., $g(X, Y) = aX + bY + c$, we have

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

Correlation and Causation

Quantifying Dependence: Covariance

- The **covariance** between any two RVs (either discrete or continuous) X and Y is one measure of dependence that quantifies the degree to which there is a **linear relationship** between X and Y.

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

- If X and Y are independent then $cov(X, Y) = 0$.
- However, $cov(X, Y) = 0$ does not necessarily imply that X and Y are independent.
- Note that $\text{cov}(X, X) = \text{var}(X)$.
- For a constant a, $\text{cov}(X, aY + b) = a \cdot \text{cov}(X, Y)$.
- Note that $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$.
- More generalized equation

$$\text{cov}\left(X, \sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} \text{cov}(X, Y_i)$$

Quantifying Dependence: Correlation

- The range of $\text{cov}(X, Y)$ values depends on the means of X, Y, and XY.
- The **correlation** ρ between X and Y is closely related to the covariance, but is *normalized* to the range $[-1, 1]$:

$$\rho(X, Y) = corr(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$$

- $\rho = 1$ indicates maximum positive covariance (e.g., $\rho(X, X)$) and $\rho = -1$ indicates maximum negative covariance (e.g., $\rho(X, -X)$).

Causation

- **Question:** When two random variables are correlated does this mean one random variable causes the other?
- **Example:** In the height/weight example, height and weight were positively correlated. Does increasing your weight make you taller?
- **Example:** There are more fireman at the scene of larger fires. Do fireman cause an increase in the size of a fire?
- **Example:** More people drown on days where a lot of ice cream is sold. Does ice cream cause drowning?

Causation

Given two correlated random variables X and Y:

- X might cause Y (i.e., causation)
- Y might cause X (i.e., reverse causation)
- A third random variable Z might cause X and Y (i.e., common cause)
- A combination of all of these (e.g., self-reinforcement)
- The correlation might be spurious due to small sample size

Limit Theorems

Overview

- Let $X_1, X_2, \cdots, X_n$ be a sequence of i.i.d. (either discrete or continuous) random variables with mean of μ and variance of σ^2.
- Limit theorems are mostly concerned with the sum of these random variables (which forms another random variable):

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

  especially when n is very large.
- Then, the mean and variance of S_n can be computed as

$$E[S_n] = E[X_1] + E[X_2] + \cdots + E[X_n] = n\mu$$

$$\text{var}(S_n) = \text{var}(X_1) + \text{var}(X_2) + \cdots + \text{var}(X_n) = n\sigma^2$$

Overview

- Let us introduce a new RV:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

- The mean and variance of Z_n are

$$E[Z_n] = 0$$

$$\text{var}(Z_n) = 1$$

- The **central limit theorem** states that the distribution of Z_n becomes the **standard normal variable** as n increases.

Markov and Chebyshev Bounds

Markov Bound $\rightarrow$ Chebyshev Bound $\rightarrow$ Central Limit Theorem

- **Markov Bound:**
  - Informally: If a nonnegative RV has a small mean, then the probability that this RV takes a large value must also be small.
  - Formally: For a non-negative random variable X,

$$P(X \geq a) \leq \frac{E(X)}{a}$$

- **Chebyshev Bound:**
  - Informally: If a RV has small variance, then the probability that it takes a value far from its mean is also small. Note that the Chebyshev inequality does not require the random variable to be nonnegative.
  - Formally: For a random variable X,

$$P(|X - E(X)| \geq c) \leq \frac{Var(X)}{c^2}$$

- The mean and the variance of a RV are only a rough summary of its properties, and we cannot expect the bounds to be close approximations of the exact probabilities.

An Alternate Form of The Chebyshev Bound

- Let $c = k\sigma$, then

$$P(|X - \mu| \geq k\sigma) = P\left(\left|\frac{X - \mu}{\sigma}\right| \geq k\right) = \frac{1}{k^2}$$

- The probability that a RV takes a value more than k standard deviations from its mean is at most $1/k^2$.

Weak law of large numbers & Convergence in probability

Sample Mean

- Let $X_1, X_2, \cdots, X_n$ be a sequence of i.i.d. (either discrete or continuous) random variables with mean of μ and variance of σ^2.
- Its sample (empirical) mean can be computed as

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  Note that $\overline{X}_n$ is also a random variable.
- We know that the expected value of the sample mean is

$$E\left[\overline{X}_n\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu$$
$$= \mu$$

- We also know that the variance and standard deviations of the sample mean are

$$\text{Var}(\overline{X}_n) = \text{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2} \cdot n \cdot \sigma^2$$
$$= \frac{\sigma^2}{n}$$
$$Std(\overline{X}_n) = \frac{\sigma}{\sqrt{n}}$$

The Weak Law of Large Numbers

- Let $X_1, X_2, \ldots$ be a sequence of i.i.d. (either discrete or continuous) random variables with mean μ and variance σ^2. For every $\epsilon > 0$, we have

$$P\left(|\overline{X}_n - \mu| \geq \epsilon\right) \to 0 \text{ as } n \to \infty.$$

- The weak law of large numbers states that if n is large, the bulk of the distribution of $\overline{X}_n$ will converge to (be concentrated around) μ.
- That is, if we consider a positive length interval $[\mu - \epsilon, \mu + \epsilon]$ around μ, then there is high probability that $\overline{X}_n$ will fall in that interval; as $n \to \infty$, this probability converges to 1. If ϵ is very small, we may have to wait longer (i.e., need a larger value of n) before this probability converges to 1.

Example 1

- Consider an event A with probability $p = P(A)$.
- We repeat the experiment n times.
- Let $\overline{X}_n$ be the fraction of time that event A occurs. This is the **empirical frequency** of A

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n},$$

  where $X_i = 1$ whenever A occurs, and 0 otherwise; thus $E[X_i] = p$.
- The weak law applies and shows that when n is large, the empirical frequency is most likely to be within ϵ of p.
- Loosely speaking, this allows us to conclude that empirical frequencies are faithful estimates of p.
- Alternatively, this is a step towards interpreting the probability p as the frequency of occurrence of A.

Convergence in probability

- Let $Y_1, Y2, \ldots$ be a sequence of random variables (not necessarily independent), and let a be a real number.
- We say that the sequence Y_n **converges to** a **in probability**, if for every $\epsilon > 0$, we have

$$\lim_{n \to \infty} P(|Y_n - a| \geq \epsilon) = 0$$

- Put it another way: $\forall \epsilon, \delta > 0, \exists n_0$ such that $\forall n \geq n_0$

$$P(|Y_n - a| \geq \epsilon) \leq \delta$$

Our measurement is accurate, with this much confidence.

Usefulness of limit theorems

- Conceptually, they provide an interpretation of expectations (as well as probabilities) is terms of a long sequence of identical independent experiments.
- They allow for an *approximate* analysis of the properties of random variables such as X_n. This is to be contrasted with an *exact* analysis which would require a formula for the PMF or PDF of X_n, a complicated and tedious task when n is large.
- They play a major role in inference and statistics, in the presence of large data sets.

Central limit theorem &
The strong law of large numbers

The Strong Law of Large Numbers

- Let $X_1, X_2, \cdots$ be a sequence of i.i.d. (either discrete or continuous) random variable with mean μ and variance σ^2.
- Then, the sequence of sample mean $\overline{X}_n$ converges to μ as $n \to \infty$, with probability 1:

$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

- Its sample mean $\overline{X}_n$, which is a RV, will **converge** to the true mean μ, which is a constant, with a probability 1 when we have an infinitely large sample size.
  - More specifically, an event of $\overline{X}_n = \mu$ has a probability of 1.
- **Example**: Let $X_i \sim \text{Bern}(p)$, then

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = p\right) = 1.$$

The Central Limit Theorem

- Let us define a variable by normalizing $\overline{X}_n$ with its mean and standard deviation
  - In the same manner as we normalized a Normal RV to derive the Standard Normal RV.

$$Z_n = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

or equivalently

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

- Then, the PDF of Z_n converges to the standard normal PDF as $n \to \infty$

$$Z_n \sim N(0,1) \text{ as } n \to \infty$$

The Central Limit Theorem

- The CLT is surprisingly general and extremely powerful.
- It states that X_i can have any forms of (discrete, continuous, or a mixture) probability distribution, but its sample mean converges to a Standard Normal distribution as n becomes large.
- Conceptually, this is important as it indicates that the sum of a large number of i.i.d RV is approximately normal.
- Practically, this is important as it eliminates the need for detailed probabilistic models as long as we have a large sample size. We can still approximate its sample mean using the Standard Normal distribution as long as we know μ and σ.

Game Theory

Game Theory

- Game theory studies what happens when self-interested agents interact.
- Each player (or agent) has his/her own description of which states of the world he/she likes (or the states that benefits the agent the most).
- A player's benefits can be represented using a **payoff matrix** that maps the *states* to *real numbers*.
- All players have pure strategy if all players select a single action and play it.

Strict Domination

- A strategy s_i of a player P_i strictly dominates another strategy s_i' of the player, if s_i generates a greater payoff than s_i'.
- More formally, let P_i represent a player i and P_{-i} represent all other players but i. Furthermore, Let s_i and s_i' be two strategies of P_i, and S_{-i} be the set of all strategies of the remaining players.
- Then, s_i strictly dominates s_i' if for all $s_{-i} \in S_{-i}$, it is the case that $u_i(s_i, s_{-i}) \geq u_i(s_i', s_{-i})$, where u_i is the payoff function.

Iterated Elimination of Strictly Dominated Strategies (IESDS)

- A strategy s_i of a player P_i is **strictly dominated** if some (not all) other strategies s_i' strictly
- Intuitively, all strategies that are strictly dominated by other strategies can be ignored, since they can never be best responses to any moves by the other players.
- Games that cannot be solved with strictly dominating strategies can be solved with Nash equilibrium.

Nash Equilibrium

Definition

A Nash Equilibrium is a set of strategies for *each player* where no change by one player alone can improve his/her outcome; each player has no incentive to change his/her rather stable strategy.

- We are considering Nash equilibrium where players do not randomize between two or more strategies (called Pure Strategy Nash Equilibrium).
- We only care about alternating strategies in an individual level, not in a group level where everyone collectively change strategies toward a single strategy.
- A Nash Equilibrium is a law that no one would want to break, even in the absence of an effective police force.
- **There may exist multiple Nash equilibria.**

Theorem (Nash)

Every game where each player has a finite number of options, has at least one Nash equilibrium.

Iterated Elimination of Strictly Dominated Strategies vs. Nash Equilibrium

- Strictly dominated strategies cannot be a part of a Nash equilibrium.
- After completing the IESDS, if there exists only one strategy for each player remaining, that strategy set is the unique Nash equilibrium.
- Even if there exists no solutions from the IESDS, there may exist Nash Equilibria.

Pure Strategies vs. Mixed Strategies

- **Pure Strategy**: Players choose a strategy to select a single action and play it - so far we have considered this scenario.
- **Mixed Strategy**: Players randomize over the set of available actions according to some probability distribution - a player randomizes and mixes between different actions.

Zero-Sum Games

Definition

A *Two-player zero-sum game* consists of a set of actions A_i for Player P_i and A_j for Player P_j, where each strategy profile $a \in A_i \times A_j$ has the payoff function $u_1(a) + u_2(a) = 0$.

E.g., two-finger Morra, with payoff matrix

	1 B Finger	2 B Finger
1 A Finger	$+2, -2$	$-3, +3$
2 A Finger	$-3, +3$	$+4, -4$

There exists no clear Nash Equilibrium when we consider pure strategies. However, based on Nash theorem, one must exist in **mixed strategies**.

Analysis of Two-finger Morra

- Suppose Bob randomizes his action by playing "1" with probability q and "2" with probability $1 - q$

$$P(B = 1) = q \text{ and } P(B = 2) = 1 - q.$$

- If Alice plays "1" then Alice has expected payoff

$$2q - 3(1 - q) = 5q - 3$$

- If Alice plays "2" then Alice has expected payoff

$$-3q + 4(1 - q) = 4 - 7q$$

- If Alice's payoffs are equal, then Alice does not have to prefer one action over the other, such that she does not expect to do better by changing her strategy (i.e., changing the value of p).
  - This meets the definition of a Nash equilibrium.
  - Do not confuse between the definitions of a strategy vs. an action.

Analysis of Two-finger Morra

- That is, when Bob's strategy makes the payoffs of Alice's actions equal

$$5q - 3 = 4 - 7q$$

  or equivalently when

$$q = 7/12$$

- Then, Alice can choose also a strategy (since the payoffs are indifferent for her) to make Bob's actions to have the equal payoff.
- Similarly, that is when Alice's strategy is $p = 7/12$.
- This also means that Bob is completely satisfied for playing a mixed strategy with $q = 7/12$.
- Hence $p = q = 7/12$ is a mixed-strategy Nash equilibrium; players do not have any incentives to change their strategies.

Hawks and Doves

- Previous discussed hawks and doves example with the following payoff matrix have both the 1) pure strategy and 2) mixed strategy Nash equilibrium

	B is a Hawk	B is a Dove
A is a Hawk	$-25, -25$	$50, 0$
A is a Dove	$0, 50$	$15, 15$

- A plays hawk and B plays Dove (or vice versa) is a pure-strategy Nash equilibrium.
- A and B play hawks with $p = q = 7/12$ is a mixed-strategy Nash equilibrium.
- The three Nash equilibria can be summarized as
  - $p = 0$ and $q = 1$ (Pure Strategy)
  - $p = 1$ and $q = 0$ (Pure Strategy)
  - $p = 7/12$ and $q = 7/12$ (Mixed Strategy)