CSE 312

Foundations of Computing II

Lecture 28: Victory Lap, What’s Next, & Review

What you've learned …

The essentials of probability and some statistics,

hands-on applications,

- Naïve Bayes SPAM filtering
- Bloom Filters
- MinHash for Distinct Elements
- Markov Chains and PageRank

some societal impact,

- Differential privacy
- Algorithmic fairness

and some Python…

a great headstart for CSE 446 (ML)

What’s next?

- Some places to apply and extend your knowledge
  - CSE 421 Algorithms - counting and more basic probability
  - CSE 422 Toolkit for Modern Algorithms – probability everywhere
  - CSE 426 Cryptography – randomness, reasoning about probability essential
  - CSE 427 Computational Biology
  - CSE 446 Machine Learning – this course + linear algebra essential
  - CSE 447 Natural Language Processing
  - CSE 473 Artificial Intelligence – Bayes nets, probability, etc.
  - CSE 490Q Quantum Computing – the quantum world is inherently random

Agenda

- What you've learned
- What's next
- Review

Counting: Sum & Product Rules

- **Sum rule:**

  If you can choose from

  - EITHER one of n options,
  - OR one of m options with NO overlap with the previous n,

  then the number of possible outcomes of the experiment is $n + m$

- **Product rule:**

  In a sequential process, if there are

  - n_1 choices for the 1^{st} step,
  - n_2 choices for the 2^{nd} step (given the first choice), ..., and
  - n_k choices for the k^{th} step (given the previous choices),

  then the total number of outcomes is $n_1 \times n_2 \times n_3 \times \cdots \times n_k$

Counting: Permutations & Combinations

Permutations. The number of orderings of n distinct objects

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

Example: How many sequences in $\{1,2,3\}^3$ *with no repeating elements?*

k-Permutations. The number of orderings of **only** k out of n distinct objects

$$P(n,k)$$

$$= n \times (n-1) \times \cdots \times (n-k+1)$$

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

Example: How many sequences of 5 distinct alphabet letters from $\{A, B, \ldots, Z\}$*?*

Combinations / Binomial Coefficient. The number of ways to select k out of n objects, where ordering of the selected k does not matter:

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

Example: How many size-5 ***subsets*** *of* $\{A, B, \ldots, Z\}$*?*

Example: How many shortest paths from Gates to Starbucks?

Example: How many solutions $(x_1, \ldots, x_k)$ *such that* $x_1, \ldots, x_k \geq 0$ *and* $\sum_{i=1}^{k} x_i = n$*?*

Counting: When order only *partly* matters

We often want to count # of partly ordered lists:

Let M = # of ways to produce fully ordered lists

P = # of partly ordered lists

N = # of ways to produce corresponding fully ordered list given a partly ordered list

Then $M = P \cdot N$ by the product rule. Often M and N are easy to compute:

$$P = M/N$$

Dividing by N "removes" part of the order.

Multinomial Coefficients

If we have k types of objects ($\boldsymbol{n}$ total), with $\boldsymbol{n_1}$ of the first type, $\boldsymbol{n_2}$ of the second, …, and $\boldsymbol{n_k}$ of the k^{th}, then the number of orderings possible is

$$\binom{n}{n_1, n_2, \cdots, n_k} = \frac{n!}{n_1!\, n_2! \cdots n_k!}$$

Counting using binary encoding/star and bars

The number of ways to distribute n indistinguishable balls into k distinguishable bins is

$$\binom{n+k-1}{k-1} = \binom{n+k-1}{n}$$

E.g., # of ways to add k non-negative integers up to n

Encode using one symbol (1 or *) for items, the other (0 or |) for dividers

Counting: Binomial Theorem

Theorem. Let $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$ a positive integer. Then,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Counting: Inclusion-Exclusion

Let A, B be sets. Then

$$|A \cup B| = |A| + |B| - |A \cap B|$$

In general, if $A_1, A_2, \ldots, A_n$ are sets, then

$$|A_1 \cup A_2 \cup \cdots \cup A_n| = singles - doubles + triples - quads + \ldots$$
$$= (|A_1| + \cdots + |A_n|) - (|A_1 \cap A_2| + \ldots + |A_{n-1} \cap A_n|) + \ldots$$

Counting: Pigeonhole Principle

If there are n pigeons in $k < n$ holes, then one hole must contain at least $\left\lceil \frac{n}{k} \right\rceil$ pigeons!

Reason. Can't have fractional number of pigeons

Syntax reminder:

- Ceiling: $\lceil x \rceil$ is x rounded up to the nearest integer (e.g., $\lceil 2.731 \rceil = 3$)
- Floor: $\lfloor x \rfloor$ is x rounded down to the nearest integer (e.g., $\lfloor 2.731 \rfloor = 2$)

Probability

Definition. A **sample space** Ω is the set of all possible outcomes of an experiment.

Examples:

- Single coin flip: $\Omega = \{H, T\}$
- Two coin flips: $\Omega = \{HH, HT, TH, TT\}$
- Roll of a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$

Definition. An **event** $E \subseteq \Omega$ is a subset of possible outcomes.

Examples:

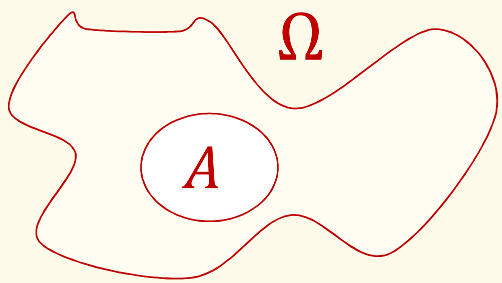
- Getting at least one head in two coin flips: $E = \{HH, HT, TH\}$
- Rolling an even number on a die : $E = \{2, 4, 6\}$

Discrete Probability

Definition. A (discrete) **probability space** is a pair (Ω, P) where:

- Ω is a set called the **sample space.**
- P is the **probability measure,**

  a function $P: \Omega \to \mathbb{R}$ such that:
  - $P(x) \geq 0$ for all $x \in \Omega$
  - $\sum_{x\in\Omega} P(x) = 1$



For $A \subseteq \Omega$:

$$P(A) = \sum_{x\in A} P(x)$$

Random Variables (Discrete Case)

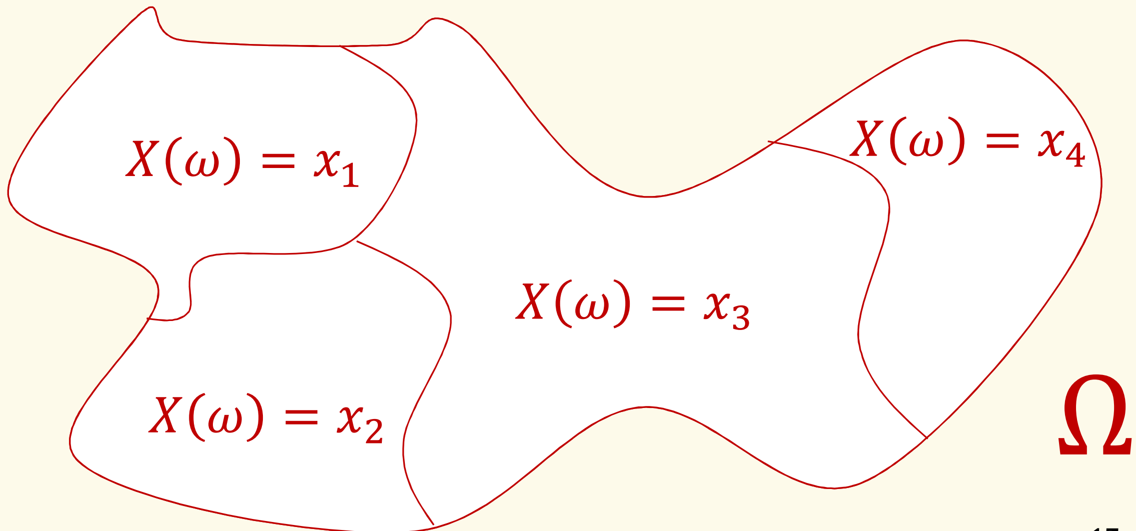
Definition. A **random variable (RV)** for a probability space (Ω, P) is a function $X: \Omega \to \mathbb{R}$.

The set of values that X can take on is its *range/support*: $X(\Omega)$ or Ω_X

$\{X = x_i\} = \{\omega \in \Omega \mid X(\omega) = x_i\}$

Random variables **partition** the sample space.

$\Sigma_{x \in X(\Omega)} P(X = x) = 1$



Probability Mass Function (PMF) and CDF (Discrete Case)

Definitions:

For a RV $X: \Omega \to \mathbb{R}$, the probability mass function (pmf) of X specifies, for any real number x, the probability that $X = x$

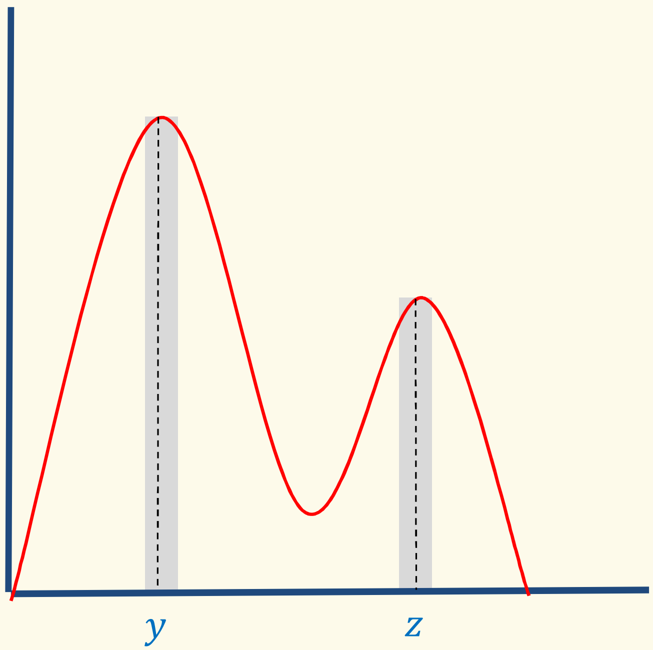
$$p_X(x) = P(X = x) = P(\{\omega \in \Omega \mid X(\omega) = x\})$$

$$\boxed{\sum_{x \in \Omega_X} p_X(x) = 1}$$

For a RV $X: \Omega \to \mathbb{R}$, the cumulative distribution function (cdf) of X specifies, for any real number x, the probability that $X \leq x$

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

Probability Density Function



Non-negativity: $f_X(x) \geq 0$ for all $x \in \mathbb{R}$

Normalization: $\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1$

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x$$

$$\frac{P(X \approx y)}{P(X \approx z)} \approx \frac{\epsilon f_X(y)}{\epsilon f_X(z)} = \frac{f_X(y)}{f_X(z)}$$

What $f_X(x)$ measures: The local ***rate*** at which probability accumulates

Cumulative Distribution Function (Continuous Case)

Definition. The **cumulative distribution function (cdf)** of X is

$$F_X(a) = P(X \leq a) = \int_{-\infty}^{a} f_X(x)\,\mathrm{d}x$$

By the fundamental theorem of Calculus $f_X(x) = \frac{d}{dx} F_X(x)$

Therefore: $P(X \in [a, b]) = F_X(b) - F_X(a)$

F_X is monotone increasing, since $f_X(x) \geq 0$. That is $F_X(c) \leq F_X(d)$ for $c \leq d$

Continuous Random Variables

Probability Density Function (PDF).

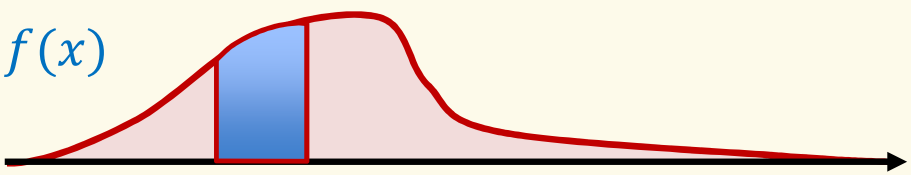
$f: \mathbb{R} \to \mathbb{R}$ s.t.

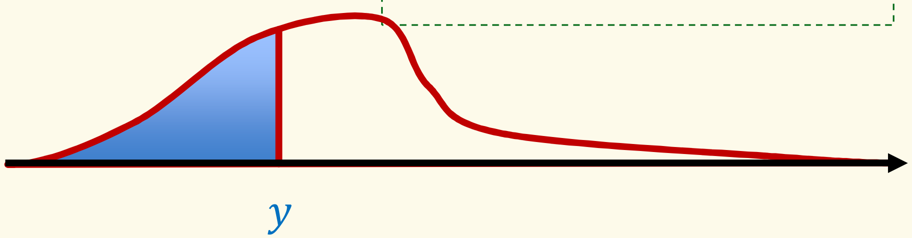
- $f(x) \geq 0$ for all $x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$

Cumulative Distribution Function (CDF).

$$F(y) = \int_{-\infty}^{y} f(x)\,\mathrm{d}x$$

Theorem. $f(x) = \frac{dF(x)}{dx}$





Density ≠ Probability !

$$P(X \in [a,b]) = \int_a^b f_X(x)\mathrm{d}x$$
$$= F_X(b) - F_X(a)$$

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

Probability: Inclusion-Exclusion

Let A, B be events. Then

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

In general, if $A_1, A_2, \ldots, A_n$ are events, then

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) &= \mathit{singles} - \mathit{doubles} + \mathit{triples} - \mathit{quads} + \ \ldots \\ &= (P(A_1) + \cdots + P(A_n)) \\ &\quad -(P(A_1 \cap A_2) + \ \ldots + P(A_{n-2} \cap A_n) + \ P(A_{n-1} \cap A_n)) \\ &\quad + \ \ldots \end{aligned}$$

Conditional Probability

- **Conditional Probability**

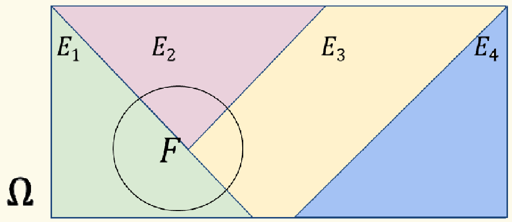
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- **Bayes Theorem**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{if } P(A) \neq 0, P(B) \neq 0$$

- **Law of Total Probability**

$E_1, \dots, E_n$ partition Ω



$$P(F) = \sum_{i=1}^{n} P(F \cap E_i) = \sum_{i=1}^{n} P(F|E_i)\, P(E_i)$$

Bayes Theorem with Law of Total Probability

Bayes Theorem with LTP: Let $E_1, E_2, \dots, E_n$ be a partition of the sample space, and F and event. Then,

$$P(E_1|F) = \frac{P(F|E_1)P(E_1)}{P(F)} = \frac{P(F|E_1)P(E_1)}{\sum_{i=1}^{n} P(F|E_i)P(E_i)}$$

Simple Partition: In particular, if E is an event with non-zero probability, then

$$P(E|F) = \frac{P(F|E)P(E)}{P(F|E)P(E) + P(F|E^C)P(E^C)}$$

Chain rule & Independence

Theorem. (Chain Rule) For events $A_1, A_2, \dots, A_n$,

$$P(A_1 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \\ \cdots P(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1})$$

Definition. Two events A and A are (statistically) **independent** if

$$P(A \cap B) = P(A) \cdot P(B).$$

"Equivalently." $P(A|B) = P(A)$.

Definition. Two events A and B are **independent conditioned on C** if $P(C) \neq 0$ and $P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C)$.

Multiple Events – Mutual Independence

Definition. Events $A_1, \dots, A_n$ are **mutually independent** if for every non-empty subset $I \subseteq \{1, \dots, n\}$, we have

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i).$$

Expected Value of a Random Variable (Discrete Case)

Definition. Given a discrete RV $X: \Omega \to \mathbb{R}$, the **expectation** or **expected value** or **mean** of X is

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot P(\omega)$$

or equivalently

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x \cdot P(X = x) = \sum_{x \in \Omega_X} x \cdot p_X(x)$$

Intuition: "Weighted average" of the possible outcomes (weighted by probability)

Linearity of Expectation

Theorem. For any random variables $X_1, \dots, X_n$, and real numbers $a_1, \dots, a_n \in \mathbb{R}$,

$$\mathbb{E}[a_1X_1 + \cdots + a_nX_n] = a_1\mathbb{E}[X_1] + \cdots + a_n\mathbb{E}[X_n].$$

Very important: In general, we do <u>not</u> have $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Linearity of Expectation with Indicator Variables.

We flip n coins, each one heads with probability p

Z is the number of heads, what is $\mathbb{E}[Z]$?

- $X_i = \begin{cases} 1, & i^{\text{th}} \text{ coin flip is heads} \\ 0, & i^{\text{th}} \text{ coin flip is tails.} \end{cases}$

Fact. $Z = X_1 + \cdots + X_n$

Linearity of Expectation:

$$\mathbb{E}[Z] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] \quad = n \cdot p$$

$P(X_i = 1) = p$
$P(X_i = 0) = 1 - p$

$$\mathbb{E}[X_i] = p \cdot 1 + (1 - p) \cdot 0 = p$$

No independence required for Linearity of Expectation

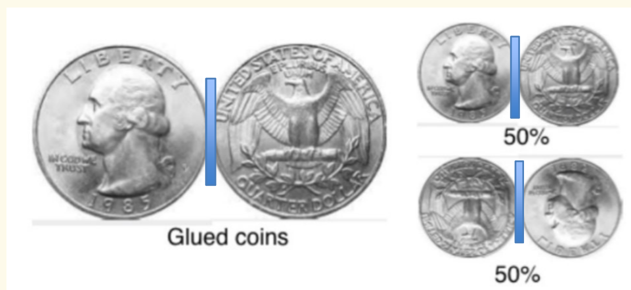
Each coin shows up heads half the time.

Two fair coins

$P(HT) = P(TH) = 0.25$
$P(HH) = P(TT) = 0.25$

$\mathbb{E}(X) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$

Glued coins



$P(HT) = P(TH) = 0.5$
$P(HH) = P(TT) = 0$

$\mathbb{E}(X) = 1 \cdot 1 = 1$

Attached coins

$P(HH) = P(TT) = 0.4$
$P(HT) = P(TH) = 0.1$

$\mathbb{E}(X) = 1 \cdot 0.2 + 2 \cdot 0.4 = 1$

LOTUS: Expected Value of $g(X)$ (Discrete Case)

Definition. Given a discrete RV $X : \Omega \to \mathbb{R}$, the **expectation** or **expected value** or **mean** of $g(X)$ is

$$\mathbb{E}[g(X)] = \sum_{\omega \in \Omega} g\big(X(\omega)\big) \cdot P(\omega)$$

or equivalently

$$\mathbb{E}[g(X)] = \sum_{x \in X(\Omega)} g(x) \cdot P(X = x) = \sum_{x \in \Omega_X} g(x) \cdot p_X(x)$$

Also known as **LOTUS:** "Law of the unconscious statistician

Linearity is special!

In general $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$

E.g., $X = \begin{cases} +1 \text{ with prob } 1/2 \\ -1 \text{ with prob } 1/2 \end{cases}$

Then: $\mathbb{E}[X^2] \neq \mathbb{E}[X]^2$

Variance (Discrete Case)

Definition. The **variance** of a (discrete) RV X is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_x p_X(x) \cdot (x - \mathbb{E}[X])^2$$

Theorem. For any $a, b \in \mathbb{R}$, $\mathrm{Var}(a \cdot X + b) = a^2 \cdot \mathrm{Var}(X)$

Theorem. $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Definition. The **standard deviation** of a (discrete) RV X is $\sigma_X = \sqrt{Var(X)}$

Note. For any $a \geq 0, b \in \mathbb{R}$, $\sigma_{a \cdot X + \mathrm{b}} = a \cdot \sigma_X$

Expectation & Variance of a Continuous Random Variable

Definition. The **expected value** of a continuous RV X is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} f_X(x) \cdot x \, \mathrm{d}x$$

Fact. $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$

Definition. The **variance** of a continuous RV X is defined as

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} f_X(x) \cdot (x - \mathbb{E}[X])^2 \, \mathrm{d}x = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

LOTUS: Expected Value of $g(X)$ (Continuous)

Definition. Given a continuous RV $X: \mathbb{R} \to \mathbb{R}$, the **expectation** or **expected value** or **mean** of $g(X)$ is

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

Review: From Discrete to Continuous

	Discrete	Continuous
PMF/PDF	$p_X(x) = P(X = x)$	$f_X(x) \neq P(X = x) = 0$
CDF	$F_X(x) = \sum_{t \leq x} p_X(t)$	$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$
Normalization	$\sum_{x} p_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$
Expectation	$\mathbb{E}[g(X)] = \sum_{x} g(x)\, p_X(x)$	$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$

Properties of Independent Random Variables

Theorem. If X, Y independent, $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Theorem. If X, Y independent, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Corollary. If $X_1, X_2, \ldots, X_n$ mutually independent,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i}^{n} \mathrm{Var}(X_i)$$

Joint PMFs and Joint Range

Definition. Let X and Y be discrete random variables. The **Joint PMF** of X and Y is

$$p_{X,Y}(a,b) = P(X = a, Y = b)$$

Definition. The **joint range** of $p_{X,Y}$ is

$$\Omega_{X,Y} = \{(c,d) : p_{X,Y}(c,d) > 0\} \subseteq \Omega_X \times \Omega_Y$$

Note that

$$\sum_{(s,t)\in\Omega_{X,Y}} p_{X,Y}(s,t) = 1$$

Marginal PMF

Definition. Let X and Y be discrete random variables and $p_{X,Y}(a,b)$ their joint PMF. The **marginal PMF** of X

$$p_X(a) = \sum_{b \in \Omega_Y} p_{X,Y}(a,b)$$

Similarly, $p_Y(b) = \sum_{a \in \Omega_X} p_{X,Y}(a,b)$

Continuous distributions on $\mathbb{R} \times \mathbb{R}$

Definition. The **joint probability density function (PDF)** of continuous random variables X and Y is a function $f_{X,Y}$ defined on $\mathbb{R} \times \mathbb{R}$ such that

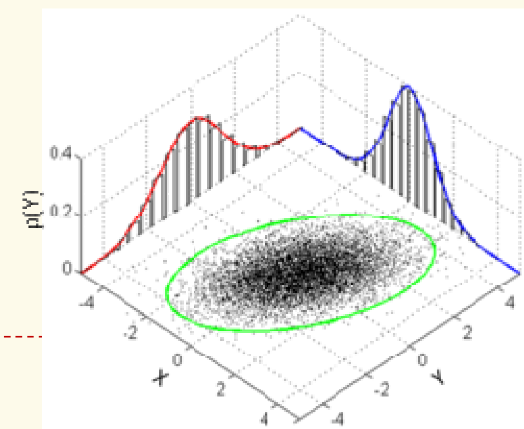
- $f_{X,Y}(x, y) \geq 0$ for all $x, y \in \mathbb{R}$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \mathrm{d}x \, \mathrm{d}y = 1$

for $A \subseteq \mathbb{R} \times \mathbb{R}$ the probability that $(X, Y) \in A$ is $\iint_A \; f_{X,Y}(x, y) \, \mathrm{d}x\mathrm{d}y$

The **(marginal) PDFs** f_X and f_Y are given by

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

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



Independence and joint distributions

Discrete random variables X and Y are independent iff

- $p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$ for all $x \in \Omega_X, y \in \Omega_Y$

Continuous random variables X and Y are independent iff

- $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$ for all $x, y \in \mathbb{R}$

Conditional Expectation

Definition. Let X be a discrete random variable then the **conditional expectation** of X given event A is

$$\mathbb{E}[X \mid A] = \sum_{x \in \Omega_X} x \cdot P(X = x \mid A)$$

Notes:

- Can be phrased as a "random variable version"

$$\mathbb{E}[X | Y = y]$$

- Linearity of expectation still applies here

$$\mathbb{E}[aX + bY + c \mid A] = a\,\mathbb{E}[X \mid A] + b\,\mathbb{E}[Y \mid A] + c$$

Law of Total Expectation

Law of Total Expectation (event version). Let X be a random variable and let events $A_1, \ldots, A_n$ partition the sample space. Then,

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

Law of Total Expectation (random variable version). Let X be a random variable and Y be a discrete random variable. Then,

$$\mathbb{E}[X] = \sum_{y \in \Omega_Y} \mathbb{E}[X \mid Y = y] \cdot P(Y = y)$$

Reference Sheet

	Discrete	Continuous
Joint PMF/PDF	$p_{X,Y}(x,y) = P(X=x, Y=y)$	$f_{X,Y}(x,y) \neq P(X=x, Y=y)$
Joint CDF	$F_{X,Y}(x,y) = \sum_{t \le x} \sum_{s \le y} p_{X,Y}(t,s)$	$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(t,s) ds dt$
Normalization	$\sum_{x} \sum_{y} p_{X,Y}(x,y) = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$
Marginal PMF/PDF	$p_X(x) = \sum_{y} p_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy$
Expectation	$E[g(X,Y)] = \sum_{x} \sum_{y} g(x,y) p_{X,Y}(x,y)$	$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dx dy$
Independence	$\forall x, y, p_{X,Y}(x,y) = p_X(x) p_Y(y)$	$\forall x, y, f_{X,Y}(x,y) = f_X(x) f_Y(y)$

Markov's and Chebyshev's Inequalities

Theorem (Markov's Inequality). Let X be a random variable taking only non-negative values. Then, for any $t > 0$,

$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Theorem (Chebyshev's Inequality). Let X be a random variable. Then, for any $t > 0$,

$$P(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Chernoff-Hoeffding Bound

Theorem. Let $X = X_1 + \cdots + X_n$ be a sum of independent RVs, each taking values in $[0,1]$, such that $\mathbb{E}[X] = \mu$. Then, for every $\delta \in [0,1]$,

$$P(|X - \mu| \geq \delta \cdot \mu) \leq e^{-\frac{\delta^2 \mu}{4}}.$$

Herman Chernoff, Herman Rubin, Wassily Hoeffding

Example: If $X \sim \mathrm{Bin}(n, p)$, then $X = X_1 + \cdots + X_n$ is a sum of independent $\{0,1\}$-Bernoulli variables, and $\mu = np$

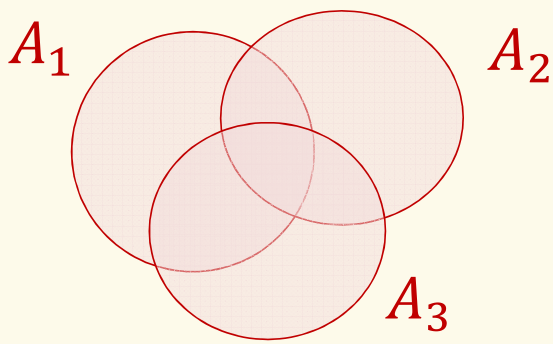
Note: More accurate versions are possible, but with more cumbersome right-hand side (e.g., see textbook)

Union Bound

Theorem (Union Bound). Let $A_1, \dots, A_n$ be arbitrary events. Then,

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

Intuition (3 evts.):



Bernoulli Random Variables

A random variable X that takes value 1 ("Success") with probability p, and 0 ("Failure") otherwise. X is called a **Bernoulli random variable**.

Notation: $X \sim \mathrm{Ber}(p)$

PMF: $P(X = 1) = p,\ P(X = 0) = 1 - p$

Expectation: $\mathbb{E}[X] = p$ Note: $\mathbb{E}[X^2] = p$

Variance: $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p)$

Examples:

- Coin flip
- Randomly guessing on a MC test question
- A server in a cluster fails
- Any indicator RV

Binomial Random Variables

A discrete random variable X that is the number of successes in n independent random variables $Y_i \sim \mathrm{Ber}(p)$.
X is a **Binomial random variable** where $X = \sum_{i=1}^{n} Y_i$

Notation: $X \sim \mathrm{Bin}(n, p)$

PMF: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

Expectation: $\mathbb{E}[X] = np$

Variance: $\mathrm{Var}(X) = np(1-p)$

Geometric Random Variables

A discrete random variable X that models the number of independent trials $Y_i \sim \mathrm{Ber}(p)$ before seeing the first success.

X is called a Geometric random variable with parameter p.

Notation: $X \sim \mathrm{Geo}(p)$

PMF: $P(X = k) = (1-p)^{k-1}p$

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

Variance: $\mathrm{Var}(X) = \frac{1-p}{p^2}$

Examples:

- # of coin flips until first head
- # of random guesses on MC questions until you get one right
- # of random guesses at a password until you hit it

Uniform Distribution (Discrete)

A discrete random variable X **equally likely** to take any (integer) value between integers a and b (inclusive), is **uniform.**

Notation: $X \sim \mathrm{Unif}[a, b]$

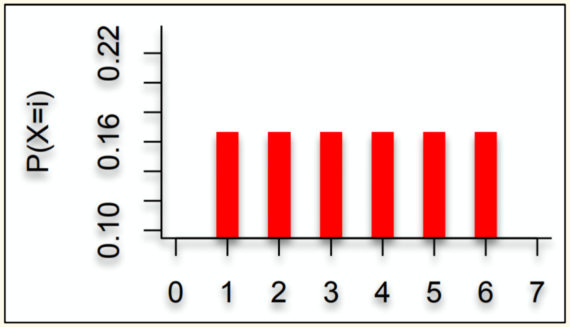
PMF: $\mathrm{P}(X = i) = \dfrac{1}{b - a + 1}$

Expectation: $\mathbb{E}[X] = \dfrac{a+b}{2}$

Variance: $\mathrm{Var}(X) = \dfrac{(b-a)(b-a+2)}{12}$

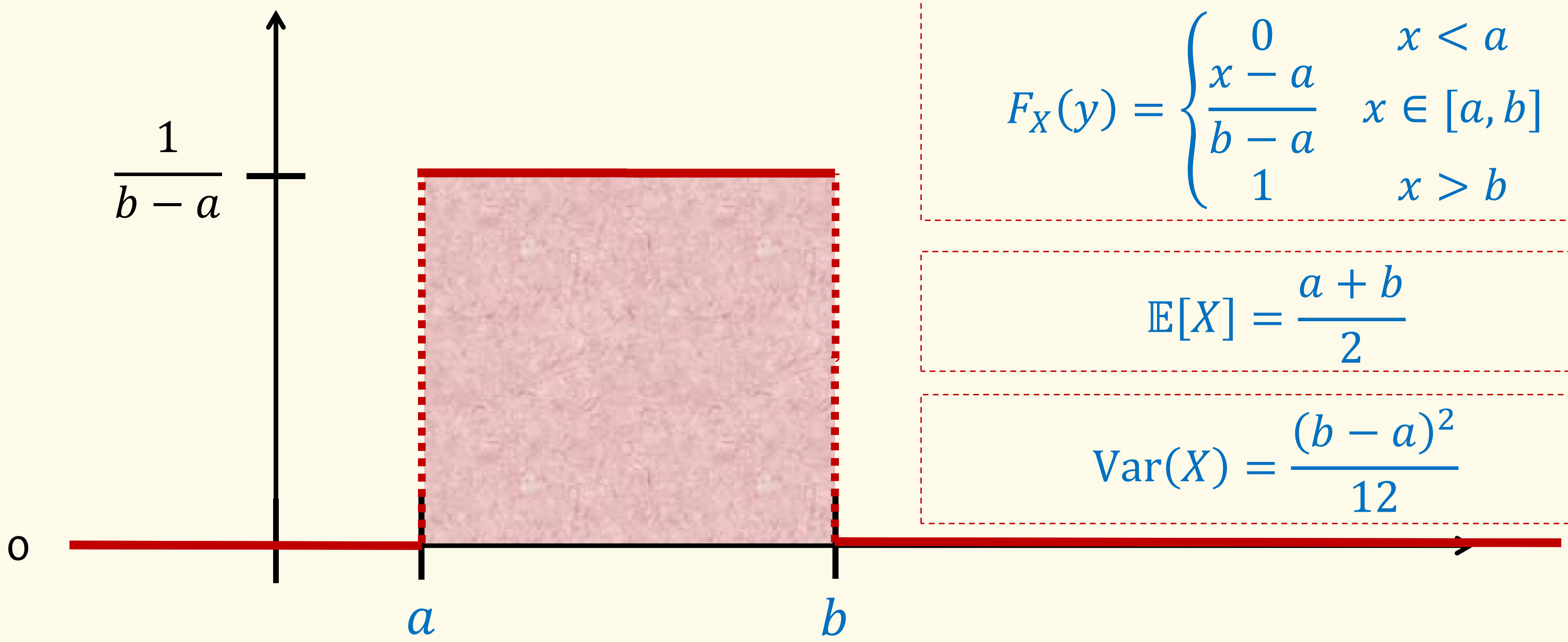
Example: value shown on one roll of a fair die is $\mathrm{Unif}[1,6]$:

- $P(X = i) = 1/6$
- $\mathbb{E}[X] = 7/2$
- $\mathrm{Var}(X) = 35/12$



Uniform Distribution (Continuous)

$X \sim \mathrm{Unif}(a, b)$



$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{else} \end{cases}$$

$$F_X(y) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x > b \end{cases}$$

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

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

Poisson Distribution

Siméon Denis Poisson
1781-1840

- X is a **Poisson r.v. with parameter** λ (denoted $X \sim \text{Poi}(\lambda)$) with this distribution (PMF): For all non-negative integers $k = 0, 1, 2, \ldots$

$$P(Z = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

- $\mathbb{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$

Limit as $n \to \infty$ of $\text{Bin}(n, p)$ for $p = \lambda/n$

Distribution of the # of events that happen, independently, at an *average* rate of λ per unit time: car arrivals, customers, radioactive decay

Theorem. Let $X_1 \sim \text{Poi}(\lambda_1), \cdots, X_n \sim \text{Poi}(\lambda_n)$ be independent. Set $Z = \Sigma_i X_i$. Then $Z \sim \text{Poi}(\lambda)$ for $\lambda = \Sigma_i \lambda_i$.

$P(X > t) = e^{-t\lambda}$

Exponential Distribution

An **exponential random variable** X with parameter $\lambda \geq 0$ ($X \sim \text{Exp}(\lambda)$) follows the exponential density $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$

CDF: For $y \geq 0$,
$F_X(y) = 1 - e^{-\lambda y}$

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \text{Var}(X) = \frac{1}{\lambda^2}$$

Distribution of waiting time until next event if rate per unit time is λ

Theorem. $X \sim \text{Exp}(\lambda)$ is **memoryless:** i.e. for all $s, t > 0$,

$$P(X > s + t \mid X > s) = P(X > t).$$

The Normal Distribution

Carl Friedrich Gauss

A **Gaussian (or normal) random variable** $X \sim \mathcal{N}(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$ and $\sigma \geq 0$ has density

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

Fact. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}[X] = \mu$, and $\mathrm{Var}(X) = \sigma^2$

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

Cor: $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$

Fact: Sum of independent normals is normal

Independent and Identically Distributed (i.i.d.) RVs

Let $X_1, \ldots, X_n$ random variables, each chosen **independently** with the same (**identical**) **distribution** having expectation μ and variance σ^2

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = n\mu$$

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\sigma^2$$

Empirical observation: $X_1 + \cdots + X_n$ looks like a normal RV as n grows.

Central Limit Theorem

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

Theorem. (Central Limit Theorem) The CDF of Y_n converges to the CDF of the standard normal $\mathcal{N}(0,1)$, i.e.,

$$\lim_{n\to\infty} P(Y_n \leq y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-x^2/2}\mathrm{d}x$$

Also stated as:

- $\lim_{n\to\infty} Y_n \to \mathcal{N}(0,1)$
- $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ for $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathrm{Var}(X_i)$

Normal approximation

- Let $\overline{X}$ be the average of i.i.d. random variables $X_1, \dots, X_n$ with mean μ and variance σ^2.
- CLT says that $\frac{\sqrt{n} \cdot (\overline{X} - \mu)}{\sigma}$ approaches $\mathcal{N}(0,1)$ **standard unit normal**
- Approximate using **CDF of** $\mathcal{N}(0,1)$

  $\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} \mathrm{d}x$ for $Z \sim \mathcal{N}(0,1)$

  Note: $\Phi(z)$ has no closed form – generally given via tables

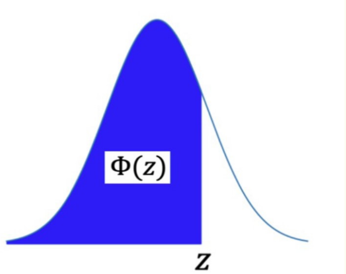
  Within 1 standard deviation 68% within 2 standard deviations 95%, 3 s.d.'s 99%

Review

Table of $\Phi(z)$ CDF of Standard Normal Distribution

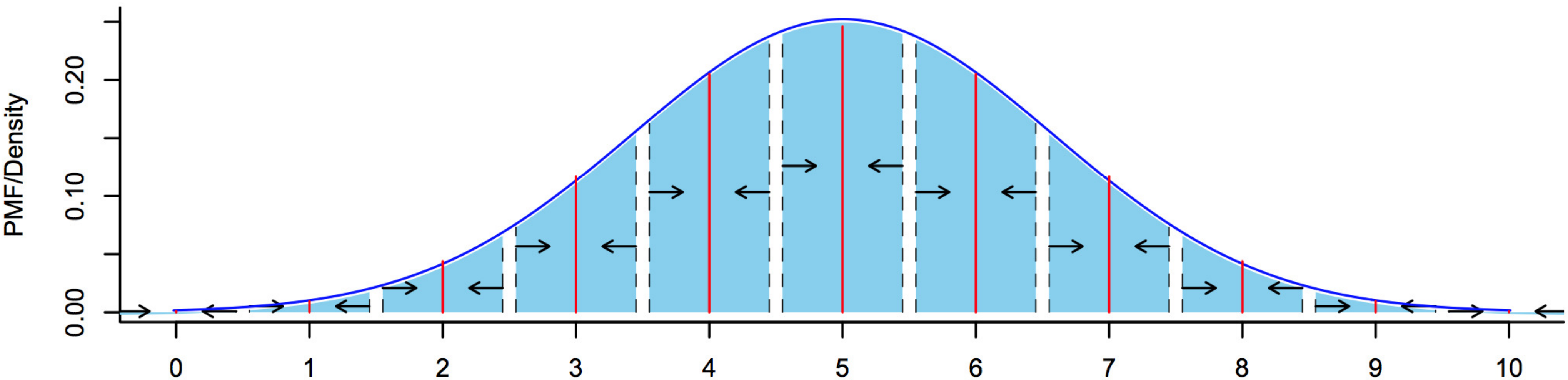
Φ Table: $\mathbb{P}(Z \leq z)$ when $Z \sim \mathcal{N}(0,1)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999



Continuity Correction

Round to next integer!



To estimate probability that discrete RV lands in set S of integers include all surrounding values that round to S.

For interval $\{a, \ldots, b\}$, compute probability for interval $\left[a - \frac{1}{2}, b + \frac{1}{2}\right]$.

Parameter Estimation – Workflow

Distribution $P(x;\theta)$ → Independent samples $X_1, \ldots, X_n$ from $P(x;\theta)$ → **Estimation Algorithm** → $\hat{\theta}$

Parameter estimate

θ = unknown parameter

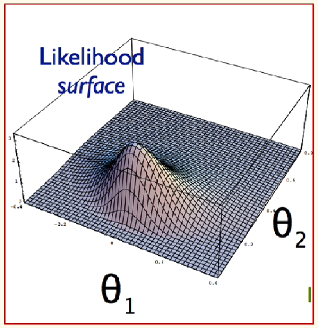
Example: coin flip distribution with unknown θ = probability of heads

Observation: $HTTHHHTHTHTTTTHTHTTTTTHT$

Goal: Estimate θ

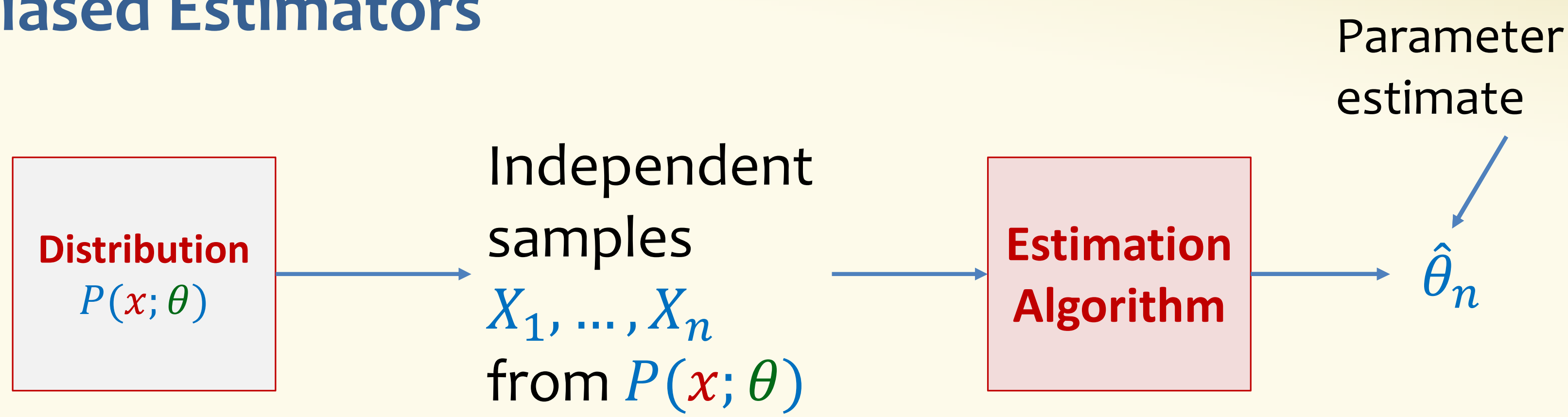
Maximum Likelihood Estimation (MLE)

1. **Input** Given n i.i.d. samples $x_1, \dots, x_n$ from parametric model with parameter (or vector of parameters) θ.
2. **Likelihood** Define your likelihood $\mathcal{L}(x_1, \dots, x_n | \theta)$.
   - For discrete $\mathcal{L}(x_1, \dots, x_n | \theta) = \prod_{i=1}^{n} P(x_i ; \theta)$
   - For continuous $\mathcal{L}(x_1, \dots, x_n | \theta) = \prod_{i=1}^{n} f(x_i ; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, \dots, x_n | \theta)$
4. **Differentiate** Compute $\frac{\partial}{\partial \theta_j} \ln \mathcal{L}(x_1, \dots, x_n | \theta)$ for each parameter in θ (also check discontinuities)
5. **Solve for** $\hat{\theta}$ by setting derivatives to 0 and solving for max.



Generally, you need to do a second derivative test to verify it is a maximum, but we won't ask you to do that in CSE 312.

Unbiased Estimators



θ = unknown parameter

An estimation algorithm like MLE defines $\hat{\theta}_n$ as a function of the random variables $X_1, \ldots, X_n$.
$\hat{\theta}_n(X_1, \ldots, X_n)$ is a r.v. whose expectation we can evaluate using LOTUS.

Definition. An estimator is **unbiased** if $\mathbb{E}\left[\hat{\theta}_n\right] = \theta$ for all $n \geq 1$.

Estimators for the Normal Distribution

Normal outcomes $X_1, \ldots, X_n$ i.i.d. according to $\mathcal{N}(\mu, \sigma^2)$ **Assume:** $\sigma^2 > 0$

$$\widehat{\Theta}_\mu = \frac{\sum_i^n X_i}{n}$$

Sample mean (MLE) – Unbiased!

$$\widehat{\Theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_\mu\right)^2$$

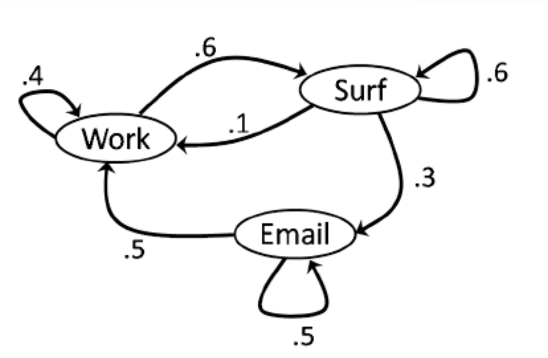
Population variance (MLE) – Biased!

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_\mu\right)^2$$

Sample variance – Unbiased!

But population variance (like every MLE) is **consistent** in that $\lim_{n\to\infty} \mathbb{E}\left[\hat{\theta}_{\sigma^2}\right] = \sigma^2$.

Markov chain



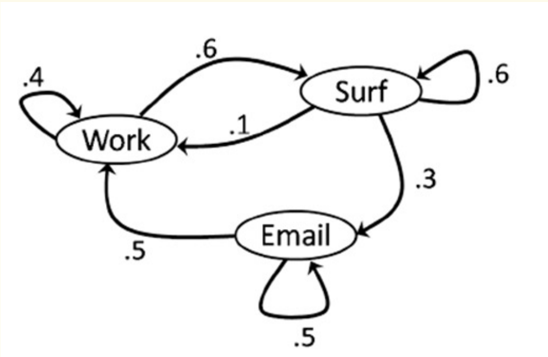
This kind of random process is called a **Markov Chain**

At *each* time step t

- Can be in one of a set of **states**
  - Work, Surf, Email
- If I am in some state s at time t
  - the **labels of out-edges** of s **give the probabilities** of moving to each of the states at time $t + 1$ (as well as staying the same)
    - so **labels on out-edges sum to** 1

e.g. If in Email, there is a 50-50 chance it will be in each of Work or Email at the next time step, but it will never be in state Surf in the next step.

Transition Probability Matrix and distribution of $X^{(t)}$



$$[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}] \overset{\boldsymbol{M}}{\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}} = [q_W^{(t+1)}, q_S^{(t+1)}, q_E^{(t+1)}]$$

Vector-matrix multiplication

$\boldsymbol{M}$ is the Transition Probability Matrix

Probability vector for state variable $X^{(t)}$ at time t: $\boldsymbol{q}^{(t)} = [q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$

For all $t \geq 0$, $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\boldsymbol{M}$

Equivalently, $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\boldsymbol{M}^t$ for all $t \geq 0$

Stationary Distribution of a Markov Chain

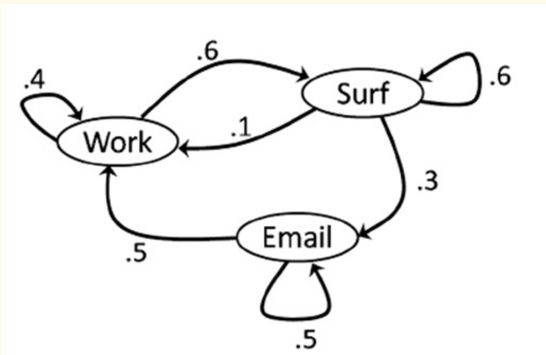
Definition. The **stationary distribution of a Markov Chain** with n states is the n-dimensional row vector $\boldsymbol{\pi}$ such that

$$\boldsymbol{\pi M} = \boldsymbol{\pi}$$

and $\boldsymbol{\pi}$ is a probability distribution

Intuition: Distribution over states at next step is the same as the distribution over states at the current step

Computing a Stationary Distribution



$$[\pi_W, \pi_S, \pi_E] \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix} = [\pi_W, \pi_S, \pi_E]$$

Stationary Distribution satisfies

- $\boldsymbol{\pi} = \boldsymbol{\pi M}$, where $\boldsymbol{\pi} = (\pi_W, \pi_S, \pi_E)$
- $\pi_W + \pi_S + \pi_E = 1$

$\Rightarrow \pi_W = \frac{10}{34},\ \pi_S = \frac{15}{34},\ \pi_E = \frac{9}{34}$

Solve system of equations:

$$\begin{cases} 0.4 \cdot \pi_W + 0.1 \cdot \pi_S + 0.5 \cdot \pi_E = \pi_W \\ 0.6 \cdot \pi_W + 0.6 \cdot \pi_S = \pi_S \\ 0.3 \cdot \pi_S + 0.5 \cdot \pi_E = \pi_E \end{cases}$$

$$\pi_W + \pi_S + \pi_E = 1$$

Fundamental Theorem of Markov Chains

Intuition: $\boldsymbol{q}^{(t)}$ is the distribution of being at each state at time t computed by $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\boldsymbol{M}^t$. Often as t gets large $\boldsymbol{q}^{(t)} \approx \boldsymbol{q}^{(t+1)}$.

Fundamental Theorem of Markov Chains : For a Markov Chain that is aperiodic* and irreducible*, with transition probabilities $\boldsymbol{M}$ and for any starting distribution $\boldsymbol{q}^{(0)}$ over the states

$$\lim_{t\to\infty} \boldsymbol{q}^{(0)}\boldsymbol{M}^t = \boldsymbol{\pi}$$

where $\boldsymbol{\pi}$ is the stationary distribution of $\boldsymbol{M}$ (i.e., $\boldsymbol{\pi}\boldsymbol{M} = \boldsymbol{\pi}$)

**These concepts are way beyond us but they turn out to cover a very large class of Markov chains of practical importance.*