DEFENCE INSTITUTE OF ADVANCED TECHNOLOGY
DEEMED UNIVERSITY
शास्त्रेण
शस्त्रं प्रकरोति
रक्षाम्

Machine Learning for Cyber Security (CS-602)

L#02-Part1

Mathematics Revisited - Probability

By

Dr Sunita Dhavale

Syllabus

- Data Analytics Foundations: R programming, Python Basics -Expressions and Variables, String Operations, Lists and Tuples, Sets, Dictionaries Conditions and Branching, Loops, Functions, Objects and Classes, Reading/Writing files, Hand ling data with Pandas, Scikit Library, Numpy Library, Matplotlib, scikit programming for data analysis, setting up lab environment, study of standard datasets. Introduction to Machine Learning- Applications of Machine Learning, Supervised, unsupervised classification and regression analysis
- Python libraries suitable for Machine Learning Feature Extraction. Data pre-processing, feature analysis etc., Dimensionality Reduction & Feature Selection Methods, Linear Discriminant Analysis and Principal Component Analysis, tackle data class imbalance problem

Syllabus

- Supervised and regression analysis, Regression, Linear Regression, Non-linear Regression, Model evaluation methods, Classification, K-Nearest Neighbor, Naïve Bayes, Decision Trees, Logistic Regression, Support Vector Machines, Artificial Neural Networks, Model Evaluation. Ensemble Learning, Convolutional Neural Networks, Spectral Embedding, Manifold detection and Anomaly Detection
- Unsupervised classification K-Means Clustering, Hierarchical Clustering, Density-Based Clustering, Recommender Systems-Content-based recommender systems, Collaborative Filtering, machine learning techniques for standard dataset, ML applications, Case studies on Cyber Security problems that can be solved using Machine learning like Malware Analysis, Intrusion Detection, Spam detection, Phishing detection, Financial Fraud detection, Denial of Service Detection.

Text/Reference Books

1. Building Machine Learning Systems with Python – Willi Richert, Luis Pedro Coelho
2. Alessandro Parisi, Hands-On Artificial Intelligence for Cybersecurity: Implement smart AI systems for preventing cyber attacks and detecting threats and network anomalies Publication date :Aug 2, 2019, Packt, ISBN-13, 9781789804027
3. Machine Learning: An Algorithmic Perspective – Stephen Marsland
4. Sunita Vikrant Dhavale, “Advanced Image-based Spam Detection and Filtering Techniques”, IGI Global, 2017
5. Soma Halder , Sinan Ozdemir, Hands-On Machine Learning for Cybersecurity: Safeguard your system by making your machines intelligent using the Python ecosystem, By Publication date : Dec 31, 2018, Packt, ISBN-13 :9781788992282

1. Stuart Russell, Peter Norvig (2009), “Artificial Intelligence – A Modern Approach”, Pearson Elaine Rich & Kevin Knight (1999), “Artificial Intelligence”, TMH, 2nd Edition
2. NP Padhy (2010), “Artificial Intelligence & Intelligent System”, Oxford
3. ZM Zurada (1992), “Introduction to Artificial Neural Systems”, West Publishing Company
4. Research paper for study (if any) – White papers on multimedia from IEEE/ACM/Elsevier/Spinger/ Nvidia sources.

Lab assignments

1	**Python Programming part-1**
2	Python Programming part-2
3	Study and Implement Linear Regression Algorithm for any standard dataset like in cyber security domain
4	Study and Implement KMeans Algorithm for any standard dataset in cyber security domain
5	Study and Implement KNN for any standard dataset in cyber security domain
6	Study and Implement ANN for any standard dataset in cyber security domain
7	Study and Implement PCA for any standard dataset in cyber security domain
8	Case Study: Use of ML along with Fuzzy Logic/GA to solve real world Problem in cyber security domain
9	Mini assignment: Apply ML along with PSO/ACO to solve any real world problem in cyber security domain
10	ML Practice Test – 1 Quiz

Defence Institute of Advanced Technology

School of Computer Engineering & Mathematical Sciences

SEMESTER-I TIME TABLE (AUTUMN 2024)$

PROGRAMMES: (I) CS [M.TECH IN CYBER SECURITY] (II) AI [M.TECH CSE (ARTIFICIAL INTELLIGENCE)] BATCH: 2024-2026

Lecture Day	L1 0900-1000	L2 1000-1100	L3 1100-1200	L4 1200-1300		L4 1400-1500	L4 1500-1600	L4 1600-1700	L4 1700-1800
Monday	CE-602 (AI) CS-602 (CS)	CE-604 (AI) CS-603 (CS)	CE-601 (AI) CS-604 (CS)	CE-601 (AI) LAB CS-603 (CS)	Lunch Break 1300-1400	LAB CE-601 (AI) LAB CS-602 (CS)		AM607	
Tuesday	CE-603 (AI) LAB CS-603 (CS)	CE-602 (AI) CS-602 (CS)	CE-601 (AI) CS-605 (CS)	CE-604 (AI) CS-604 (CS)		PGC 601		AM607	
Wednesday	CS-605 (CS)	CE-603 (AI) CS-602 (CS)	CE-602 (AI) CS-603 (CS)	CE-604 (AI) CS-604 (CS)		CE-605(AI) LAB CS-605 (CS)	LAB CS-605 (CS)	AM607	
Thursday	LAB CE-604 (AI) CS-603 (CS)	LAB CE-604 (AI) CS-605 (CS)	LAB CE-602 (AI) CS-601 (CS)	CE-603 (AI) CS-601 (CS)		PGC 601		AM607	
Friday	LAB CE-603 (AI) LAB CS-601 (CS)		LAB CE-602 (AI) CS-601 (CS)	LAB CS-604 (CS)		CE-605(AI) LAB CS-604 (CS)	CE-605(AI)	LAB CE-605(AI)	

COURSE CODE & COURSE NAME		FACULTY
Programme: CS [M.Tech in Cyber Security] Classroom: Arjun	Programme: AI [M.Tech CSE (Artificial Intelligence)] Classroom: Kaveri	
CS-601 Data Security & Privacy	CE-601 Responsible Artificial Intelligence;	MJN: Dr. Manisha J. Nene
CS-602 ML for Cyber Security	CE-604 Practical Machine Learning;	SVD: Dr. Sunita V. Dhavale
CS-605 Network and Cloud Security	CE-602 Intelligent Algorithms	CRS: Prof. CRS Kumar
CS-604 Advanced System Security	---------	DVV: Dr. Deepti V. Vidyarthi
CS-603 Applied Cryptography	---------	AM: Dr. Arun Mishra
---------	CE-603 Deep Neural Network;	US: Dr. Upasna Singh
---------	CE-605 Mathematics for ML;	Unit-2: Dr Upasna, Unit 4: Dr Sunita, Unit3:MJN, Unit 1: Faculty To be Nominated
AM-607 Mathematics for Engineers	AM-607 Mathematics for Engineers	OO/DS/DP: Dr Odellu O., Dr Dasari S., Dr. Debasis P.
PGC-601 Research Methodology	PGC-601 Research Methodology	Common Subject for All

$ TENTATIVE T.T. SUBJECT TO CHANGE

Program Coordinator,
M.Tech (CS & AI), Batch 2024-26

Director, SoCE&MS

Mathematics Revisited

Why Important?

- Machine learning is interdisciplinary field, intersecting diverse areas of mathematics and computer science
  - makes research in Machine Learning so exciting and challenging!
- Selecting the right algorithm which includes giving considerations to accuracy, training time, model complexity, number of parameters and number of features.
- Choosing parameter settings and validation strategies.
- Identifying under-fitting and over-fitting by understanding the Bias-Variance tradeoff.
- Estimating optimal parameter values.

Notations

$a \in A$	*Set Membership*
|B|	cardinality/number of items in Set B
||v||	norm: length of a vector
$\sum$	summation
$\int$	integral
R	set of real numbers
R^n	real number space of n dimension,
(Here n=2 is 2D space, n=3 is 3D space and n>3 is called hyperspace)	
A	Matrix
Y=f(x)	Function that maps Domain of x to Range of Y
$\frac{dY}{dx}$	derivative of Y with respect to single variable x
$\frac{\partial Y}{\partial x}$	Partial derivative of Y with respect to x

Probability Theory

Random Phenomena

- Deterministic phenomenon: Phenomenon whose outcome can be predicted with a very high degree of confidence
- Example: Age of a person (using date of birth stated in Aadhaar card)
- Stochastic phenomenon: Phenomenon which can have many possible outcomes for same experimental conditions.
- Outcome can be predicted with limited confidence
- Example: Outcome of a coin toss

Characterizing random phenomena

- Sources of error in observed outcomes
  - Lack of knowledge of generating process (model error)
  - Errors in sensors used for observing outcomes (measurement error)
  - Modeled using pdf generally
- Types of random phenomena
  - Discrete:
    - Outcomes are finite Coin toss : {H, T}
    - Throw of a dice : {I, 2, 3, 4, 5, 6}
  - Continuous:
    - Infinite number of outcomes
    - Body temperature measurement in deg F

Probability

Is widely used in mathematics, science, engineering, finance and philosophy to draw conclusions about the likelihood of potential events and the underlying mechanics of complex systems

Probability

- Probability is a measure of how likely it is for an event to happen
- Probability measure is a function that assigns a real value to every outcome of a random phenomena which satisfies following axioms
  - 0 < P(A) < 1 (Probabilities are non-negative and less than 1 for any event A)
  - P(S) = 1
  - For two mutually exclusive events A and B : P(A U B) = P(A) + P(B)
- We measure probability with a number between 0 and 1
- If an event is certain to happen, then the probability of the event is 1
- If an event is certain not to happen, then the probability of the event is 0
- Interpretation of probability as a frequency :
  - Conduct an experiment (coin toss) N times.
  - If NA is number of times outcome A occurs then P(A) = NA/N
  - As n->infinity, p(A)->0.5 for fair coin

Random Experiment

- An experiment is called *random* if the outcome of the experiment is uncertain
- For a random experiment:
  - The set of all possible outcomes is known before the experiment
  - The outcome of the experiment is not known in advance
- *Sample space* Ω of an experiment is the set of all possible outcomes of the experiment
- Example: Consider random experiment of tossing a coin twice. Sample space is:

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$$

Probability of Events

- An *event* is a subset of sample space

Example 1: in tossing a coin twice, $E=\{(H,H)\}$ is the event of having two heads

Example 2: in tossing a coin twice, $E=\{(H,H), (H,T)\}$ is the event of having a head in the first toss

- *Probability* of an event E is a numerical measure of the likelihood that event E will occur, expressed as a number between 0 and 1,

$$0 \leq \mathbb{P}(E) \leq 1$$

  - If all possible outcomes are equally likely: $\mathbb{P}(E) = |E|/|\Omega|$
  - Probability of the sample space is 1: $\mathbb{P}(\Omega) = 1$

R code

- //dice roll
- sample(1:6, size = 1, replace = TRUE)
- //two fair dice roll and compute sum
- d <- sample(1:6, size = 2, replace = TRUE)
- sum(d)
- sum(sample(1:6, size = 2, replace = TRUE))
- //create a function
- function to roll two fair, six-sided dice and return their sum
- fun1 <- function(){ return(sum(sample(1:6, size = 2, replace = TRUE))) }
- fun1()
- **Call The Function Repeatedly/Repeat the experiment**
- replicate(n = 20, expr = fun1())
- in general, get a different result each time you run this. That’s because R is simulating a *random experiment*.
- Increase n and see the effect on histogram

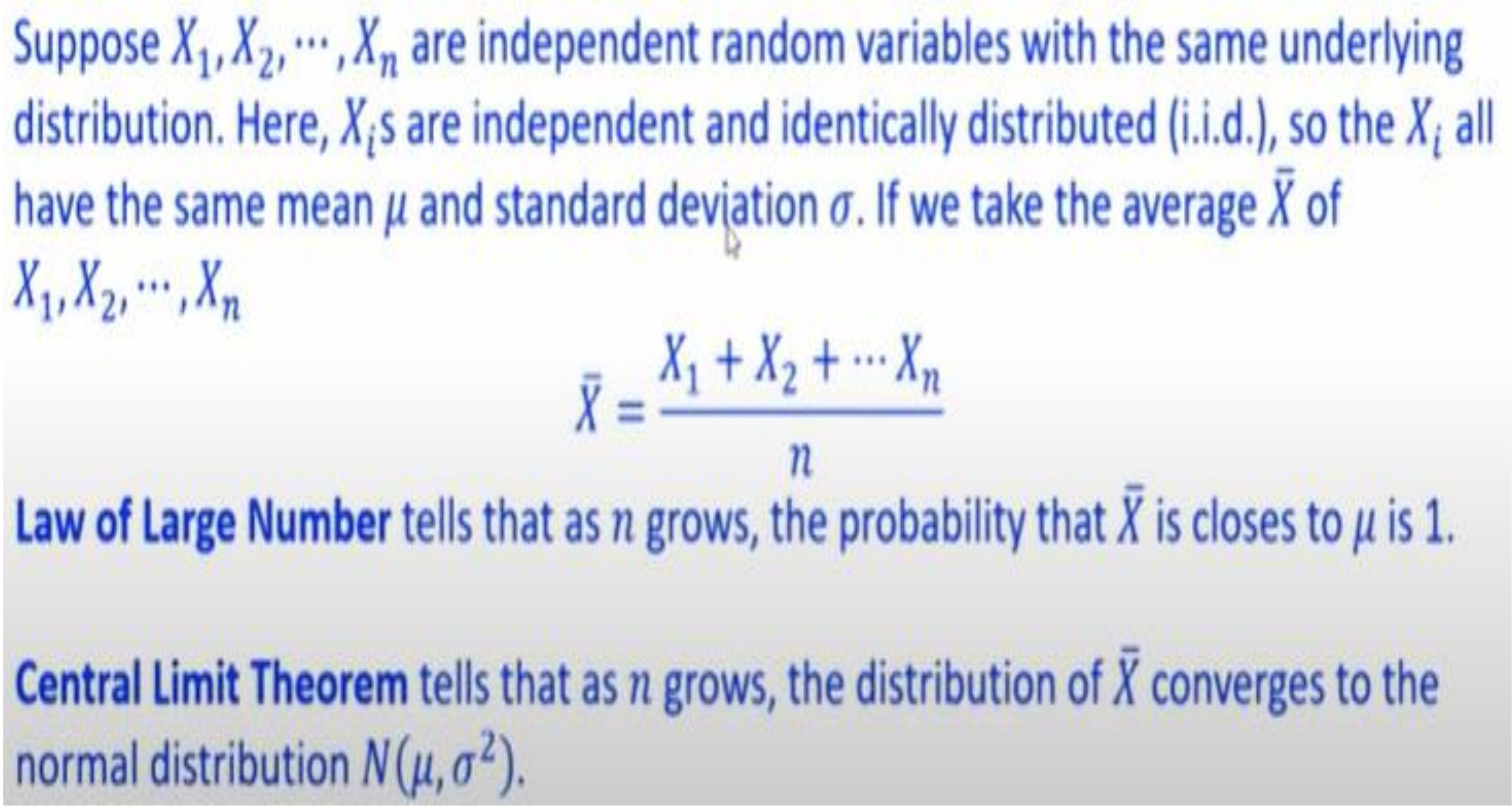
R code

- sims <- replicate(100, fun1())
- table(sims)
- convert this to relative frequencies, we need to divide by the number of times we carried out the experiment
- table(sims)/length(sims)
- plot(table(sims), xlab = 'Sum', ylab = 'Frequency', main = '100 Rolls of 2 Fair Dice')
- plot(table(sims)/length(sims), xlab = 'Sum', ylab = 'Relative Frequency', main = '100 Rolls of 2 Fair Dice')

Imp

- As we increased n, the results got closer to what we know are the true probabilities for rolling dice.
- Increasing the number of simulation replications takes us closer to the idea of “long-run” and hence gives more accurate results, but also takes more time on the computer.
- How many simulation replications are “enough” depends on the problem at hand, but for examples in this class 10,000 will usually suffice.

Imp



Suppose $X_1, X_2, \cdots, X_n$ are independent random variables with the same underlying distribution. Here, X_is are independent and identically distributed (i.i.d.), so the X_i all have the same mean μ and standard deviation σ. If we take the average $\bar{X}$ of $X_1, X_2, \cdots, X_n$

$$\bar{X} = \frac{X_1 + X_2 + \cdots X_n}{n}$$

Law of Large Number tells that as n grows, the probability that $\bar{X}$ is closes to μ is 1.

Central Limit Theorem tells that as n grows, the distribution of $\bar{X}$ converges to the normal distribution $N(\mu, \sigma^2)$.

R code

- runif (n,min,max) is the typical code for generate R.V. from uniform dist.
- x=runif(10,1,10)
- plot(x)
- dotchart(x)
- plot(x,type="l")
- plot(x,type="h")
- plot(x,type="s")
- plot(x,type=“o")
- hist(x)
- barplot(x)
- pie(x)

Dotchart-to draw a Cleveland dot plot

- They are useful for highlighting clusters and gaps, as well as outliers.
- plots of points that each belong to one of several categories. They are an alternative to bar charts or pie charts, and look somewhat like a horizontal bar chart where the bars are replaced by a dots at the values associated with each category.
- # Plot and color by groups cyl
- grps <- as.factor(mtcars$cyl)
- my_cols <- c("blue", "darkgreen", "orange")
- dotchart(mtcars$mpg, labels = row.names(mtcars),
-         groups = grps, gcolor = my_cols,
-         color = my_cols[grps],
-         cex = 0.9, pch = 22, xlab = "mpg")

R code

```
install.packages('plotrix')
library(plotrix)
pie3D(x)
boxplot(x)
x=runif(10,1,10)
plot(x)
x=sample(17:23,100,TRUE)
plot(x)
hist(x)
hist(x, breaks=5, col="red")
table(x)
table(x=sample(17:23,100,TRUE) )
```

R code

- x=seq(2,10,0.1)
- sum(x)
- length(x)
- mean(x)
- sd(x)
- min(x)
- max(x)
- var(x)
- normalize=function(x){y=(x-min(x))/(max(x)-min(x))+return (y)}
- Normalize(x)

R code

- library(help = "datasets")
- data("rivers")
- Example dataset: (Lengths of Major North American Rivers). The U.S. Geological Survey recorded the lengths (in miles) of several rivers in North America. They are stored in the vector rivers in the datasets package (which ships with base R).
- Rr=rivers
- #Apply previous functions to it
- Rr1=normalize(rr)
- Max(rr1)
- Hist(rr)
- Hist(rr,100)

RCode

- Example dataset: (Level of Lake Huron 1875-1972). Brockwell and Davis [10] give the annual measurements of the level (in feet) of Lake Huron from 1875–1972. The data are stored in the time series LakeHuron
- data("LakeHuron")
- rr=LakeHuron
- plot(rr)
- //for categorical data+heights of the bars proportional to the frequencies of observations falling in the respective categories+levels are ordered alphabetically (by default)
- barplot(table(state.region), cex.names = 1.20)
- pie(table(state.region), cex.names = 1.20)

Joint Probability

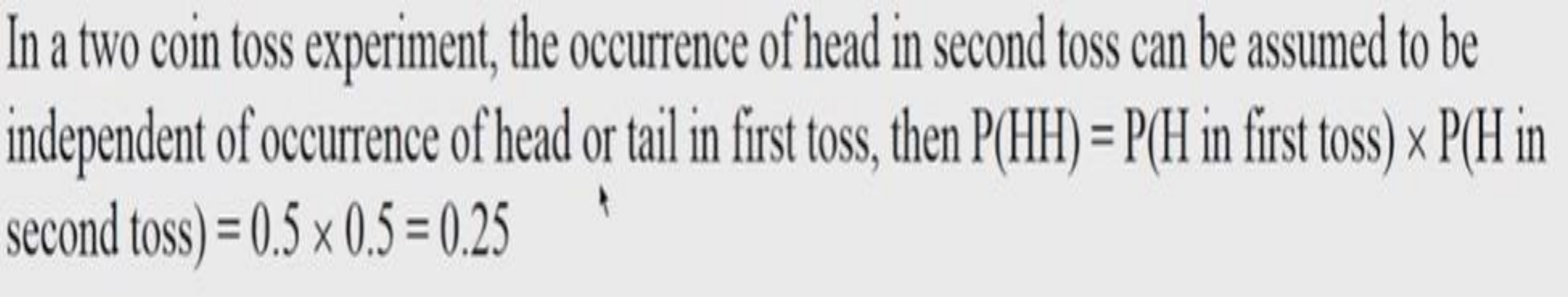
- Probability that two events A and B occur in a single experiment:

$$\mathbb{P}(A \textbf{ and } B) = \mathbb{P}(A \cap B)$$

- Example: drawing a single card at random from a regular deck of cards, probability of getting a red king
  - A: getting a red card
  - B: getting a king
  - $\mathbb{P}(A \cap B) = \frac{2}{52}$

Independent Events

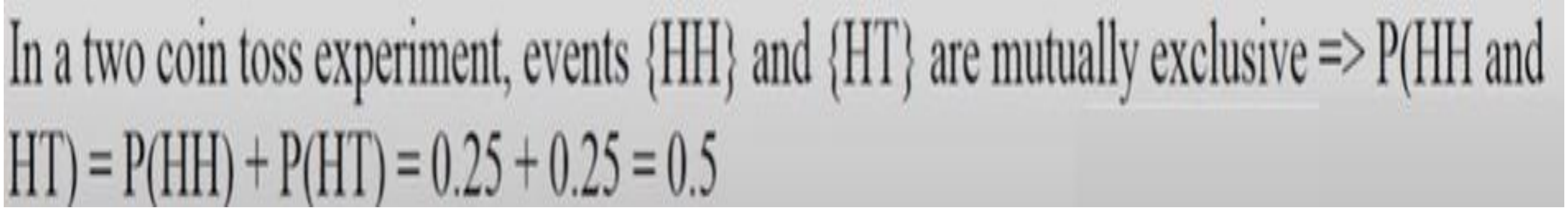
- Two events A and B are independent if the occurrence of one does not affect the occurrence of the other:
$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$
- Example: drawing a single card at random from a regular deck of cards, probability of getting a red king
  - A: getting a red card ⇒ $\mathbb{P}(A) = 26/52$
  - B: getting a king ⇒ $\mathbb{P}(B) = 4/52$
  - $\mathbb{P}(A \cap B) = \frac{2}{52} = \mathbb{P}(A)\mathbb{P}(B) \Rightarrow A$ and B are not independent



In a two coin toss experiment, the occurrence of head in second toss can be assumed to be independent of occurrence of head or tail in first toss, then P(HH) = P(H in first toss) × P(H in second toss) = 0.5 × 0.5 = 0.25

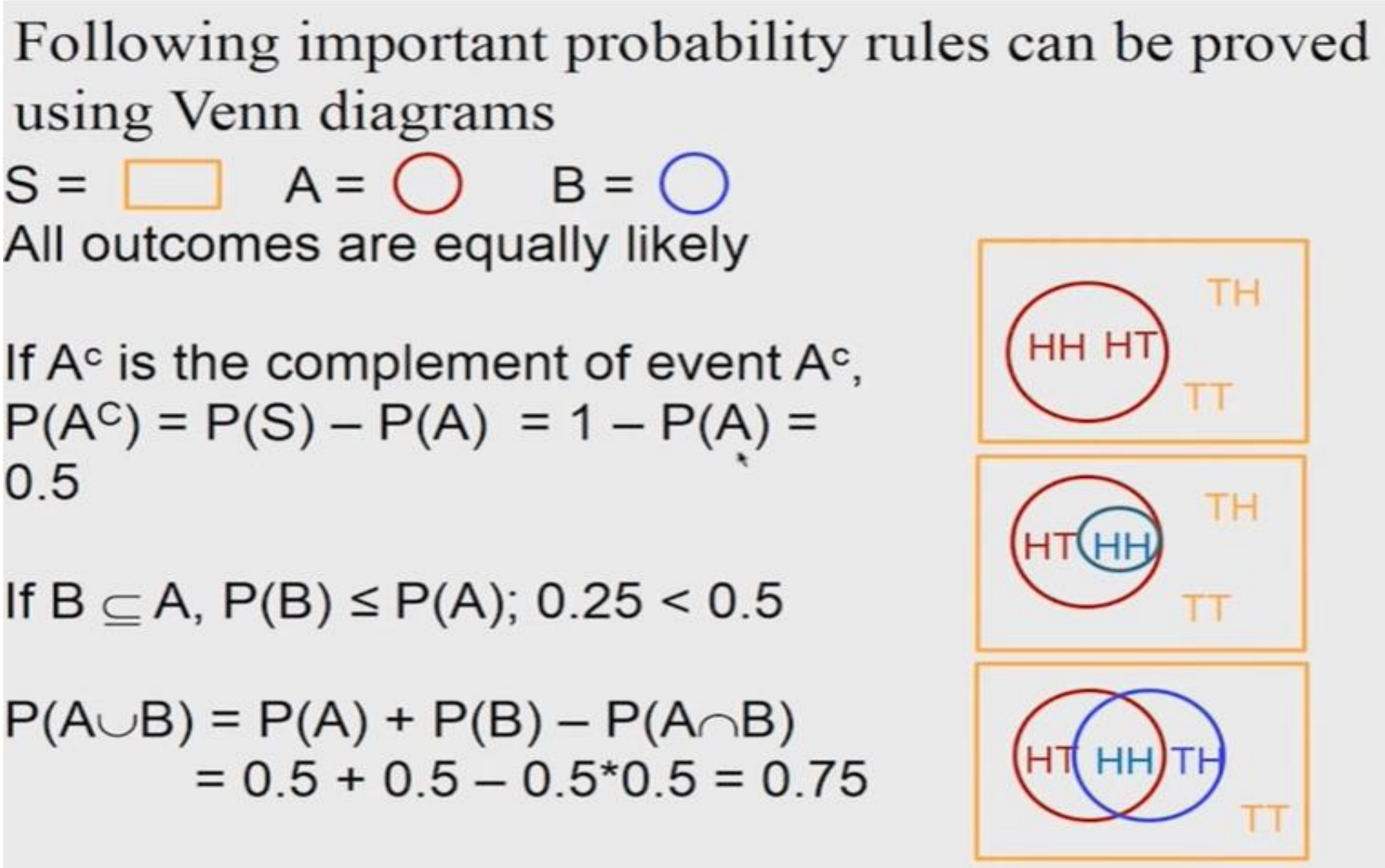
Mutually Exclusive Events

- Events A and B are mutually exclusive if the occurrence of one implies the non-occurrence of the other, i.e., $A \cap B = \phi$:
$$\mathbb{P}(A \cap B) = 0$$
- Example: drawing a single card at random from a regular deck of cards, probability of getting a red club
  - A: getting a red card
  - B: getting a club
  - $\mathbb{P}(A \cap B) = 0$
- Complementary event of event A is event $[not\ A]$, i.e., the event that A does not occur, denoted by $\bar{A}$
  - Events A and $\bar{A}$ are mutually exclusive
  - $\mathbb{P}(\bar{A}) = 1\ - \mathbb{P}(A)$



In a two coin toss experiment, events {HH} and {HT} are mutually exclusive => P(HH and HT) = P(HH) + P(HT) = 0.25 + 0.25 = 0.5

Probability rules



Following important probability rules can be proved using Venn diagrams

S = ▭ A = ○ B = ○

All outcomes are equally likely

If A^c is the complement of event A^c,
$P(A^C) = P(S) – P(A) = 1 – P(A) = 0.5$

If $B \subseteq A$, $P(B) \leq P(A)$; $0.25 < 0.5$

$$P(A \cup B) = P(A) + P(B) – P(A \cap B)$$
$$= 0.5 + 0.5 – 0.5*0.5 = 0.75$$

Union Probability

- Union of events A and B:

$$\mathbb{P}(A \textbf{ or } B) = \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

- If A and B are mutually exclusive:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

- Example: drawing a single card at random from a regular deck of cards, probability of getting a red card or a king
  - A: getting a red card ⇒ $\mathbb{P}(A) = 26/52$
  - B: getting a king ⇒ $\mathbb{P}(B) = 4/52$
  - $\mathbb{P}(A \cap B) = \frac{2}{52}$
  - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = \frac{26}{52} + \frac{4}{52} - \frac{2}{52} = \frac{28}{52}$

Conditional Probability

- Used when 2 events are not independent
- Probability of event A given the occurrence of some event B:

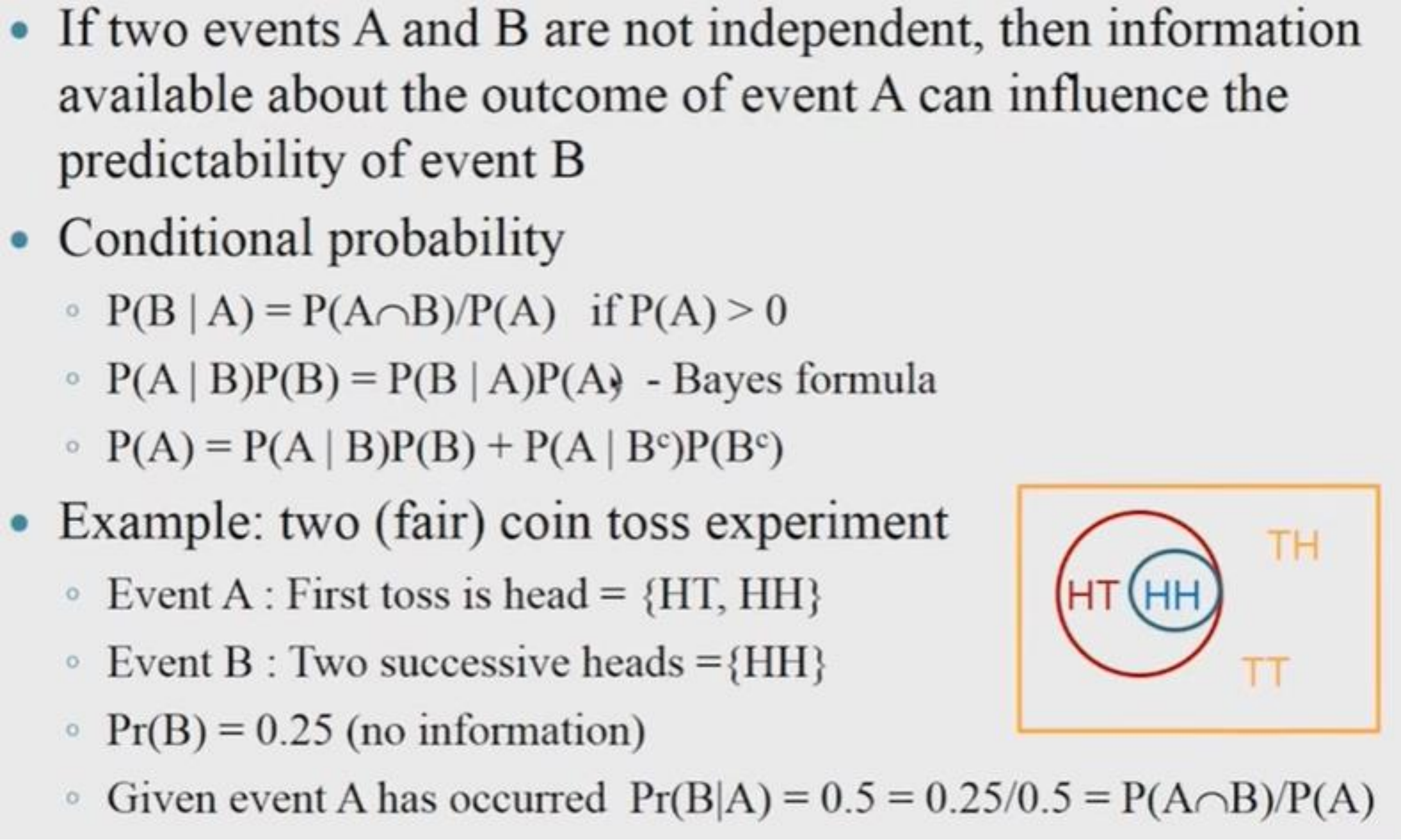
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

- If events A and B are independent:

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

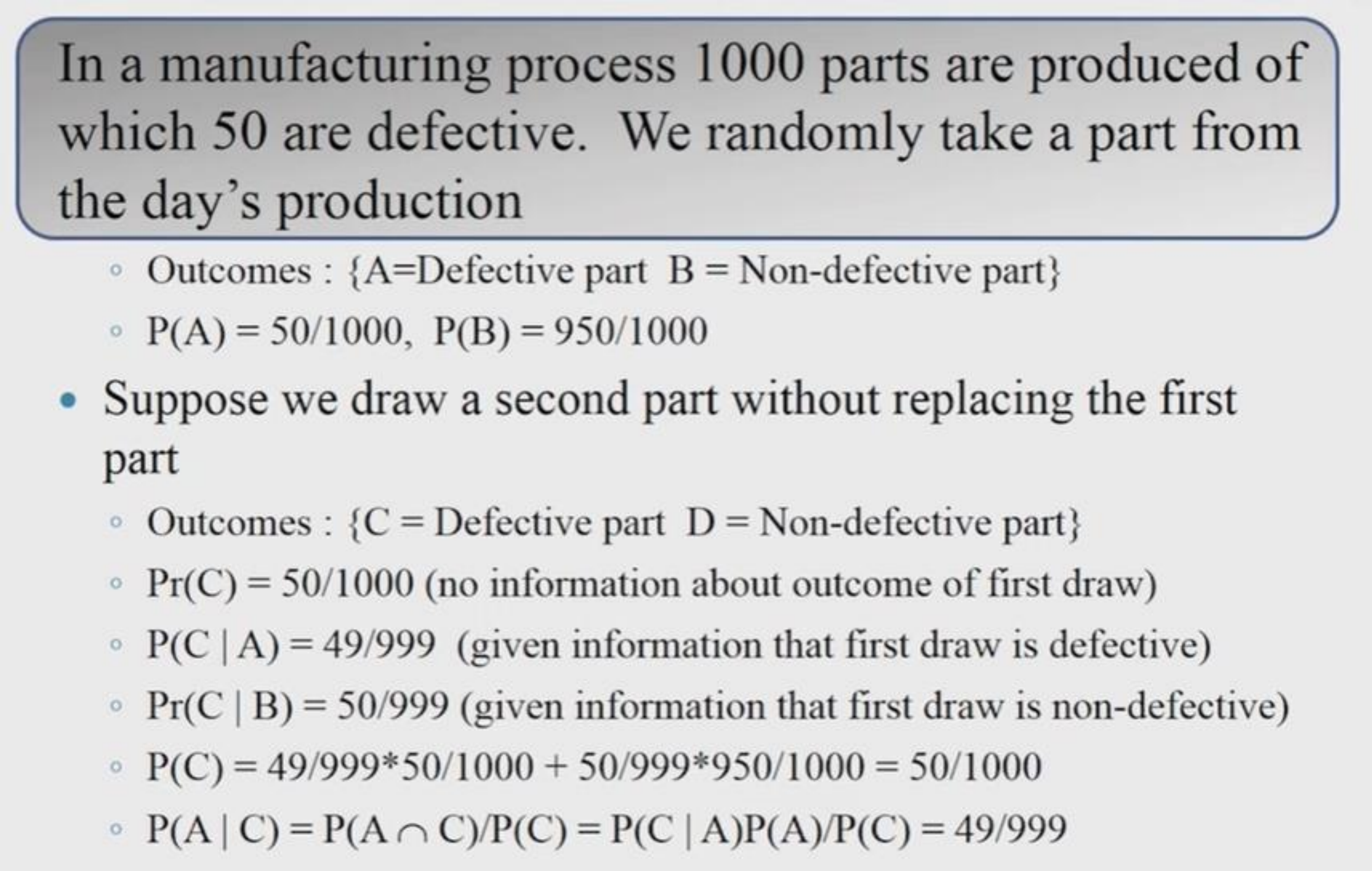
- Example: drawing a single card at random from a regular deck of cards, probability of getting a king given that the card is red
  - A: getting a red card ⇒ $\mathbb{P}(A) = 26/52$
  - B: getting a king ⇒ $\mathbb{P}(B) = 4/52$
  - $\mathbb{P}(A \cap B) = \frac{2}{52}$
  - $\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{2}{26} = \mathbb{P}(B)$

Conditional Probability



- If two events A and B are not independent, then information available about the outcome of event A can influence the predictability of event B
- Conditional probability
  - $P(B \mid A) = P(A \cap B)/P(A)$ if $P(A) > 0$
  - $P(A \mid B)P(B) = P(B \mid A)P(A)$ - Bayes formula
  - $P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$
- Example: two (fair) coin toss experiment
  - Event A : First toss is head = {HT, HH}
  - Event B : Two successive heads ={HH}
  - Pr(B) = 0.25 (no information)
  - Given event A has occurred $\Pr(B|A) = 0.5 = 0.25/0.5 = P(A \cap B)/P(A)$

Example



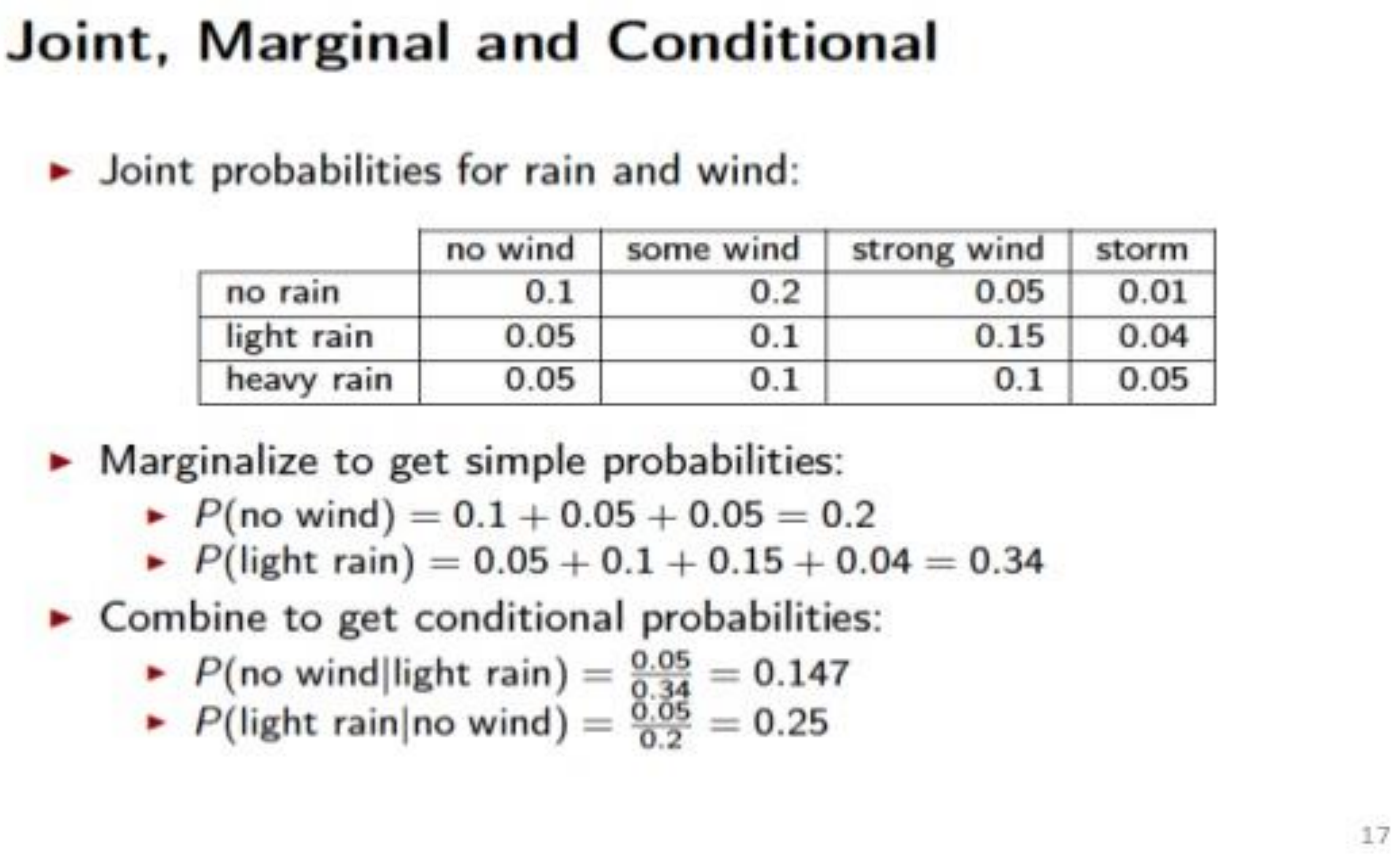
In a manufacturing process 1000 parts are produced of which 50 are defective. We randomly take a part from the day's production

- Outcomes : {A=Defective part B = Non-defective part}
- $P(A) = 50/1000$, $P(B) = 950/1000$

- Suppose we draw a second part without replacing the first part
  - Outcomes : {C = Defective part D = Non-defective part}
  - $Pr(C) = 50/1000$ (no information about outcome of first draw)
  - $P(C \mid A) = 49/999$ (given information that first draw is defective)
  - $Pr(C \mid B) = 50/999$ (given information that first draw is non-defective)
  - $P(C) = 49/999*50/1000 + 50/999*950/1000 = 50/1000$
  - $P(A \mid C) = P(A \cap C)/P(C) = P(C \mid A)P(A)/P(C) = 49/999$

Example

Joint, Marginal & Conditional Probabilities



Joint, Marginal and Conditional

- Joint probabilities for rain and wind:

	no wind	some wind	strong wind	storm
no rain	0.1	0.2	0.05	0.01
light rain	0.05	0.1	0.15	0.04
heavy rain	0.05	0.1	0.1	0.05

- Marginalize to get simple probabilities:
  - $P(\text{no wind}) = 0.1 + 0.05 + 0.05 = 0.2$
  - $P(\text{light rain}) = 0.05 + 0.1 + 0.15 + 0.04 = 0.34$
- Combine to get conditional probabilities:
  - $P(\text{no wind}|\text{light rain}) = \frac{0.05}{0.34} = 0.147$
  - $P(\text{light rain}|\text{no wind}) = \frac{0.05}{0.2} = 0.25$

Random Variable

- A numerical value can be associated with each outcome of an experiment
- A *random variable* X is a function from the sample space Ω to the real line that assigns a real number $X(s)$ to each element s of Ω

$$X\colon \Omega \rightarrow R$$

- Random variable takes on its values with some probability

Random Variable

- Example: Consider random experiment of tossing a coin twice. Sample space is:

  $$\Omega = \{(H,H), (H,T), (T,H), (T,T)\}$$

  Define random variable X as the number of heads in the experiment:

  $$X((T,T)) = 0, \ \ X((H,T))=1,$$
  $$X((T,H)) = 1, X((H,H))=2$$

- Example: Rolling a die.
  Sample space $\Omega = \{1,2,3,4,5,6)$.

  Define random variable X as the number rolled:

  $$X(j) = j, \qquad 1 \leq j \leq 6$$

Random Variable

- Example: roll two fair dice and observe the outcome

  Sample space = $\{(i,j) \mid 1 \leq i \leq 6,\ 1 \leq j \leq 6\}$

  i: integer from the first die

  j: integer from the second die

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Possible outcomes

Random Variable

- Random variable X: sum of the two faces of the dice

$X(i,j) = i+j$

  - $\mathbb{P}(X = 12) = \mathbb{P}(\ (6,6)\ ) = 1/36$
  - $\mathbb{P}(X = 10) = \mathbb{P}(\ (5,5), (4,6), (6,4)\ ) = 3/36$

- Random variable Y: value of the first die
  - $\mathbb{P}(Y = 1) = 1/6$
  - $\mathbb{P}(\ Y = i) = 1/6, \qquad 1 \leq i \leq 6$

Types of Random Variables

- Discrete
  - Random variables whose set of possible values can be written as a finite or infinite sequence
  - Example: number of requests sent to a web server
- Continuous
  - Random variables that take a continuum of possible values
  - Example: time between requests sent to a web server

Probability Mass Function (PMF)

- X: discrete random variable
- $p(x_i)$: probability mass function of X, where

$$p(x_i) = \mathbb{P}(X = x_i)$$

- Properties:

$$0 \leq p(x_i) \leq 1$$

$$\sum_{x_i} p(x_i) = 1$$

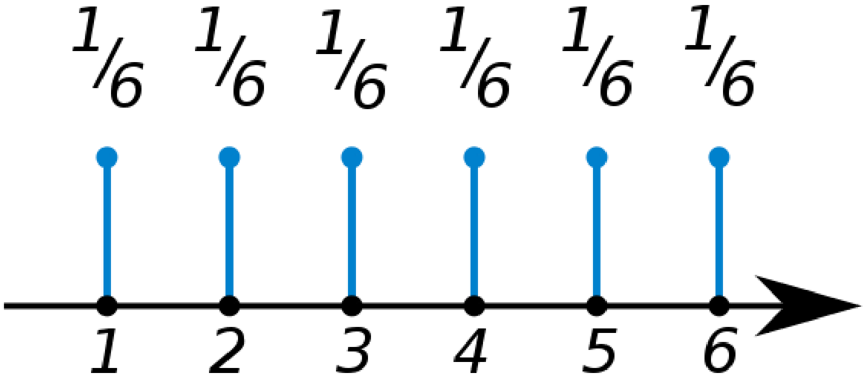
PMF Examples

- Number of heads in tossing three coins

x_i	$p(x_i)$
0	1/8
1	3/8
2	3/8
3	1/8

$$\sum_{x_i} p(x_i) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$$

- Number rolled in rolling a fair die



$$\sum_{x_i} p(x_i) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

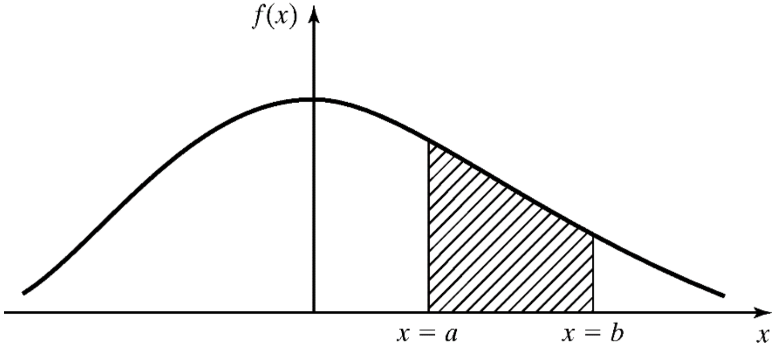
Probability Density Function (PDF)

- X: continuous random variable
- $f(x)$: probability density function of X

$$f(x) = \frac{d}{dx} F(x)$$

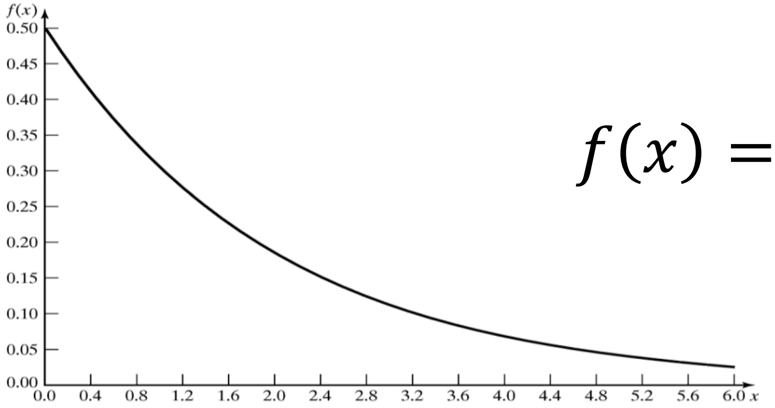
CDF of X

- Properties:
  - $\mathbb{P}(a \le X \le b) = \int_a^b f(x)dx$
  - $\int_{-\infty}^{+\infty} f(x)dx = 1$



Probability Density Function

Example: Life of an inspection device is given by X, a continuous random variable with PDF:

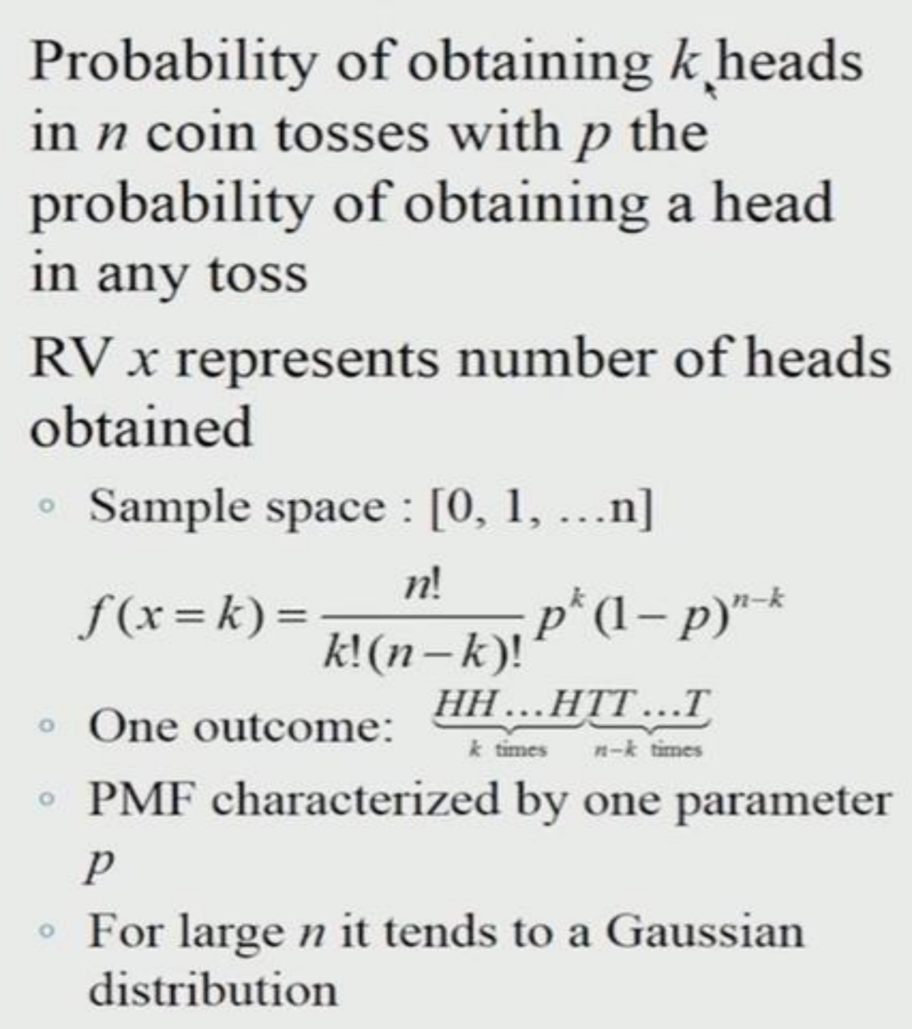


$$f(x) = \frac{1}{2} e^{-\frac{x}{2}}, \quad \text{for } x \geq 0$$

- X has an exponential distribution
- Probability that the device's life is between 2 and 3 years:

$$\mathbb{P}(2 \leq X \leq 3) = \frac{1}{2} \int_2^3 e^{-\frac{x}{2}} dx = 0.14$$

Binomial mass function



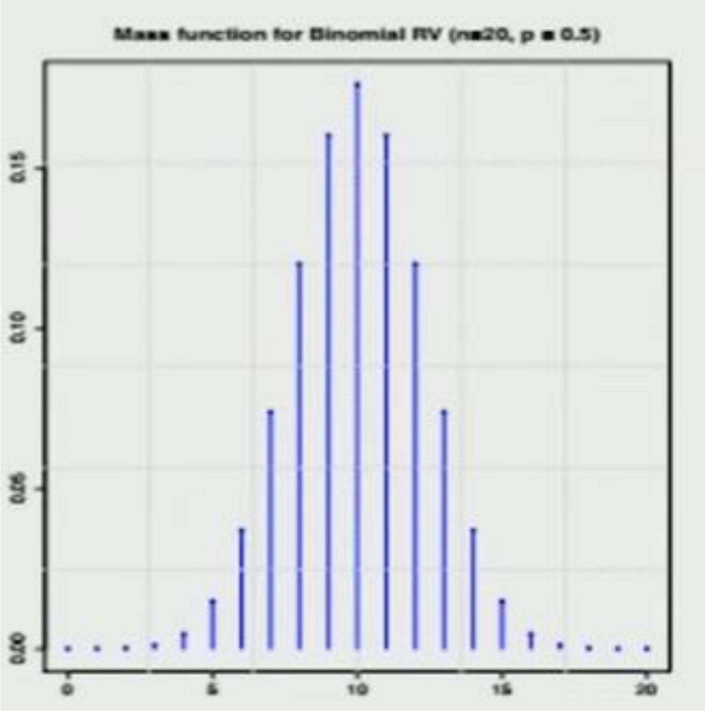
Probability of obtaining k heads in n coin tosses with p the probability of obtaining a head in any toss

RV x represents number of heads obtained

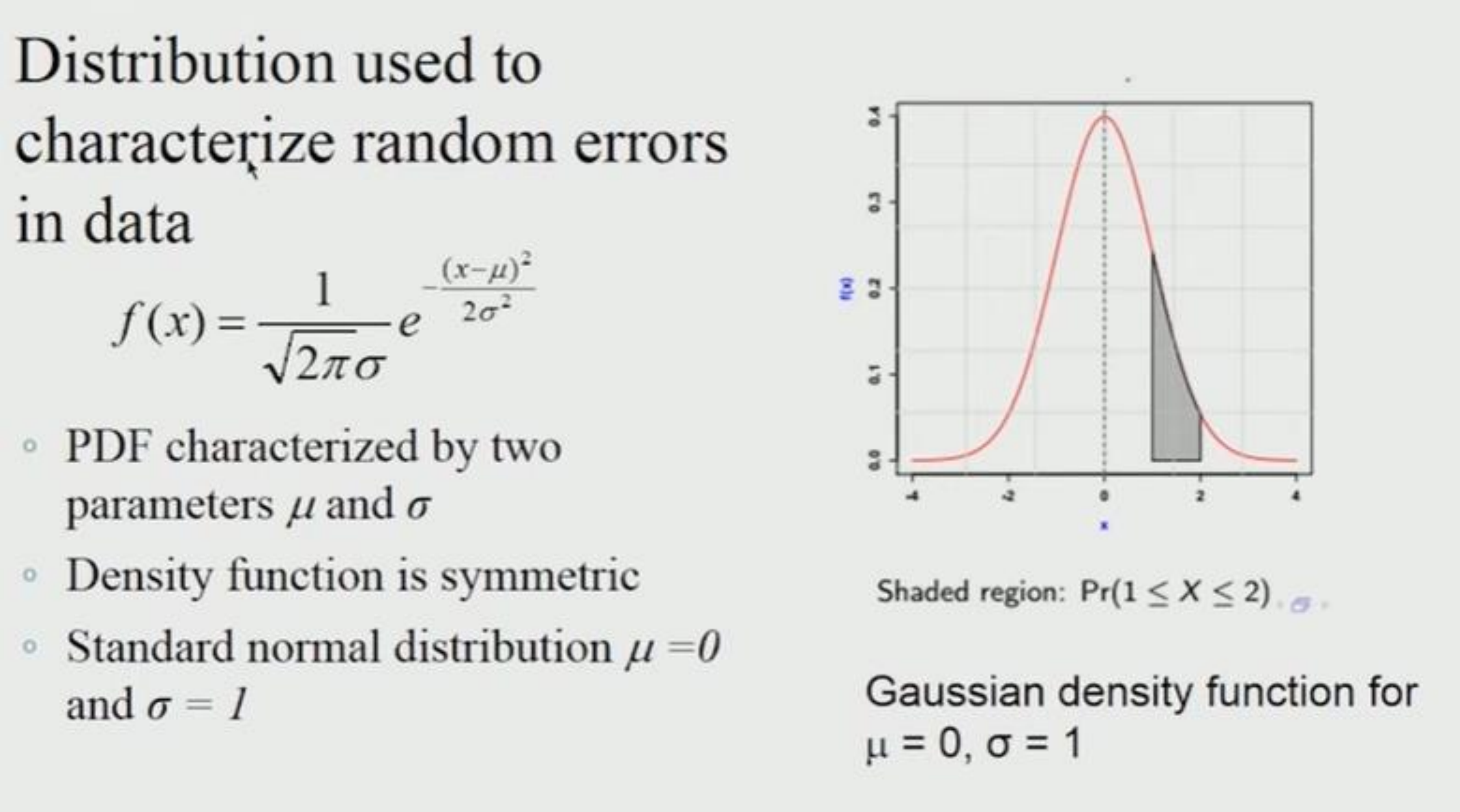
- Sample space : [0, 1, …n]

$$f(x=k)=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

- One outcome: $\underbrace{HH\ldots H}_{k\ \text{times}}\underbrace{TT\ldots T}_{n-k\ \text{times}}$
- PMF characterized by one parameter p
- For large n it tends to a Gaussian distribution



Normal or Gaussian density function



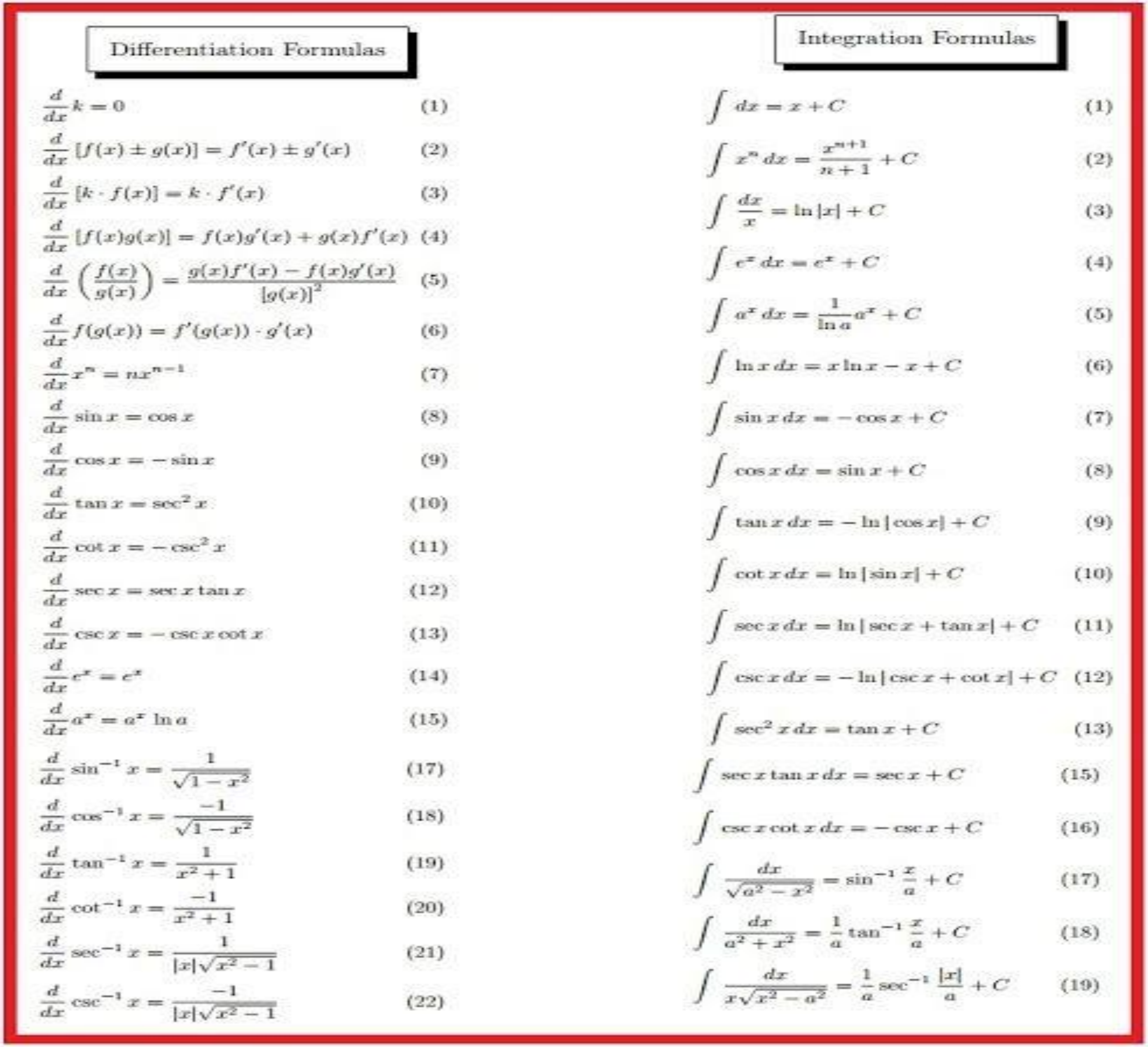
Distribution used to characterize random errors in data

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

- PDF characterized by two parameters μ and σ
- Density function is symmetric
- Standard normal distribution $\mu = 0$ and $\sigma = 1$

Shaded region: $\Pr(1 \leq X \leq 2)$

Gaussian density function for $\mu = 0$, $\sigma = 1$



Differentiation Formulas

$$\frac{d}{dx}k = 0 \tag{1}$$

$$\frac{d}{dx}[f(x) \pm g(x)] = f'(x) \pm g'(x) \tag{2}$$

$$\frac{d}{dx}[k \cdot f(x)] = k \cdot f'(x) \tag{3}$$

$$\frac{d}{dx}[f(x)g(x)] = f(x)g'(x) + g(x)f'(x) \tag{4}$$

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{g(x)f'(x) - f(x)g'(x)}{[g(x)]^2} \tag{5}$$

$$\frac{d}{dx}f(g(x)) = f'(g(x)) \cdot g'(x) \tag{6}$$

$$\frac{d}{dx}x^n = nx^{n-1} \tag{7}$$

$$\frac{d}{dx}\sin x = \cos x \tag{8}$$

$$\frac{d}{dx}\cos x = -\sin x \tag{9}$$

$$\frac{d}{dx}\tan x = \sec^2 x \tag{10}$$

$$\frac{d}{dx}\cot x = -\csc^2 x \tag{11}$$

$$\frac{d}{dx}\sec x = \sec x \tan x \tag{12}$$

$$\frac{d}{dx}\csc x = -\csc x \cot x \tag{13}$$

$$\frac{d}{dx}e^x = e^x \tag{14}$$

$$\frac{d}{dx}a^x = a^x \ln a \tag{15}$$

$$\frac{d}{dx}\sin^{-1} x = \frac{1}{\sqrt{1-x^2}} \tag{17}$$

$$\frac{d}{dx}\cos^{-1} x = \frac{-1}{\sqrt{1-x^2}} \tag{18}$$

$$\frac{d}{dx}\tan^{-1} x = \frac{1}{x^2+1} \tag{19}$$

$$\frac{d}{dx}\cot^{-1} x = \frac{-1}{x^2+1} \tag{20}$$

$$\frac{d}{dx}\sec^{-1} x = \frac{1}{|x|\sqrt{x^2-1}} \tag{21}$$

$$\frac{d}{dx}\csc^{-1} x = \frac{-1}{|x|\sqrt{x^2-1}} \tag{22}$$

Integration Formulas

$$\int dx = x + C \tag{1}$$

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C \tag{2}$$

$$\int \frac{dx}{x} = \ln|x| + C \tag{3}$$

$$\int e^x\,dx = e^x + C \tag{4}$$

$$\int a^x\,dx = \frac{1}{\ln a}a^x + C \tag{5}$$

$$\int \ln x\,dx = x\ln x - x + C \tag{6}$$

$$\int \sin x\,dx = -\cos x + C \tag{7}$$

$$\int \cos x\,dx = \sin x + C \tag{8}$$

$$\int \tan x\,dx = -\ln|\cos x| + C \tag{9}$$

$$\int \cot x\,dx = \ln|\sin x| + C \tag{10}$$

$$\int \sec x\,dx = \ln|\sec x + \tan x| + C \tag{11}$$

$$\int \csc x\,dx = -\ln|\csc x + \cot x| + C \tag{12}$$

$$\int \sec^2 x\,dx = \tan x + C \tag{13}$$

$$\int \sec x \tan x\,dx = \sec x + C \tag{15}$$

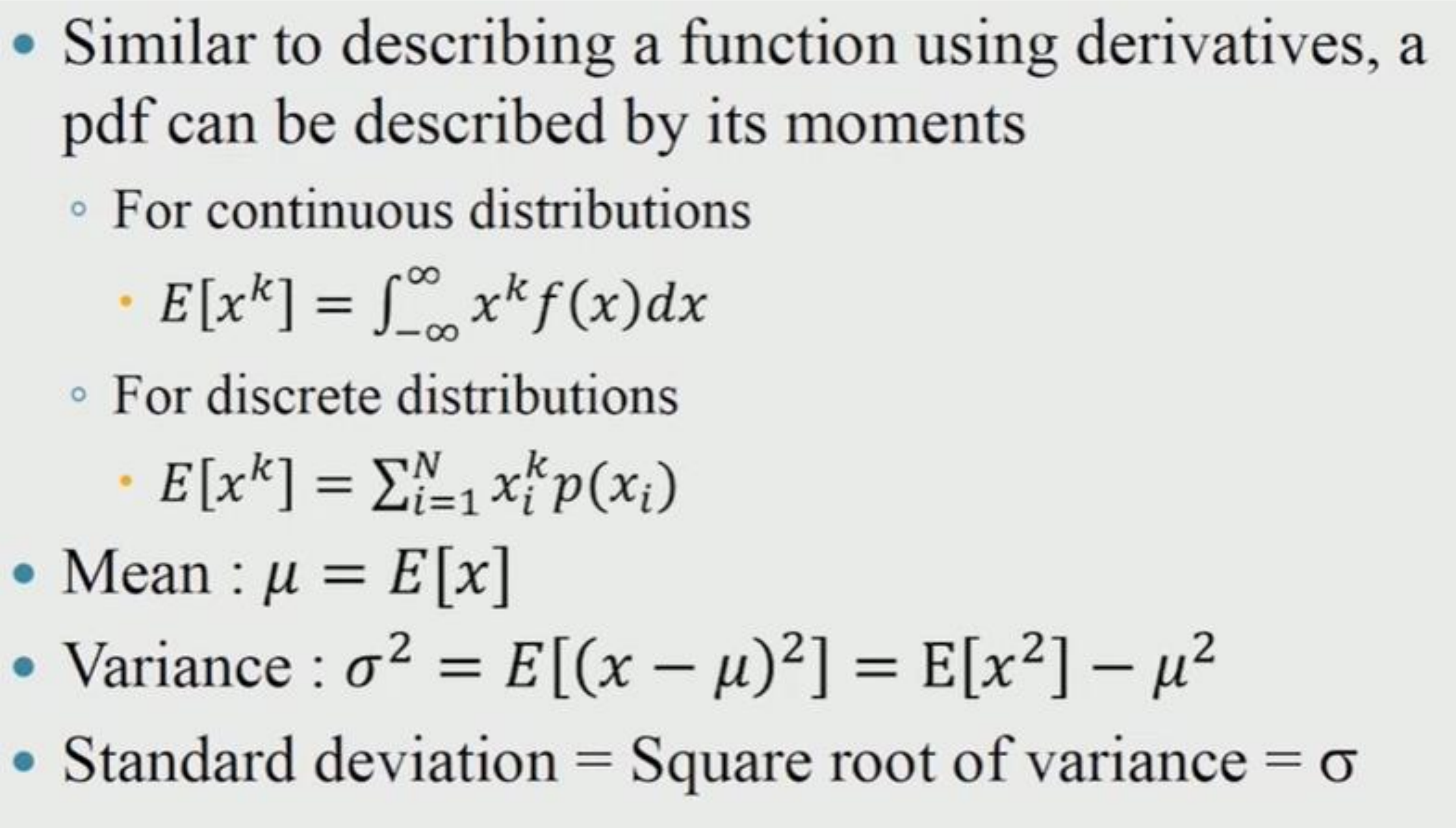
$$\int \csc x \cot x\,dx = -\csc x + C \tag{16}$$

$$\int \frac{dx}{\sqrt{a^2-x^2}} = \sin^{-1}\frac{x}{a} + C \tag{17}$$

$$\int \frac{dx}{a^2+x^2} = \frac{1}{a}\tan^{-1}\frac{x}{a} + C \tag{18}$$

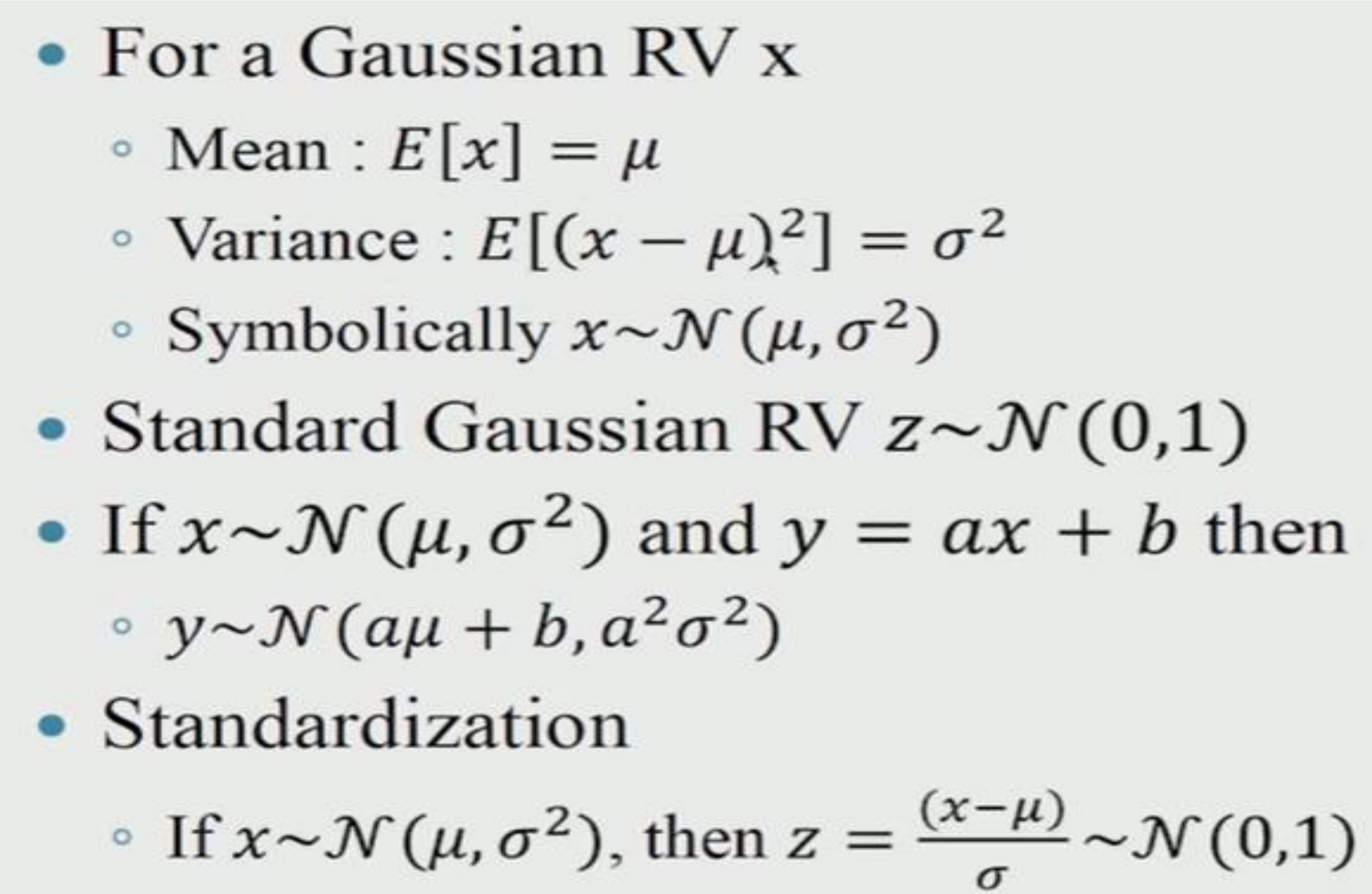
$$\int \frac{dx}{x\sqrt{x^2-a^2}} = \frac{1}{a}\sec^{-1}\frac{|x|}{a} + C \tag{19}$$

Moments of pdf



- Similar to describing a function using derivatives, a pdf can be described by its moments
  - For continuous distributions
    - $E[x^k] = \int_{-\infty}^{\infty} x^k f(x) dx$
  - For discrete distributions
    - $E[x^k] = \sum_{i=1}^{N} x_i^k p(x_i)$
- Mean : $\mu = E[x]$
- Variance : $\sigma^2 = E[(x-\mu)^2] = \mathrm{E}[x^2] - \mu^2$
- Standard deviation = Square root of variance = σ

Properties of Gaussian RVs



- For a Gaussian RV x
  - Mean : $E[x] = \mu$
  - Variance : $E[(x - \mu)^2] = \sigma^2$
  - Symbolically $x \sim \mathcal{N}(\mu, \sigma^2)$
- Standard Gaussian RV $z \sim \mathcal{N}(0,1)$
- If $x \sim \mathcal{N}(\mu, \sigma^2)$ and $y = ax + b$ then
  - $y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
- Standardization
  - If $x \sim \mathcal{N}(\mu, \sigma^2)$, then $z = \frac{(x-\mu)}{\sigma} \sim \mathcal{N}(0,1)$

Try this

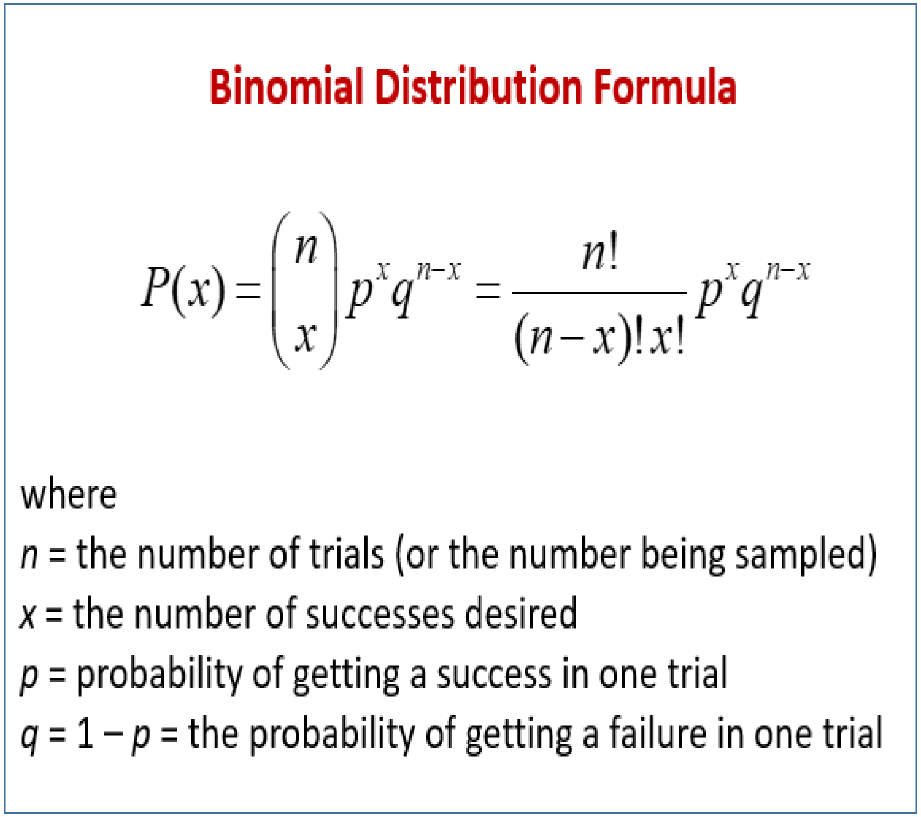
- x=1:20
- y=0.5*exp(-x/2)
- plot(y)
- //to find if subset
- x <- 1:10
- y <- 8:12
- y %in% x
- install.packages("prob")
- library(prob)
- isin(x,y) //need prob package

Try this

- Install.packages(«prob»)
- Library(prob)
- Rolldie, tosscoin, and roulette functions
- Rolldie(1),rolldie(2)
- S <- rolldie(times = 3, makespace = TRUE)
- Prob(S, X1+X2 > 9) //conditional probability
- S <- cards()
- A <- subset(S, suit == "Heart")
- B <- subset(S, rank == "A")
- setdiff(B, A)

Rcode

- R code:
- *p(x) = choose(n, x) p^x (1-p)^(n-x)* for *x = 0, …, n*.
- Pr=dbinom(0:3, size = 3 , prob = 0.5)
- Plot(Pr) //PMF
- A <- data.frame(Pr=dbinom(0:3, size = 3, prob = 0.5)) //flipping coin 3 times
- A <- data.frame(Pr=dbinom(0:4, size = 4, prob = 0.5)) //PMF for number of heads in 4 trials



Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where

n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

$q = 1 - p$ = the probability of getting a failure in one trial

Expectation of a Random Variable

- **Mean** or **Expected Value**:

$$\mu = E[X] = \begin{cases} \sum_{i=1}^{n} x_i p(x_i) & \text{discrete } X \\ \int_{-\infty}^{\infty} x f(x)dx & \text{continuous } X \end{cases}$$

- Example: number of heads in tossing three coins

$E[X] = 0\cdot p(0) + 1\cdot p(1) + 2\cdot p(2) + 3\cdot p(3)$

$= 1\cdot 3/8 + 2\cdot 3/8 + 3\cdot 1/8$

$= 12/8$

$= 1.5$

Expectation of a Function

- $g(X)$: a real-valued function of random variable X
- How to compute $E[g(X)]$?
  - If X is discrete with PMF $p(x)$:

$$E[g(X)] = \sum_{x} g(x)p(x)$$

  - If X is continuous with PDF $f(x)$:

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

- Example: X is the number rolled when rolling a die
  - PMF: $p(x) = 1/6$, for $x = 1,2, \dots, 6$

$$E[X^2] = \sum_{x=1}^{6} x^2 p(x) = \frac{1}{6}(1 + 2^2 + \cdots + 6^2) = \frac{91}{6} = 15.17$$

Properties of Expectation

- X, Y: two random variables
- a, b: two constants

$$E[aX] = aE[X]$$

$$E[X + b] = E[X] + b$$

$$E[X + Y] = E[X] + E[Y]$$

Misuses of Expectations

- ***Multiplying means to get the mean of a product***

$$E[XY] \neq E[X]E[Y]$$

- Example: tossing three coins
  - X: number of heads
  - Y: number of tails
  - $E[X] = E[Y] = 3/2 \Rightarrow E[X]E[Y] = 9/4$
  - $E[XY] = 3/2$

$$\Rightarrow E[XY] \neq E[X]E[Y]$$

- ***Dividing means to get the mean of a ratio***

$$E\left[\frac{X}{Y}\right] \neq \frac{E[X]}{E[Y]}$$

Variance of a Random Variable

- The variance is a measure of the *spread* of a distribution around its mean value
- Variance is symbolized by $V[X]$ or $Var[X]$ or σ^2:
  - Mean is a way to describe the *location* of a distribution
  - Variance is a way to capture its *scale or degree* of being spread out
  - The unit of variance is the square of the unit of the original variable
- σ: standard deviation
  - Defined as the square root of variance $V[X]$
  - Expressed in the same units as the mean

Variance of a Random Variable

- **Variance**: The expected value of the square of distance between a random variable and its mean

$$\sigma^2 = V[X]$$

$$= E[(X-\mu)^2] = \begin{cases} \sum_{i=1}^{n} (x_i - \mu)^2 p(x_i) & \text{discrete } X \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx & \text{continuous } X \end{cases}$$

where, $\mu = E[X]$

- Equivalently:

$$\sigma^2 = E[X^2] - (E[X])^2$$

Properties of Variance

- X, Y: two random variables
- a, b: two constants

$$V[X] \geq 0$$

$$V[aX] = a^2 V[X]$$

$$V[X + b] = V[X]$$

- If X and Y are independent:

$$V[X + Y] = V[X] + V[Y]$$

Covariance

- Covariance between random variables X and Y denoted by $Cov(X,Y)$ or $\sigma_{X,Y}^2$ is a measure of how much X and Y change together

$$\begin{aligned}\sigma_{X,Y}^2 &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

- For independent variables, the covariance is zero:

$$E[XY] = E[X]E[Y]$$

- Note: Although independence always implies zero covariance, the reverse is **not** true

Covariance

- Example: tossing three coins
  - X: number of heads
  - Y: number of tails
  - $E[X] = E[Y] = 3/2$
- $E[XY]$?
  - X and Y depend on each other
  - $Y = 3 – X$
  - $E[XY] = 0×P(0) + 2×P(2)$
    $= 3/2$
- $\sigma^2_{X,Y} = E[XY] – E[X]E[Y]$
  $= 3/2 – 3/2 × 3/2$
  $= – 3/4$

x	y	xy	$p(x)$
0	3	0	1/8
1	2	2	3/8
2	1	2	3/8
3	0	0	1/8

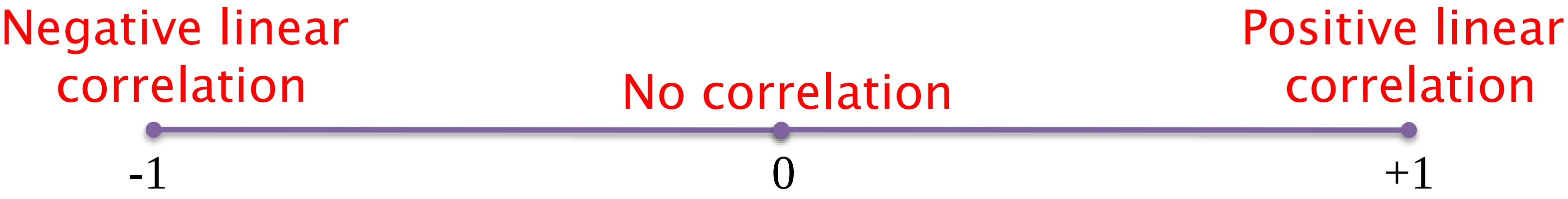
xy	$p(xy)$
0	2/8
2	6/8

Correlation

- Correlation Coefficient between random variables X and Y, denoted by $\rho_{X,Y}$, is the normalized value of their covariance:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}^2}{\sigma_X \sigma_Y}$$

- Indicates the strength and direction of a linear relationship between two random variables
- The correlation always lies between -1 and +1



Negative linear correlation

No correlation

Positive linear correlation

-1

0

+1

-

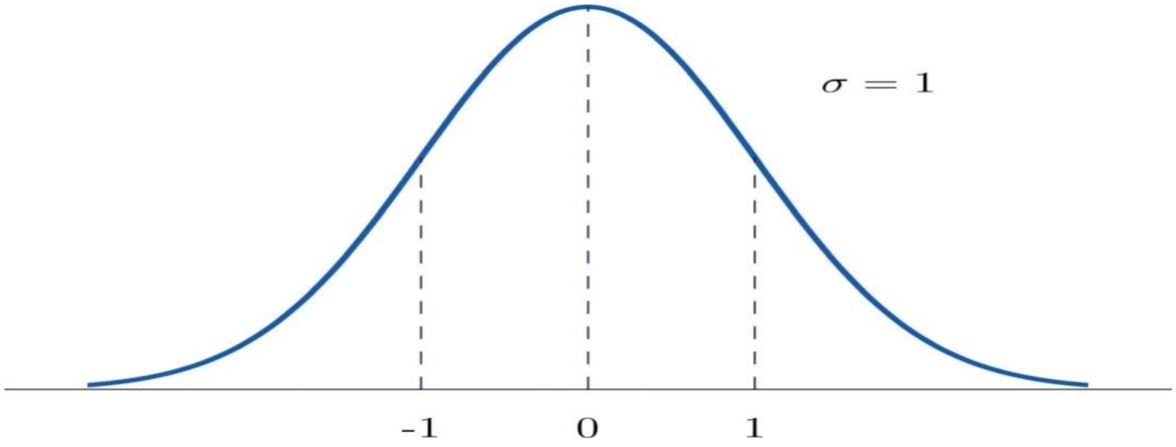
R code

- `plot( iris$Sepal.Length, type="l", col="red" )`
- `par(new=TRUE)`
- `plot( iris$Sepal.Width, type="l", col="green" )`
- `cov(iris$Sepal.Length, iris$Sepal.Width)`
- `cor(iris$Sepal.Length, iris$Sepal.Width)`

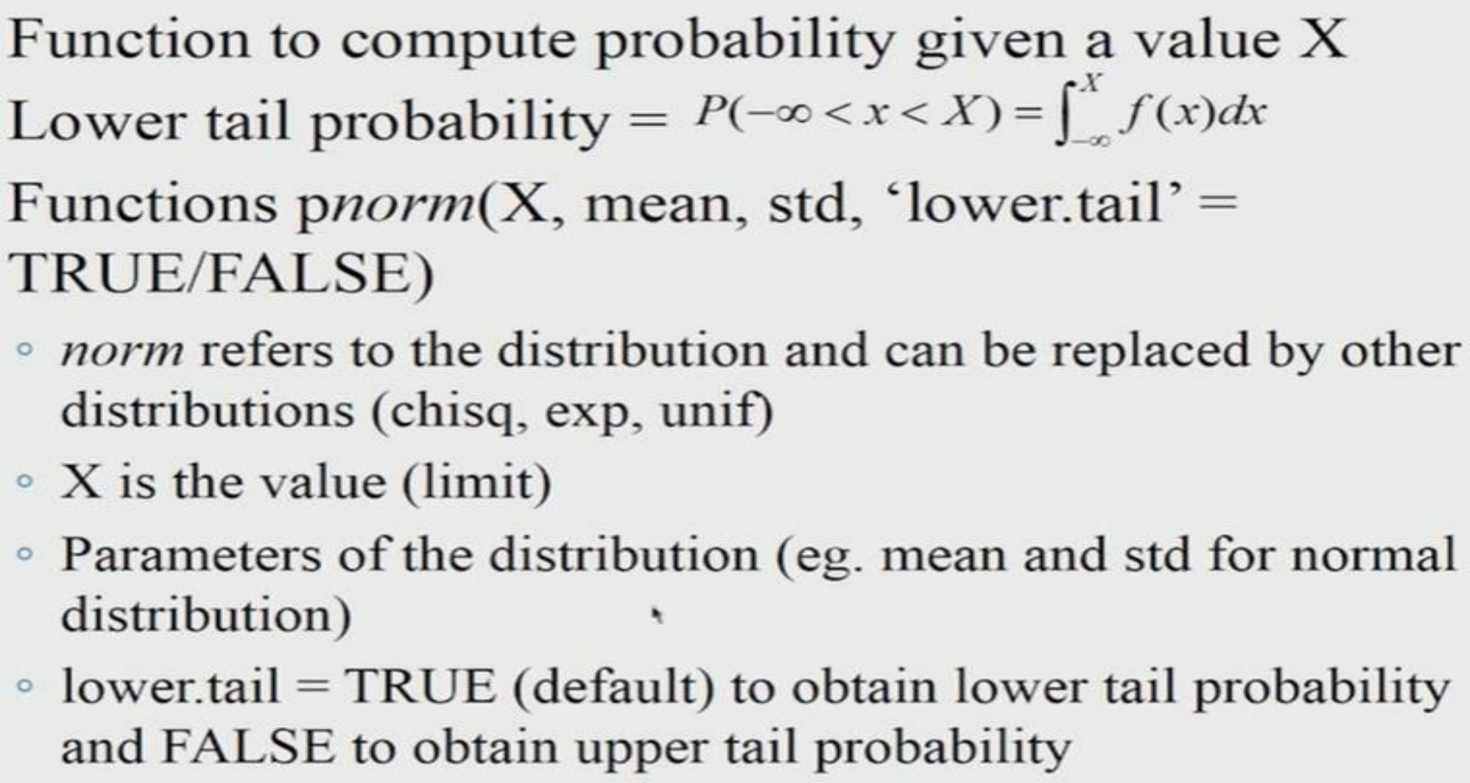
Standard Normal Distribution

- Random variable Z has Standard Normal Distribution if it is normally distributed with parameters $(0, 1)$, i.e., $Z \sim N(0, 1)$:
  - PDF: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, for $-\infty \leq x \leq +\infty$
  - CDF: commonly denoted by $\Phi(z)$:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}x^2} \, dx$$



R Code

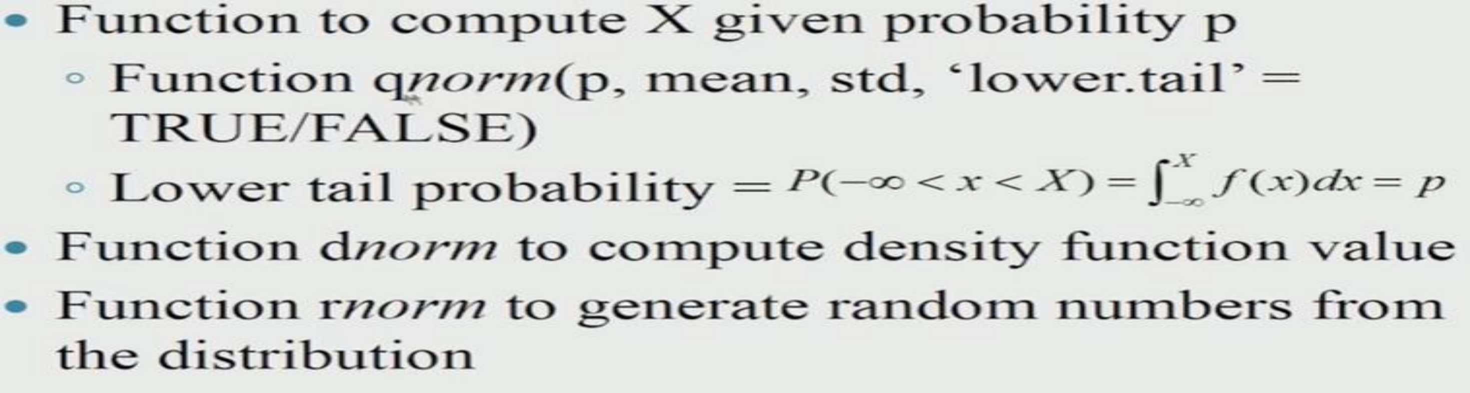


Function to compute probability given a value X

Lower tail probability = $P(-\infty < x < X) = \int_{-\infty}^{X} f(x)dx$

Functions p*norm*(X, mean, std, ‘lower.tail’ = TRUE/FALSE)

- *norm* refers to the distribution and can be replaced by other distributions (chisq, exp, unif)
- X is the value (limit)
- Parameters of the distribution (eg. mean and std for normal distribution)
- lower.tail = TRUE (default) to obtain lower tail probability and FALSE to obtain upper tail probability



- Function to compute X given probability p
  - Function q*norm*(p, mean, std, ‘lower.tail’ = TRUE/FALSE)
  - Lower tail probability = $P(-\infty < x < X) = \int_{-\infty}^{X} f(x)dx = p$
- Function d*norm* to compute density function value
- Function r*norm* to generate random numbers from the distribution

R code

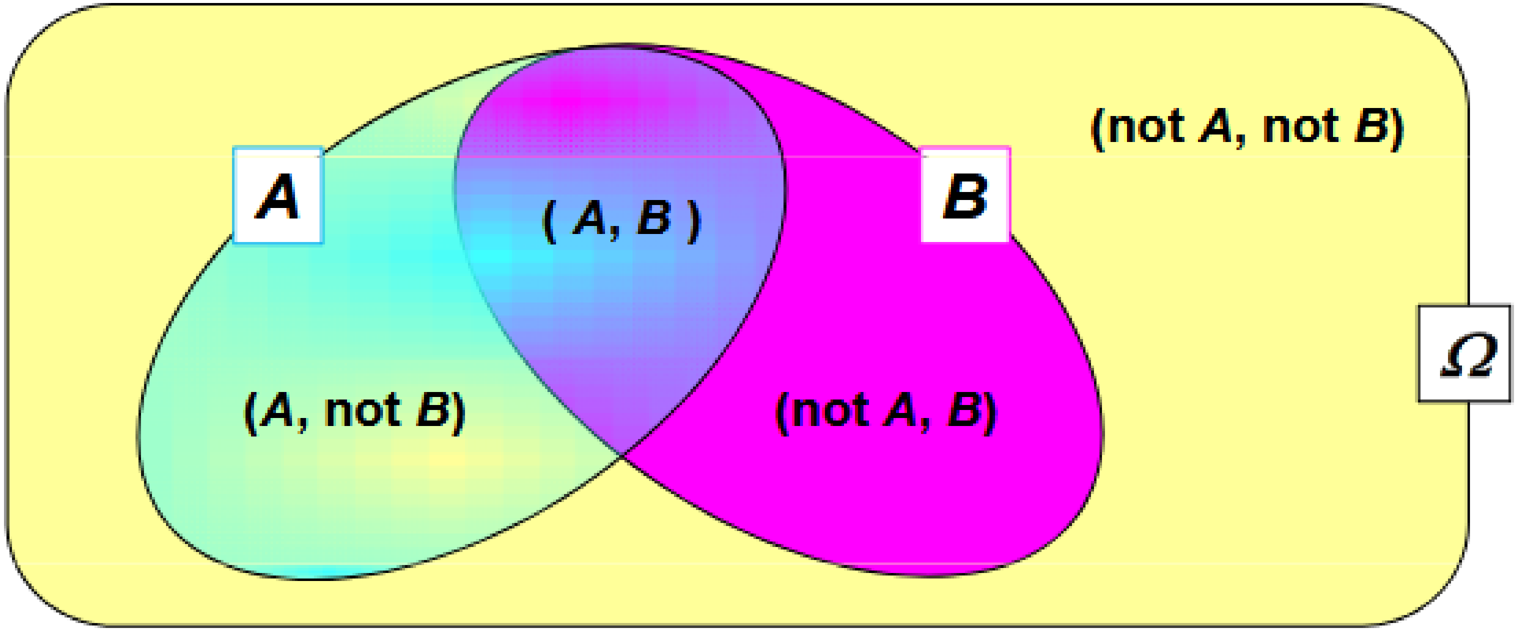
- str(rnorm)
  - **function** (n, mean = 0, sd = 1)
- mydata1 <- rnorm(100, 2, 1) *## Generate some data*
- 100 is assigned to the n argument, 2 is assigned to the mean argument, and 1 is assigned to the sd argument
- plot(mydata1)
- hist(mydata1)
- mydata2 <- rnorm(100, 4, 2)
- hist(mydata1,50,border="red")
- par(new=TRUE)
- hist(mydata2,50,border="red")
- sd(mydata)
- sd(x = mydata, na.rm = **FALSE**)
- intersect, setdiff, union, isin functions in prob package

Bayes Rule

A way to find conditional probabilities for one variable when conditional probabilities for another variable are known.

$$p(\,B \mid A\,) = p(\,A \mid B\,) \cdot p(\,B\,) \,/\, p(\,A\,)$$

where $p(\,A\,) = p(\,A, B\,) + p(\,A, \text{not } B\,)$



Bayes Rule

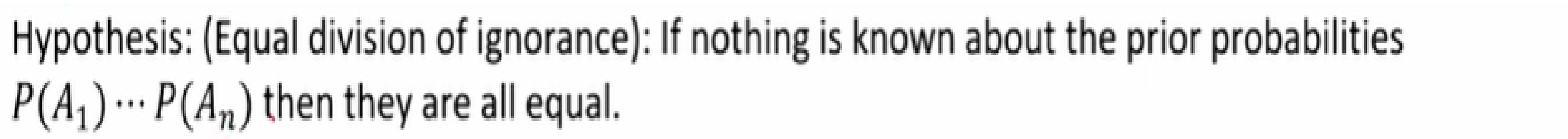
$$\overbrace{P(x/y)}^{\text{posterior}} = \frac{\overbrace{P(y/x)}^{\text{likelihood}}\overbrace{P(x)}^{\text{prior}}}{\underbrace{P(y)}_{\text{evidence}}}$$

Assume,
1. we have some prior knowledge $\boldsymbol{p}(\boldsymbol{x})$
2. a random variable $\boldsymbol{y}$, assumes value from random sample
3. We can update the knowledge $\boldsymbol{p}(\boldsymbol{x})$ after observing $\boldsymbol{y}$, $\boldsymbol{P}(\boldsymbol{x}|\boldsymbol{y})$

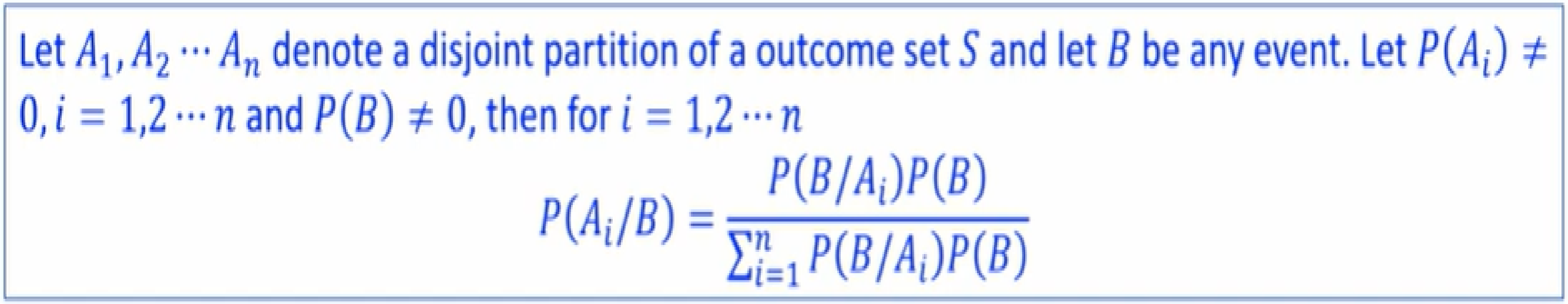
- The prior distribution represents initial beliefs before observing any data.
- The posterior distribution represents updated beliefs after taking into account both the prior information and the observed data.

The likelihood function quantifies how well different values explain the observed data.

Bayes Rule

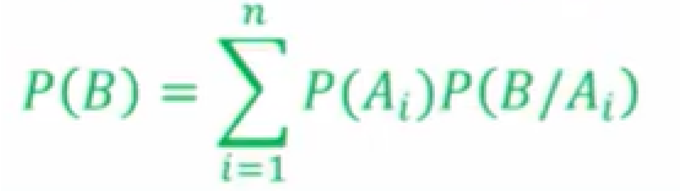


Hypothesis: (Equal division of ignorance): If nothing is known about the prior probabilities $P(A_1) \cdots P(A_n)$ then they are all equal.



Let $A_1, A_2 \cdots A_n$ denote a disjoint partition of a outcome set S and let B be any event. Let $P(A_i) \neq 0, i = 1,2 \cdots n$ and $P(B) \neq 0$, then for $i = 1,2 \cdots n$

$$P(A_i/B) = \frac{P(B/A_i)P(B)}{\sum_{i=1}^{n} P(B/A_i)P(B)}$$



$$P(B) = \sum_{i=1}^{n} P(A_i)P(B/A_i)$$

$P(Ai) \neq 0$

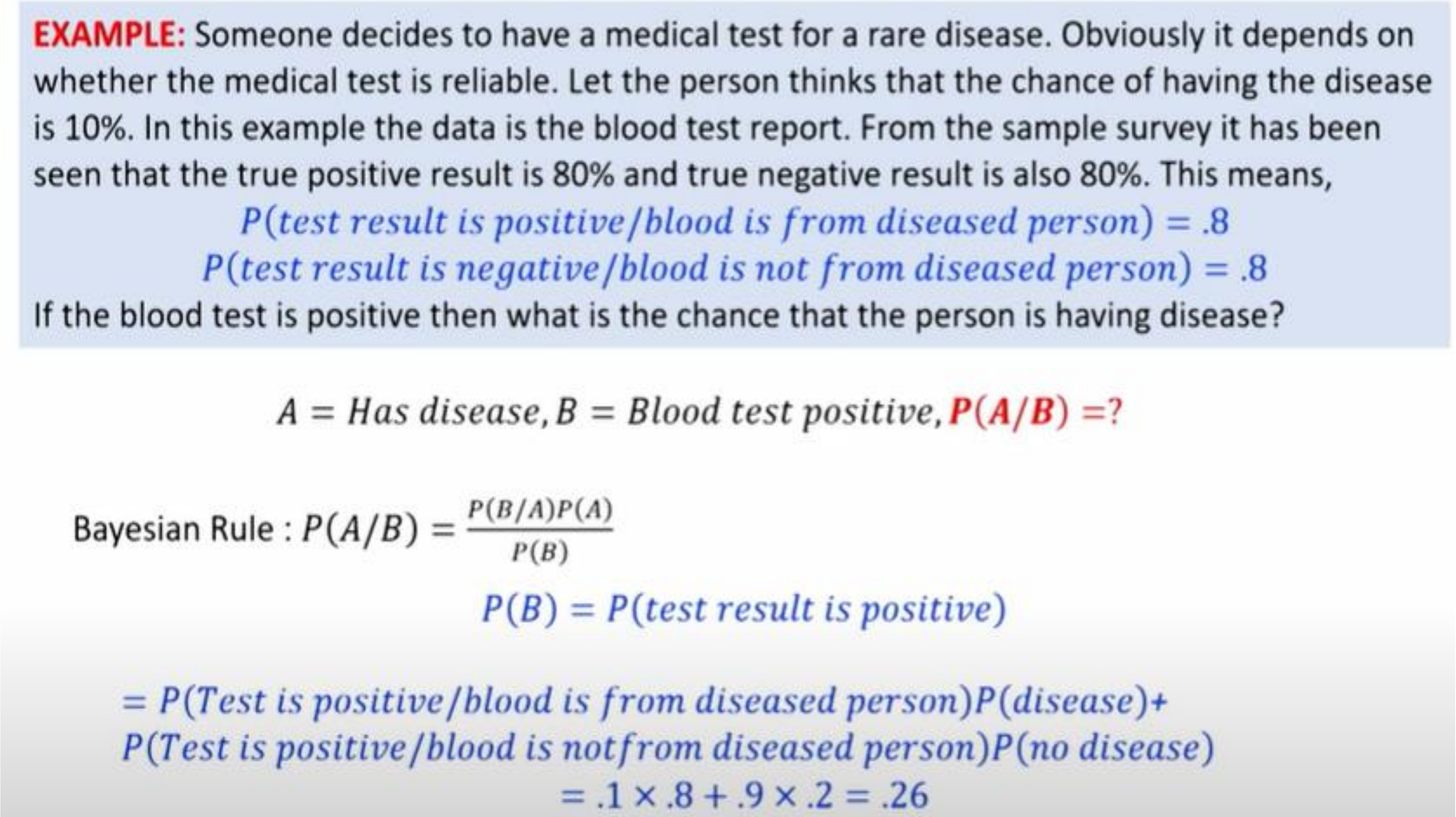
$$P(A_i/B) = \frac{P(B/A_i)P(A_i)}{\sum_{i=1}^{n} P(B/A_i)P(A_i)}$$

Example

- Marie is getting married tomorrow at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman is forecasting rain for tomorrow. When it actually rains, the weatherman has forecast rain 90% of the time. When it doesn't rain, he has forecast rain 10% of the time. What is the probability it will rain on the day of Marie's wedding? Probability it will rain on the day of Marie’s wedding?
- Event A: The weatherman has forecast rain.
- Event B: It rains.

Solution

- p(B) = 5 / 365 = 0.0137 [It rains 5 days out of the year.]
- P(not B)=1-p(B)=0.9863
- P(A|B)=0.9
- P(A|not B)=0.1
- P(B|A)=P(A|B)*P(B)/P(A)
- P(A)=P(A|B)*P(B)+P(A|not B)*P(not B)=0.9*0.0137+0.1*0.9863=0.111
- P(B|A)=0.9*0.0137/0.111=0.111
- The result seems unintuitive but is correct. Even when the weatherman predicts rain, it only rains only about 11% of the time. Despite the weatherman's gloomy prediction, it is unlikely Marie will get rained on at her wedding.



EXAMPLE: Someone decides to have a medical test for a rare disease. Obviously it depends on whether the medical test is reliable. Let the person thinks that the chance of having the disease is 10%. In this example the data is the blood test report. From the sample survey it has been seen that the true positive result is 80% and true negative result is also 80%. This means,

$$P(\textit{test result is positive}/\textit{blood is from diseased person}) = .8$$

$$P(\textit{test result is negative}/\textit{blood is not from diseased person}) = .8$$

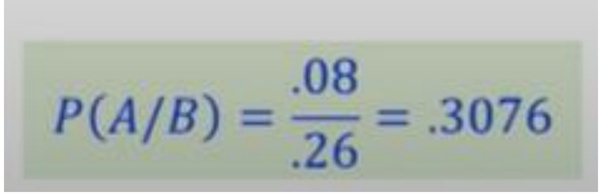
If the blood test is positive then what is the chance that the person is having disease?

$$A = \textit{Has disease}, B = \textit{Blood test positive}, \mathbf{P(A/B)} = ?$$

Bayesian Rule : $P(A/B) = \frac{P(B/A)P(A)}{P(B)}$

$$P(B) = P(\textit{test result is positive})$$

$$= P(\textit{Test is positive}/\textit{blood is from diseased person})P(\textit{disease}) +$$
$$P(\textit{Test is positive}/\textit{blood is notfrom diseased person})P(\textit{no disease})$$
$$= .1 \times .8 + .9 \times .2 = .26$$



$$P(A/B) = \frac{.08}{.26} = .3076$$

How to extend probability to reason about uncertain continuous quantities

- Suppose X is some uncertain continuous quantity.
- The probability that X lies in any interval $a \leq X \leq b$ can be computed as follows.
- Define the events: if A and W are mutually exclusive and P(B)=p(A)+p(W)

$$A = (X \leq a), \quad B = (X \leq b) \; W = (a < X \leq b) \quad B = A \cup W$$

$$P(W) = P(B) - P(A)$$

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

$$P(W) = P(a < X \leq b) = F(b) - F(a)$$

Define the function: $F(q) = P(X \leq q)$ ⟵ ***cumulative distribution function (cdf) of X, which is monotonically increasing***

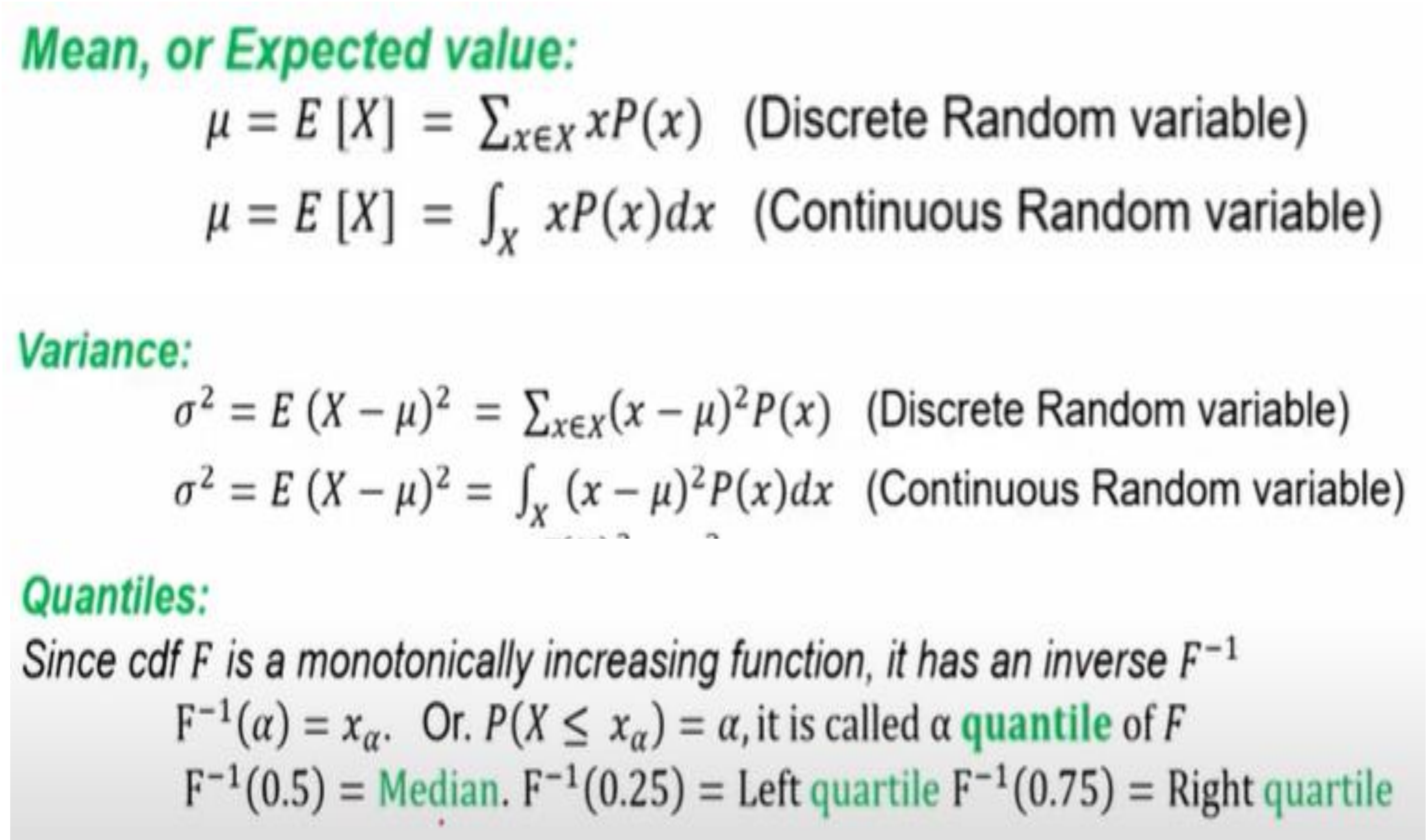
Define the function : $f(x) = \frac{d}{dx}F(x)$ (we assume this derivative exists);

this is called the probability density function or pdf

$$F(b) - F(a) = P(a < X \leq b) = \int_a^b f(x)dx$$

If the size of the interval gets smaller, we can write $P(x < X \leq x + dx) \approx p(x)dx$

How to extend probability to reason about uncertain continuous quantities



Mean, or Expected value:

$$\mu = E\,[X] = \sum_{x \in X} xP(x) \quad \text{(Discrete Random variable)}$$

$$\mu = E\,[X] = \int_X xP(x)dx \quad \text{(Continuous Random variable)}$$

Variance:

$$\sigma^2 = E\,(X-\mu)^2 = \sum_{x \in X}(x-\mu)^2 P(x) \quad \text{(Discrete Random variable)}$$

$$\sigma^2 = E\,(X-\mu)^2 = \int_X (x-\mu)^2 P(x)dx \quad \text{(Continuous Random variable)}$$

Quantiles:

Since cdf F is a monotonically increasing function, it has an inverse F^{-1}

$\mathrm{F}^{-1}(\alpha) = x_\alpha$. Or. $P(X \leq x_\alpha) = \alpha$, it is called α quantile of F

$\mathrm{F}^{-1}(0.5)$ = Median. $\mathrm{F}^{-1}(0.25)$ = Left quartile $\mathrm{F}^{-1}(0.75)$ = Right quartile

Discrete probability Distribution

- Bernoulli – used for modeling binary outcomes, such as the success or failure of an event.
- Binomial –used to describe the number of successes in a repeated number of independent Bernoulli trials
- Poisson –used to get probability of the number of occurrences of a rare event
- Geometric –used in modeling the number of trials until a certain event occurs
- Negative Binomial –used to get the probability of getting a number of failures before the rth success
- Many practical problems we examine appropriate conditions and select the probability distribution
- A range of discrete probability distributions are considered here

Statistical Estimation- MLE vs MAP

- E.g. **if your sample consists of 100 students and the average height of your sample is 160 cm, you might estimate that the average height of all students in the school is around 162 cm, with 95% confidence average height of all students is in between 155 cm to 165 cm.**
- Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP) – both used to estimate parameters of statistical model.
- Both methods assume that the data are independent and identically distributed (*i.i.d.*) samples $X1$,$X2$,…,Xk.
- The central idea behind MLE is to select that parameter θ value that maximizes the probability of observing the sample. MLE do not incorporate prior info/beliefs about model parameters.
- The likelihood function quantifies how well different values explain the observed data. Likelihood of all of our data is the product of the likelihood of each data point.
- In the case of discrete distributions, likelihood is the joint probability of data points. In the case of continuous distribution, likelihood refers to the joint probability density of data.
- In MAP process we trying to calculate the conditional probability of unobserved parameter θgiven observed random variables. MAP incorporate prior info/beliefs about model parameters by multiplying likelihood function with prior distribution of parameters.

MLE Estimation

If k random samples $X_1, X_2, .., X_k$ are drawn from a population with probability function $f(X|\theta)$, then joint probability function L of $X_1, X_2, .., X_k$ is,

$$\boldsymbol{L(\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_k|\theta) = \prod_{i=1}^{k} f(x_i|\theta)}$$

L: Likelihood function.

Since, $X_1, X_2, .., X_k$ are independently and identically distributed.

Objective: To find θ, such that Likelihood function is maximum

i.e. $\theta_{MLE} = \underset{\theta}{argmax}\ L(\theta)$

EX: Find MLE of θ in $Ber(X|\theta) = \begin{cases} \theta. & \text{if } X = 1 \\ (1-\theta) & \text{if } X = 0 \end{cases}$

$$Ber(X|\theta) = \theta^X(1-\theta)^{1-X}, X \in \{0,1\}$$

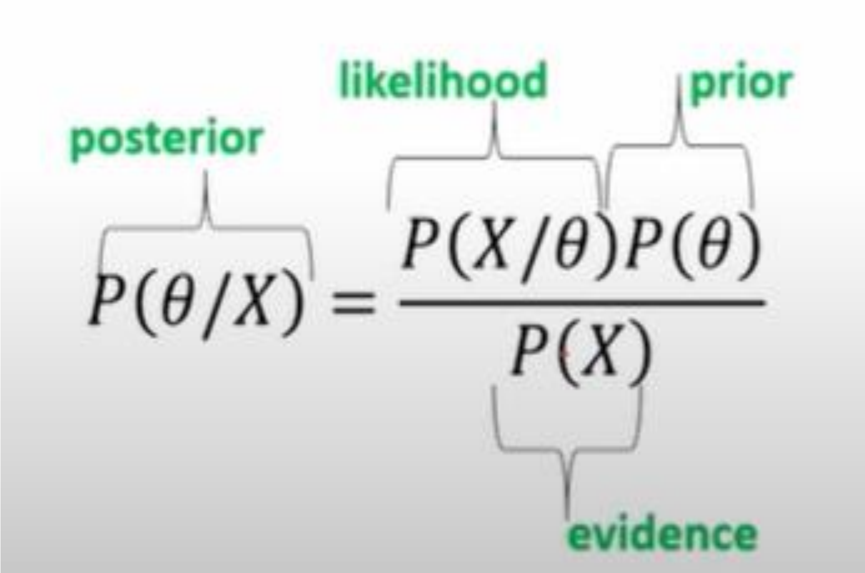
$$L(\theta) = \prod_{i=1}^{k} f(x_i|\theta) = \theta^{x_1}(1-\theta)^{1-x_1}\,\theta^{x_2}(1-\theta)^{1-x_2}\ldots\theta^{x_n}(1-\theta)^{1-x_n}$$

$$\log L = \log\{\theta^{x_1+x_2+\cdots x_n}(1-\theta)^{n-(x_1+x_2+\cdots x_n)}\}$$

$$\frac{d.\log L}{d\theta} = n\left(\frac{\bar{x}}{\theta} - \frac{\overline{1-x}}{1-\theta}\right) \text{ implies } \theta_{MLE} = \bar{X}$$

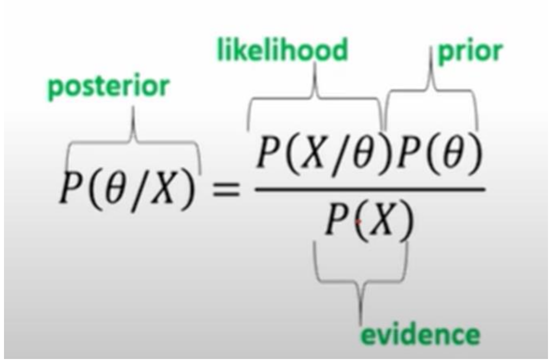
Process of Bayesian estimation

- Assume,
- we have some prior knowledge $p(\boldsymbol{\theta})$
- a random variable y, assumes value from random sample
- Prior Distribution $p(\boldsymbol{\theta})$indicates our probability of belief that $\boldsymbol{\theta}$is a true parameter, prior to seeing any evidence at all.
- We can update the knowledge $p(\boldsymbol{\theta})$after observing y,$\boldsymbol{P(\theta/X)}$



$$\overbrace{P(\theta/X)}^{\text{posterior}} = \frac{\overbrace{P(X/\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(X)}_{\text{evidence}}}$$

MAP Estimation



$$\overbrace{P(\theta/X)}^{\text{posterior}} = \frac{\overbrace{P(X/\theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(X)}_{\text{evidence}}}$$

$$\theta_{MAP} = \underset{\theta}{argmax}\ f(\theta|X_1, X_2 \cdots X_n)$$

$$= \underset{\theta}{argmax}\ \frac{f(X_1, X_2 \cdots X_n|\theta)g(\theta)}{h(X_1, X_2 \cdots X_n)}$$ applying Bayes theorem

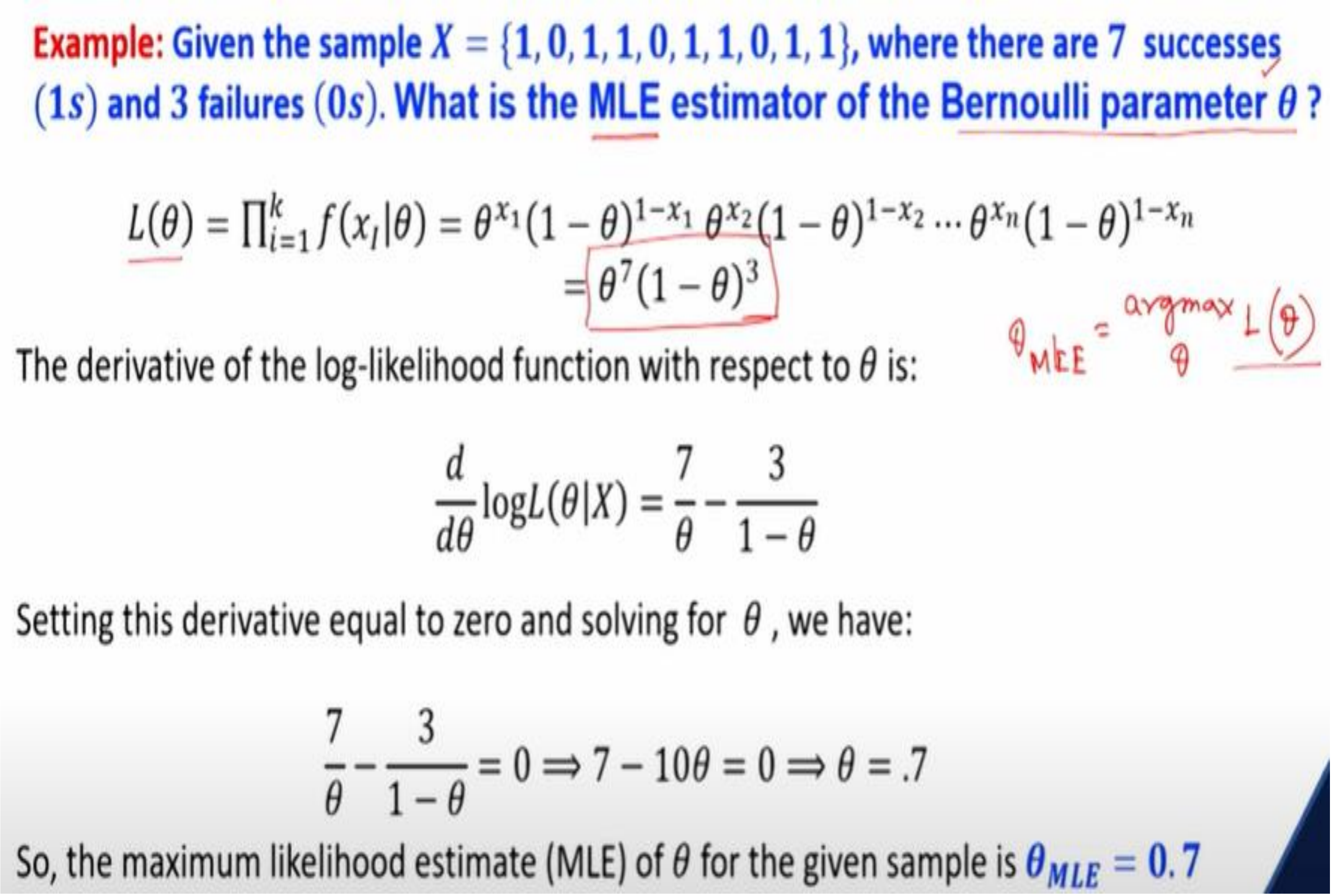
$$= \underset{\theta}{argmax}\ \prod_{i=1}^{k} f(X_i|\theta)g(\theta).$$ since h is constant with respect to θ

Considering log of MAP function

$$\theta_{MAP_c} = \underset{\theta}{argmax}\left(\log g(\theta) + \sum_{i=1}^{n} \log f(X_i|\theta)\right)$$

$g(\theta)$: Prior distribution

θ_{MAP} is the most likely θ given the data $X_1, X_2 \cdots X_n$



Example: Given the sample $X = \{1, 0, 1, 1, 0, 1, 1, 0, 1, 1\}$, where there are 7 successes ✓ $(1s)$ and 3 failures $(0s)$. **What is the MLE estimator of the Bernoulli parameter θ ?**

$$\underline{L(\theta)} = \prod_{i=1}^{k} f(x_i|\theta) = \theta^{x_1}(1-\theta)^{1-x_1}\,\theta^{x_2}(1-\theta)^{1-x_2}\cdots\theta^{x_n}(1-\theta)^{1-x_n}$$

$$= \boxed{\theta^7(1-\theta)^3}$$

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\ \underline{L(\theta)}$$

The derivative of the log-likelihood function with respect to θ is:

$$\frac{d}{d\theta}\log L(\theta|X) = \frac{7}{\theta} - \frac{3}{1-\theta}$$

Setting this derivative equal to zero and solving for θ , we have:

$$\frac{7}{\theta} - \frac{3}{1-\theta} = 0 \Rightarrow 7 - 10\theta = 0 \Rightarrow \theta = .7$$

So, the maximum likelihood estimate (MLE) of θ for the given sample is $\boldsymbol{\theta_{MLE} = 0.7}$

Consider the same example with 7 successes and 3 failures. What is the MAP estimator of the Bernoulli parameter θ, if we assume a prior on θ of $\underline{Beta\ (5,4)}$?

$\theta_{MAP} = ?$

Model: Bernoulli with parameter θ: $i.e.\ \underline{Ber(\theta|7,3) = \theta^7(1-\theta)^3}$

Prior: Consider conjugate distribution as Beta with parameter θ: $Beta\ (5,4)$

$$Beta(\theta|5,4) = \left.\frac{\theta^4(1-\theta)^3}{\Gamma(9)/\Gamma(5)\Gamma(4)}\right] n$$

$$\theta_{MAP} = \underset{\theta}{argmax}\,(\log \underline{g(\theta)} + \textstyle\sum_{i=1}^{n} \underline{\log f(X_I|\theta)})$$

$$\theta_{MAP} = \underset{\theta}{\text{argmax}} \frac{\overset{\downarrow}{f}(x_1, x_2 \cdots x_n)/\theta)\ \overset{\downarrow}{P}(\theta)}{(h(x_1, \underset{\uparrow}{,} x_2 \cdots x_n))}$$

Determine log prior + log likelihood

$$\log g(\theta) + \sum_{i=1}^{n} \log f(X_I|\theta) = \underline{\log}\left(\theta^4(1-\theta)^3\right) - \log n + \underline{\log \theta^7(1-\theta)^3}$$

Differentiate w.r.t. θ, set to 0

$$\boldsymbol{\theta_{MAP} = \frac{11}{17} = .647 \text{ and } \theta_{MLE} = .7}$$

References

- https://courses.washington.edu/css490/2012.Winter/lecture_slides/02_math_essentials.pdf
- Christopher Bishop: ”Pattern Recognition and Machine Learning” , 2006
- Kevin Murphy: ”Machine Learning: a Probabilistic Perspective”
- David Mackay: ”Information Theory, Inference, and Learning Algorithms”
- Ethem Alpaydin: ”Introduction to Machine Learning” , 2nd edition, 2010.
- R. Duda, P. Hart & D. Stork, ***Pattern Classification*** (2nd ed.), Wiley T. Mitchell, ***Machine Learning,*** McGraw-Hill

Thank you

- ???

In short -What is probability?

- **Frequentist interpretation:**
  - If we flip the coin repeatedly it is expected that half of the time there will be head.
  - Repeated identical trials probability is based on the no of favourable cases. **R**atio of frequencies as n tends to infinity gives probability measure.
- **Bayesian interpretation:**
  - It is based on our belief that the coin is equally likely to land heads or tails on the next toss.
  - Here, uncertainty about events are calculated that do not have long term frequencies.
  - This event will happen zero or one times, but cannot happen repeatedly.
  - e.g. what is the probability that the mail which has come just now is ***spam***, the idea of repeated trials does not make sense here.
- ***The basic rules of probability theory are the same, no matter which interpretation is adopted***