# Probability Review

ECE/CS 498 DS U/G

Probability Review

Lecture 2

Professor Ravi K. Iyer

University of Illinois

# Announcements

- HW0 has been released today and is due on Jan 23 by 23:59 hrs. Please upload your solved HW on Compass2G. You can either upload a scanned copy of your answers or type them.
- You should all have access to Compass2G now. Please contact us in case you are unable to.
- Groups for Mini Projects:
  - Submit the NetIDs of the members of the group via the Google Form shared on Piazza [Link: https://goo.gl/forms/Yo12IZDiEfi30Fgy1]
  - Deadline is Jan 22, Tuesday 23:59 hrs.
  - Feel free to use Piazza for finding group members
  - Students unable to form groups by the deadline will be assigned randomly by the TAs
- Please follow the 'Private Post and Email Etiquette' on Piazza for sending an in-person email

# Probability Basic Concepts

- ***Random experiment*** is an experiment the outcome of which is not certain
- ***Sample Space (S)*** is the totality of the possible outcomes of a random experiment
- ***Discrete (countable) sample space*** is a sample space which is either
  - *finite*; the set of all possible outcomes of the experiment is finite
  - *countably infinite;* the set of all outcomes can be put into a one-to-one correspondence with the natural numbers
- ***Continuous sample space*** is a sample space for which all elements constitute a continuum, such as all the points on a line, all the points in a plane
- An ***event*** is a collection of certain sample points, i.e., a subset of the sample space
  - ***Universal event*** is the entire sample space S
  - *The null set* $\varnothing$ is a **null or impossible event**

# Algebra of Events

- **Algebra of Events**
  - The ***intersection*** of $E_1$ and $E_2$ is given by:
    - $E_1 \cap E_2 = \{s \in S \mid s \text{ is an element of both } E_1 \text{ and } E_2\}$
  - The ***union*** $E_1$ and $E_2$ is given by:
    - $E_1 \cup E_2 = = \{s \in S \mid \text{either } s \in E_1 \text{ or } s \in E_2 \text{ or both}\}$
  - In general: $|E_1 \cup E_2| \leq |E_1| + |E_2|$
    - where $|A|$ = the number of elements in the set **(Cardinality)**
  - Definition of *union* and *intersection* extend to any finite number of sets:

$$\bigcup_{i=1}^{n} E_i = E_1 \cup E_2 \cup E_3 \cup \ldots \cup E_n$$

$$\bigcap_{i=1}^{n} E_i = E_1 \cap E_2 \cap E_3 \cap \ldots \cap E_n$$

# Mutual Exclusive and Collectively Exhaustive

- ***Mutually exclusive or disjoint events*** are two events for which

$$A \cap B = \varnothing$$

- A list of events $A_1, A_2, \ldots, A_n$ is said to be
  - composed of ***mutually exclusive events*** iff:

$$A_i \cap A_j = \begin{cases} A_i, & \text{if } i = j \\ \emptyset, & \text{otherwise} \end{cases}$$

  - *collectively* ***exhaustive*** iff: $A_1 \cup A_2 \cup \cdots \cup A_n = S$

# Probability Axioms

- **Probability Axioms**
  - Let *S* be a sample space of a random experiment and *P(A)* be the probability of the event *A*
  - The probability function *P(.)* must satisfy the three following axioms:
  - **(A1)** For any event A, $P(A) \geq 0$

    *(probabilities are nonnegative real numbers)*
  - **(A2)** $P(S) = 1$

    *(probability of a certain event, an event that must happen is equal 1)*
  - **(A3)** $P(A \cup B) = P(A) + P(B)$, whenever A and B are mutually exclusive events, i.e., $A \cap B = \varnothing$

    *(probability function must be additive)*
  - **(A3')** For any countable sequence of events $A_1, A_2, \ldots, A_n \ldots$, that are mutually exclusive (that is $A_j \cap A_k = \varnothing$ whenever $j \neq k$)

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

# Probability Axioms

- **(Ra)** For any event A, $P(\bar{A}) = 1 - P(A)$
- **(Rb)** If $\varnothing$ is the impossible event, then $P(\varnothing) = 0$
- **(Rc)** If A and B are any events, not necessarily mutually exclusive, then

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

- **(Rd)**(generalization of Rc) If $A_1, A_2, \ldots, A_n$ are any events, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) + \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k) +$$

$$\ldots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n)$$

where the successive sums are over all possible events, pairs of events, triples of events, and so on. (Can prove this relation by induction (see class web site))

# Discrete/Continuous Random Variables

- Random Variable $X: S \rightarrow \mathbb{R}$
- The ***discrete*** random variables are either a finite or a countable number of possible values.
- Random variables that take on a continuum of possible values are known as ***continuous*** random variables.
- Example: A random variable denoting the time to disengagement of an AV, when the time is assumed to take on any value in some interval $(0, \infty)$ is *continuous*

# Discrete Random Variables

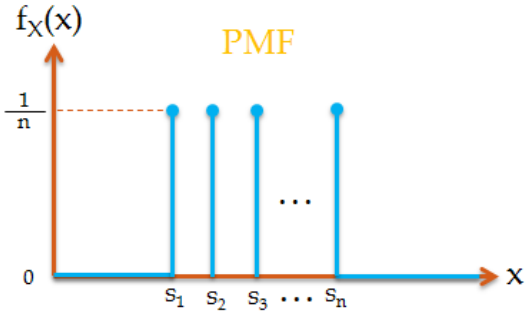
- **Probability Mass Function (PMF):**

  Properties:

$$p(a) = P\{X = a\}$$

$$p(x) = \begin{cases} > 0, & x = x_1, x_2, \ldots \\ 0, & \text{for other values of } x \end{cases}$$

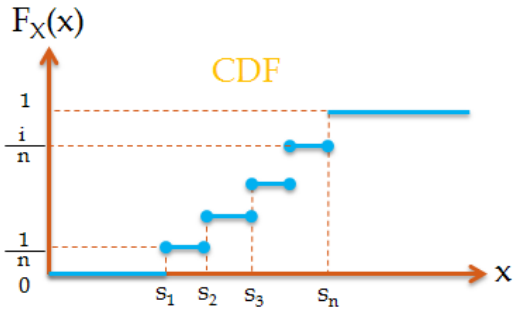
$$\sum_{i=1}^{\infty} p(x_i) = 1$$




- **Cumulative Mass Function (CMF):**

$$F(a) = \sum_{all\ x_i \le a} p(x_i)$$

  - Staircase function



# Discrete Random Variables: Probability Mass Function (pmf)

- A random variable that can take on at most countable number of possible values is said to be *discrete*.
- For a discrete random variable $X$, we define the ***probability mass function*** $p(a)$ of $X$ by:

$$p(a) = P\{X = a\}$$

- $p(a)$ is positive for at most a countable number of values of $a$. i.e., if $X$ must assume one of the values $x_1, x_2, \ldots$ then

$$p(x) \begin{cases} > 0, & x = x_1, x_2, \ldots \\ = 0, & \text{for other values of } x \end{cases}$$

- Since $X$ takes values $x_i$: $\sum_{i=1}^{\infty} p(x_i) = 1$

# An Example of Discrete RV - *Geometric Distribution*

- To find the pmf of a Geometric R.V $Z$, note that the event $[Z = i]$ occurs if and only if we have a sequence of $(i - 1)$ failures followed by one success - a sequence of independent Bernoulli trials with the probability of success equal to $p$ and failure $q$.
- Hence, we have

$$p_Z(i) = q^{i-1}p = p(1-p)^{i-1} \quad \text{for } i = 1,2, \ldots$$

- Using the formula for the sum of a *geometric* series, we have:

$$\sum_{i=1}^{\infty} p_Z(i) = \sum_{i=1}^{\infty} pq^{i-1} = \frac{p}{1-q} = \frac{p}{p} = 1$$

- The corresponding CDF is:

$$F_Z(t) = \sum_{i=1}^{\lfloor t \rfloor} p(1-p)^{i-1} = 1 - (1-p)^{\lfloor t \rfloor} \quad \text{for } t \geq 0$$

# Continuous Random Variables

- **Continuous Random Variables:**
  - **Probability density function (pdf)**:

$$P\{X \in B\} = \int_B f(x)dx$$

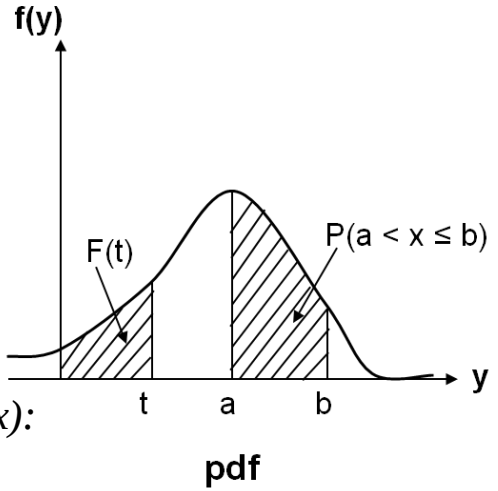
  - Properties:

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)dx$$

  - All probability statements about *X* can be answered by *f(x):*

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

$$P\{X = a\} = \int_a^a f(x)dx = 0$$

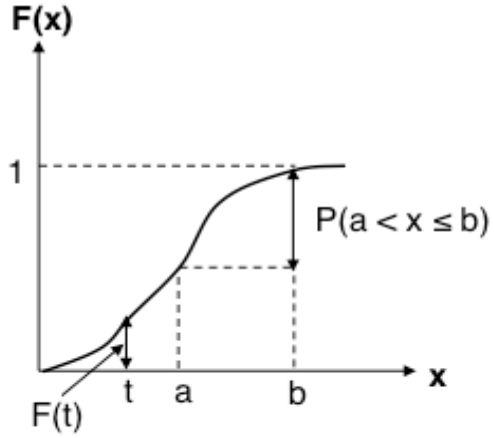


pdf

  - **Cumulative distribution function (CDF)**:

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)dt = 0, \qquad -\infty < x < \infty$$

  - Properties: $\frac{d}{da}F(a) = f(a)$
  - **A continuous function**



CDF

# Continuous Random Variables

- Random variables whose set of possible values is uncountable

- *X* is a continuous random variable if there exists a nonnegative function *f(x)* defined for all real $x \hat{I}$ (-¥,¥), having the property that for any set of *B* real numbers

$$P\{X \in B\} = \int_B f(x)dx$$

- *f(x)* is called the *probability density function (pdf)* of the random variable $X$

- The probability that $X$ will be in $B$ may be obtained by integrating the probability density function over the set $B$. Since $X$ must assume some value, *f(x)* must satisfy

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)dx$$

# Continuous Random Variables Cont'd

- All probability statements about $X$ can be answered in terms of $f(x)$
  *e.g.* letting $B=[a,b]$, we obtain $P\{a \le X \le b\} = \int_a^b f(x)dx$
- If we let $a=b$ in the preceding, then $P\{X = a\} = \int_a^a f(x)dx = 0$
- This equation states that the probability that a continuous random variable will assume any *particular* value is zero
- The relationship between the cumulative distribution $F(\cdot)$ and the probability density $f(\cdot)$

$$F_X(a) = P(X \in (-\infty, a]) = \int_{-\infty}^{a} f_X(t)dt$$

- Differentiating both sides of the preceding yields

$$\frac{d}{da}F(a) = f(a)$$

# Continuous Random Variables Cont'd

- That is, the density function is the derivative of the cumulative distribution function.

- A somewhat more intuitive interpretation of the density function

$$P\left\{a - \frac{\varepsilon}{2} \leq X \leq a + \frac{\varepsilon}{2}\right\} = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f(x)dx \approx \varepsilon f(a)$$

when $\varepsilon$ is small

- The probability that $X$ will be contained in an interval of length $\varepsilon$ around the point $a$ is approximately $\varepsilon f(a)$

# Uniform Distribution

- A continuous random variable *X* is said to have a uniform distribution over the interval (*a,b)* if its density is given by:

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

- And the distribution function is given by:

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

# Binomial Theorem

- **Combinatorial Problems**
  - Permutations with replacement:
    - Ordered samples of size k, with replacement *P(n, k)*
  - Permutations without replacement
    - Ordered Samples of size k, without replacement

$$n(n-1)\dots(n-k+1) = \frac{n!}{(n-k)!} \qquad k = 1,2,\dots,n$$

  - Combinations
    - Unordered sample of size k, without replacement

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

- **Binomial Theorem**

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

# Bernoulli and Binomial Distributions

- *X* is said to be a *Bernoulli* random variable if its probability mass function is given by the equation below for some $p \in [0,1]$, where is the probability that the trial is a success

$$p(0) = P\{X = 0\} = 1 - p$$

$$p(1) = P\{X = 1\} = p$$

- In *n* independent trials, each of which results in a "success" with *p* and in a "failure" with probability *1-p,*
- If *X* represents the number of successes that occur in the *n* trials, *X* is said to be a *binomial* random variable with parameters (*n,p*) and its PMF is given by:

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}, \qquad \text{where } \binom{n}{i} = \frac{n!}{i!\,(n-i)!}$$

# Binomial Random Variable Example 1

- In Alzheimer's disease (AD), the ability to retrieve memories gets compromised due to deterioration in brain health. The process of retrieving memories can be thought of as being probabilistic, and the probability reduces in AD.
- Researchers found that in AD, the probability of successfully retrieving a memory of a previously viewed picture is 0.1. Assume that the retrieval of a memory is independent of retrieval of any other memory.
- What is the probability that out of three pictures stored in memory, at most one will be retrieved?
- If $X$ is the number of pictures that are retrieved, then $X$ is a binomial random variable with parameters (3, 0.1). Hence, the desired probability is given by:

$$P\{X=0\}+P\{X=1\}=\binom{3}{0}(0.1)^0(0.9)^3+\binom{3}{1}(0.1)^1(0.9)^2=0.972$$

# Poisson Distribution

- A random variable $X$, taking on one of the values 0,1,2,…, is said to be a *Poisson* random variable with parameter $\lambda$, if for some $\lambda > 0$,

$$p(i) = P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}, \qquad i = 0,1,\ldots$$

  defines a probability mass function since

- **Relationship to Exponential Distribution**

  If the number of arrivals in an interval $t$ is Poisson distributed with parameter $\lambda$, the inter-arrival times will be Exponentially distributed with parameter $\lambda$.

- **Intuition**: Within certain time, how many events have happened
- **Example**: Number of failures occurred in a certain component up to time $t$

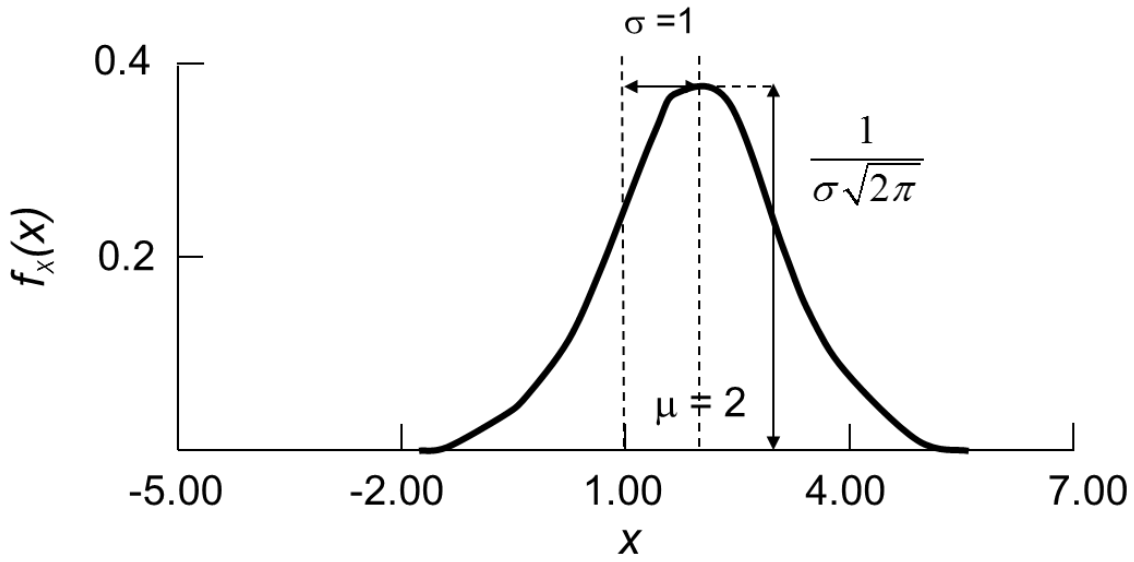
# Normal/Gaussian Distribution

- The normal density is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty$$

where $\mu$ and $\sigma$ are two parameters of the distribution.

- Normal density with parameters $\mu = 2$ and $\sigma = 1$



# Standard Normal Distribution

- The distribution function *F(x)* has no closed form, so between every pair of limits *a* and *b*, probabilities relating to normal distributions are usually obtained numerically and recorded in special tables.

- These tables apply to the **standard normal distribution** $Z \sim N(0,1)$ -- a normal distribution with parameters $\mu = 0$ , $\sigma = 1$ -- and their entries are the values of:

$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$$

# The Exponential Distribution

- The exponential distribution occurs in applications such as reliability theory and queuing theory. Reasons for its use include:
  - Its memoryless (Markov) property
  - Its relation to the (discrete) Poisson distribution

- The following random variables will often be modeled as exponential:
  - Time between two successive job arrivals to a computing center
  - Service time at a server in a queuing network
  - Time to failure (lifetime) of a component
  - Time required to repair a component that has malfunctioned

# Exponential Distribution

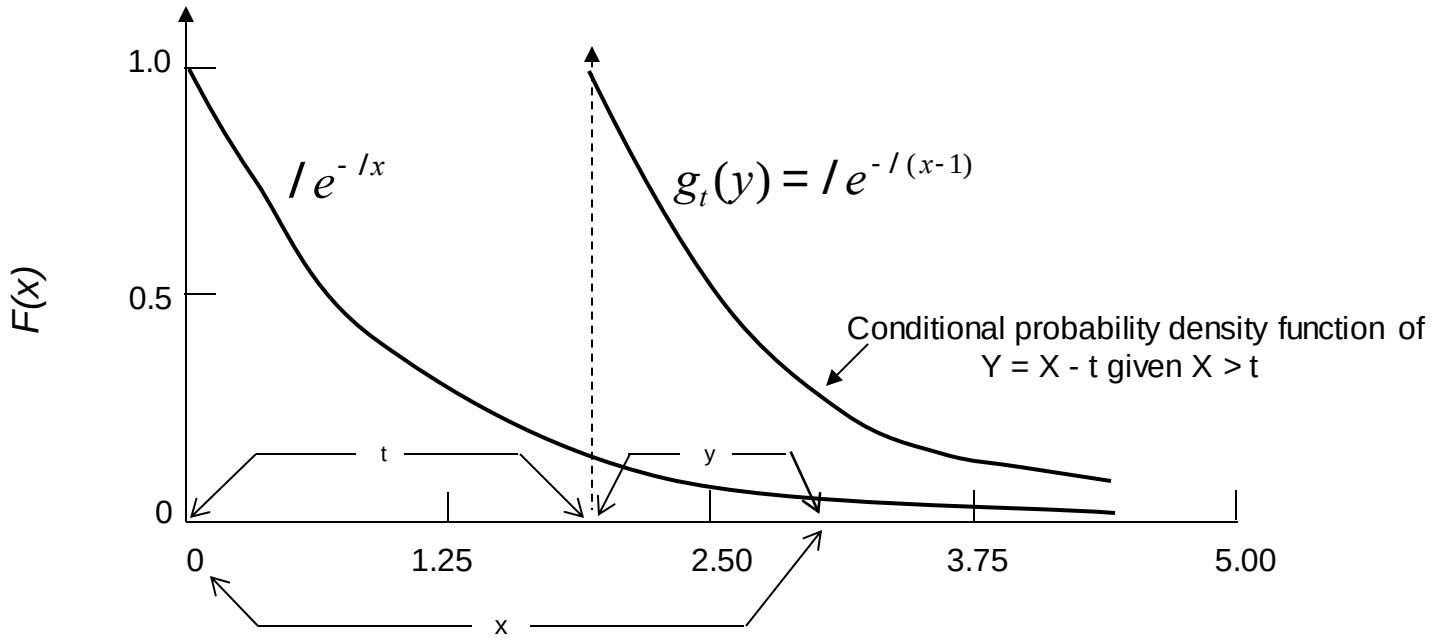
- The CDF of the exponential distribution is given by:

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

- If the CDF of a random variable $X$ is given by the above equation, we use the notation $X \sim EXP(\lambda)$, for brevity. The pdf of $X$ is given by:

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

# Memory-less Property of Exponential



# Gamma Distribution

- Two-parameter family of continuous probability distributions
  - Exponential, Erlang, and chi-squared distributions are special cases of Gamma distribution
- Shape $\alpha$ and rate $\beta$

$$X \sim \Gamma(\alpha, \beta) \equiv Gamma(\alpha, \beta)$$

PDF: $f(x; \alpha, \beta) = \frac{\beta x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$, where, $\Gamma(\alpha)$ is the gamma function

$$\Gamma(z) = \begin{cases} (z-1)!, & \text{positive integer} \\ \int_0^\infty x^{z-1} e^{-x}\, dx, & \text{otherwise} \end{cases}$$

- Intuition: how much time it takes for $\alpha$ events to happen
- Example: the number of requests on web servers

# Beta Distribution

- Two-parameter continuous probability distributions defined on [0,1]
  - A special case of the Dirichlet distribution

- Shape parameters $\alpha$ and $\beta$

$$PDF:\ f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

where $\mathrm{B}(\alpha,\beta)$ is the beta function: $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

($\Gamma(\cdot)$ is the gamma function shown in last slide)

- Intuition: the likelihood of simulated Bernoulli experiments (with a finite sequence of probabilities) that agrees with the observation
- Example: Density of product rating

# Weibull Distribution

- Two-parameter continuous probability distributions
  - $\lambda$ defines scale; $k$ defines shape
- Shape parameters $\alpha$ and $\beta$

$$pdf: f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^{k}} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

  - Shape parameter $k$:
    - $k < 1$: failure rate decreases over time
    - $k = 1$: failure rate is constant over time
    - $k > 1$: failure rate increases over time
- Example: time to failure; reaction time for an AV disengagement

https://en.wikipedia.org/wiki/Weibull_distribution

# General Analysis

- **Summary of important distributions:**

| Distribution | PDF or PMF | Mean | Variance |
|---|---|---|---|
| $Bernoulli(p)$ | $\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$ | $p$ | $p(1-p)$ |
| $Binomial(n,p)$ | $\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$ | $np$ | $npq$ |
| $Geometric(p)$ | $p(1-p)^{k-1}$ for $k = 1, 2, \ldots$ | $\frac{1}{p}$ | $\frac{1-p}{p^2}$ |
| $Poisson(\lambda)$ | $e^{-\lambda}\lambda^x / x!$ for $k = 1, 2, \ldots$ | $\lambda$ | $\lambda$ |
| $Uniform(a,b)$ | $\frac{1}{b-a}$ $\forall x \in (a,b)$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| $Gaussian(\mu, \sigma^2)$ | $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ | $\mu$ | $\sigma^2$ |
| $Exponential(\lambda)$ | $\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |

# Conditional Probability

- **Conditional Probability** *of A given B (P(A|B))* defines the conditional probability of the event A given that the event B occurs and is given by:

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

if P(B) ≠ 0 and ***is undefined otherwise.***

- A rearrangement of the above definition gives the following ***multiplication rule (MR)***

$$P(A \cap B) = \begin{cases} P(B)P(A|B) & if\ P(B) \neq 0 \\ P(A)P(B|A) & if\ P(A) \neq 0 \\ 0 & otherwise \end{cases}$$

# Theorem of Total Probability, Bayes Formula

- **Theorem of Total Probability**
- Any event A can be partitioned into two disjoint subsets:

$$A = (A \cap B) \cup (A \cap \bar{B})$$

- Then:

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

- In general:

$$\boxed{P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

- **Bayes Formula:**

$$\boxed{P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}}$$

# Bayes Formula Example

- A certain security software monitors the system for suspicious activity and raises a flag if the probability of the user's account being compromised is > 0.5 based on the activity.
- A user recently downloaded an executable file onto the system. From past data, the security software knows the following:
  - The probability of the normal user downloading an executable file = 0.2
  - The probability of a hacker downloading an executable file = 0.8
  - The prior probability of an account being compromised (i.e., hacker logs in instead of the true user) = 0.1
- Based on the above, should the security software raise a flag?
- $A$: event that an executable file is downloaded
  $B$: user's account is not compromised
- $P(A|B) = 0.2, P(A|\bar{B}) = 0.8, P(\bar{B}) = 0.1, P(B) = 1 - P(\bar{B}) = 0.9$
- From Baye's formula,

$$P(\bar{B}|A) = \frac{0.8 * 0.1}{0.8 * 0.1 + 0.2 * 0.9} = \frac{0.08}{0.26} = 0.31 < 0.5$$

# Independence of Events

- **Independence of Events:**
- Two events A and B are independent if and only if:

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

- Or events A and B are said to be independent if:

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

# Binary Hypothesis Testing (ML)

- Maximum Likelihood (ML) Decision Rule
  - Declares the hypothesis which maximizes the probability (or likelihood) of the probability
    - Likelihood Ratio Test (LRT)

$$\Lambda(k) = \frac{p_1(k)}{p_0(k)} \qquad \Lambda(k) = \begin{cases} > 1 \text{ declare } \mathrm{H}_1 \text{ is true} \\ \leq 1 \text{ declare } \mathrm{H}_0 \text{ is true} \end{cases}$$

  - More generally

$$\Lambda(k) = \begin{cases} > \tau \text{ declare } \mathrm{H}_1 \text{ is true} \\ \leq \tau \text{ declare } \mathrm{H}_0 \text{ is true} \end{cases}$$

As $\tau$ increases, fewer observations lead to decide $H_1$ to be true, thus $p_{false\ alarm}$ decreases and $p_{miss}$ increases. Therefore, $\tau$ can be applied to select an operating point on the tradeoff between the two error probabilities

# Binary Hypothesis Testing Example

- To assist doctors in the diagnosis of breast cancer, researchers have created a classification model that uses the image of breast tissue from biopsies. The model must decide whether the image is of a normal tissue or cancerous tissue.
- For the tissue biopsy image of a new patient, the model assigned the following probabilities:
  - Probability of image given the tissue is normal = 0.3
  - Probability of image given the tissue is cancerous = 0.5
- Assume that the threshold for the model ($\tau$) = 1. Based on the model, does the patient have cancer?
- $p_0(k) = 0.3$, and $p_1(k) = 0.5$
- $\Lambda(k) = \frac{p_1(k)}{p_0(k)} = \frac{0.5}{0.3} = 1.67 > 1 = \tau$
- Therefore, the patient has cancer.

# Binary Hypothesis Testing (MAP)

- Maximum a Posteriori (MAP) Decision Rule
  - Prior, observation $\xrightarrow{Bayes\ formula}$ Posterior
    - By Baye's formula

observation prior

$$P(H_1|X=k) = \frac{P(H_1,X=k)}{P(X=k)} = \frac{P(H_1,X=k)}{P(H_1,X=k)+P(H_0,X=k)} = \frac{P(X=k|H_1)P(H_1)}{P(H_1,X=k)+P(H_0,X=k)}$$

  - Priors: $\pi_0 = P(H_0), \pi_1 = P(H_1)$
  - The MAP rule declares hypothesis $H_1$ is true if $\pi_1 p_1(k) > \pi_0 p_0(k)$, or $\Lambda(k) > \frac{\pi_0}{\pi_1}$
    - Therefore, the MAP rule is equivalent to the LRT with threshold $\tau = \frac{\pi_0}{\pi_1}$
  - The MAP rule minimizes $p_e = \pi_0 p_{false\ alarm} + \pi_1 p_{miss}$

# Joint Distribution Functions

- **Joint distribution functions:**
  - For any two random variables $X$ and $Y$, the ***joint cumulative probability distribution function*** of $X$ and $Y$ by:

$$F(a,b) = P\{X \leq a, Y \leq b\}, \qquad -\infty < a, b, < \infty$$

  - **Discrete:**
    - The ***joint probability mass function*** of $X$ and $Y$

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

    - Marginal PMFs of X and Y:

$$p_X(x) = \sum_{y:\, p(x,y)>0} p(x,y)$$

$$p_Y(y) = \sum_{x:\, p(x,y)>0} p(x,y)$$

# Joint Distribution Functions

- **Joint distribution functions:**
  - **Continuous:**
    - The ***joint probability density function*** of *X* and *Y*:

$$P\{X \in A, Y \in B\} = \int_B \int_A f(x,y)dx\,dy$$

    - Marginal PDFs of X and Y:

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

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

    - Relation between joint CDF and PDF

$$F(a,b) = P(X \le a, Y \le b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f(x,y)dy\,dx$$

# Conditional and Marginal Distribution Functions

- Joint distribution: $f(x,y) = P[X = x, Y = y]$
- Marginal distributions:

$$f(x) = \sum_{all\ y} f(x,y) \qquad f(y) = \sum_{all\ x} f(x,y)$$

- Conditional distributions:

$$f(y|x) = \frac{f(x,y)}{f(x)} \qquad f(x|y) = \frac{f(x,y)}{f(y)} = \frac{f(y|x)f(x)}{y(y)}$$

- Example

| | | $x$ | | | |
|---|---|---|---|---|---|
| | | -1 | 0 | 1 | |
| $y$ | 5 | 0.15 | 0.2 | 0.1 | 0.45 |
| | 10 | 0.05 | 0.2 | 0.3 | 0.55 |
| | | 0.2 | 0.4 | 0.4 | 1 |

# Independent Random Variables

- **Independent Random Variables:** Two random variables $X$ and $Y$ are said to be independent if:

$$F(x,y) = F_X(x)F_Y(y),\ -\infty < x < \infty, -\infty < y < \infty$$

- If $X$ and $Y$ are continuous:

$$f(x,y) = f_X(x)\,f_Y(y),\ -\infty < x < \infty, -\infty < y < \infty$$

- If $X$ is discrete and $Y$ is continuous:

$$P(X = x, Y \pounds y) = p_X(x) f_Y(y),\ \text{all } x \text{ and } y$$

# Moments

- Recall the definition of moments:

$$E[X^n] = \begin{cases} \sum x^n p(x), & if\ X\ is\ discrete \\ \int_{-\infty}^{\infty} x^n f(x) df, & if\ X\ is\ continuous \end{cases}$$

- **The first and second moments (mean and variance) of a single random variable** convey important information about the distribution of the variable.
- For example the ***variance*** of $X$ measures the expected square of the deviation of $X$ from its expected value, and is defined by:

$$Var(X) = E[X^2] - (E[X])^2$$

- Use of moments is even more important when considering more than one random variable at a time with joint distributions that are much more complex than distributions for individual random variables.

# Covariance

- The ***covariance*** of any two random variables, $X$ and $Y$, denoted by $Cov(X,Y)$, is defined by

$$\begin{aligned} Cov(X,Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - YE[X] - XE[Y] + E[X]E[Y]] \\ &= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

- Covariance generalizes variance, in the sense that ***Var(X) = Cov(X, X).***
- If either X or Y has mean zero, then E[XY] = Cov(X,Y).

- If X and Y are ***independent*** then it follows that ***Cov(X,Y) = 0.***
- **But the converse is not true:**
  - **Cov(X,Y) = 0** means X and Y are **uncorrelated, but it doesn't imply** that X and Y are **independent.**

# Properties of Covariance

- For any random variable X, Y, Z, and constant c, we have:
  1. Cov(X,X) = Var(X),
  2. Cov(X,Y) = Cov(Y,X),
  3. Cov(cX,Y) = cCov(X,Y),
  4. Cov(X,Y+Z) = Cov(X,Y) + Cov(X,Z).

  Whereas the first three properties are immediate, the final one is easily proven as follows:

$$\begin{aligned} Cov(X, Y+Z) &= E[X(Y+Z)] - E[X]E[Y+Z] \\ &= E[XY] - E[X]E[Y] + E[XZ] - E[X]E[Z] \\ &= Cov(X,Y) + Cov(X,Z) \end{aligned}$$

- The last property generalizes to give the following result:

$$Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$$

# Correlation Coefficient

- The correlation between two random variable X and Y is measured using the ***correlation coefficient***:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

($\rho_{X,Y}$ is well defined if Var(X) > 0 and Var(Y) > 0)

- If ***Cov(X,Y) = 0,*** X and Y are called ***uncorrelated,*** which implies that ***E[XY] = E[X]E[Y].***

- If ***Cov(X,Y) > 0***, X and Y are ***positively correlated,*** $Y$ **tends to increase as** $X$ **increases**.

- If ***Cov(X,Y) < 0***, X and Y are ***negatively correlated,*** $Y$ **tends to decrease as** $X$ **increases**.

# Limit Theorems

- **Markov's Inequality:** If $X$ is a random variable that takes only nonnegative values, then for any value $a>0$:

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

- **Chebyshev's Inequality:** If $X$ is a random variable with mean $\mu$ and variance $\sigma^2$ then for any value $k>0$,

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

- **Strong law of large numbers:** Let $X_1, X_2, \ldots$ be a sequence of independent random variables having an identical distribution, and let $E[X_i]=\mu$. Then, almost surely,

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \rightarrow \mu \qquad \text{as } n \rightarrow \infty$$

# Limit Theorems

- **Central Limit Theorem:** Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables, each with mean $\mu$ and variance $\sigma^2$ then the distribution of

$$\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}} \rightarrow N(0,1) \text{ as } n \rightarrow \infty$$

- That is,

$$\mathrm{P}\left(\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}} \leq a\right) \rightarrow \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}dx \quad \text{as} \quad n \rightarrow \infty$$

- Note that like the other results, this theorem holds for any distribution of the $X_i$'s ; herein lies its power.