Machine Learning
Dr. H C Vijayalakshmi

References

- http://www.pitt.edu/~super4/34011-35001/34301.ppt
- https://math.usask.ca › S244 10 Probability
- CS479/679 Pattern Recognition by Dr. George Bebis
- Pattern Recognition and image Analysis by Earl Gose, Johnsonbaugh and Steve Jost
- Data Analytics CS40003 by Dr. Debasis Mahanta

- **Pattern recognition** is the process of recognizing patterns by using a machine learning algorithm. Pattern recognition can be defined as the classification of data based on knowledge already gained or on statistical information extracted from patterns and/or their representation. One of the important aspects of pattern recognition is its application potential.
- **Pattern** is anything which has a regular sequence of occurrence. A pattern can be either through visualization or it can be derived mathematically by applying algorithms.
- **Patterns include repeated trends in various forms of data**
- Examples : Speech pattern, print on clothes, design of outfits, jewelry pattern, Sound wave, tree species, fingerprint, face, barcode, QR-code, handwriting, or character image etc.
- **Examples:** Speech recognition, speaker identification, multimedia document recognition (MDR), automatic medical diagnosis.

The data inputs for pattern recognition can be words or texts, images, or audio files. Hence, pattern recognition is broader compared to computer vision that focuses on image recognition.

In a typical pattern recognition application, the raw data is processed and converted into a form that is convenient for a machine to use. Pattern recognition involves the classification and cluster of patterns.

Definition of Pattern Recognition

- Pattern recognition is defined as the study of how machines can observe the environment, learn to distinguish various patterns of interest from their background, and make logical decisions about the categories of the patterns. **During recognition, the given objects are assigned to a specific category**.

- In general, pattern recognition can be described as an information reduction, information mapping, or information labeling process.

- In computer science, **pattern recognition refers to the process of matching information already stored in a database with incoming data based on their attributes or features.**

Input Data and Output Response for Various Applications

Task of Classification	Input Data	Output Response
Character Recognition	Optical Signals or Strokes	Name of the character
Speech Recognition	Acoustic Waveforms	Name of the word
Speaker Recognition	Voice	Name of the speaker
Weather Prediction	Weather Maps	Weather Forecasts
Medical Diagnosis	Symptoms	Disease
Stock Market Prediction	Financial News and Charts	Predicted Market ups and Downs

Applications

Application	Input Pattern	Pattern Classes
Optical character recognition	Document image	Characters, words
Internet search	Text document	Semantic categories
Junk mail filtering	Email	Junk/non-junk
Internet search	Video clip	Video genres
Telephone directory assis-tance	Speech waveform	Spoken words
Information extraction	Sentences	Parts of speech
Personal identification	Face, iris, fingerprint	Authorized users for access control
Computer aided diagnosis	Microscopic image	Cancerous/healthy cell
Automatic target recognition	Optical or infrared image	Target type
Printed circuit board inspec-tion	Intensity or range image	Defective/non-defective prod-uct
Fruit sorting	Images taken on a conveyor belt	Grade of quality
Forecasting crop yield	Multispectral image	Land use categories
Sequence analysis	DNA sequence	Known types of genes
Searching for meaningful pat-terns	Points in multidimensional space	Compact and well-separated clusters

Handwriting Recognition

From
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

Nov 10, 1999

To
Dr. Bob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

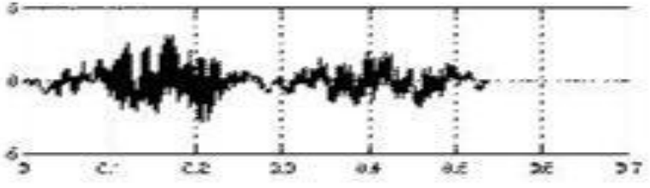
However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Jim

Number / Licence Plate Recognition

Biometric Recognition



Face Detection/Recognition

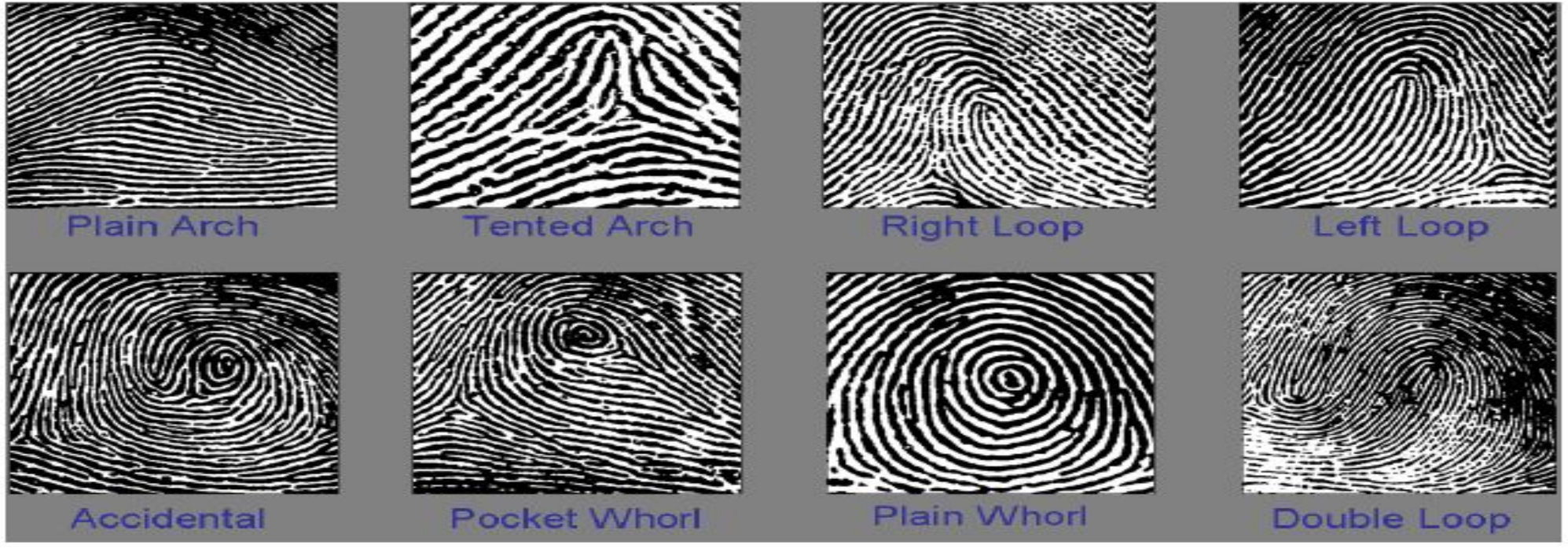
Detection

Matching

Recognition

YCbCr segmentation

Fingerprint Classification

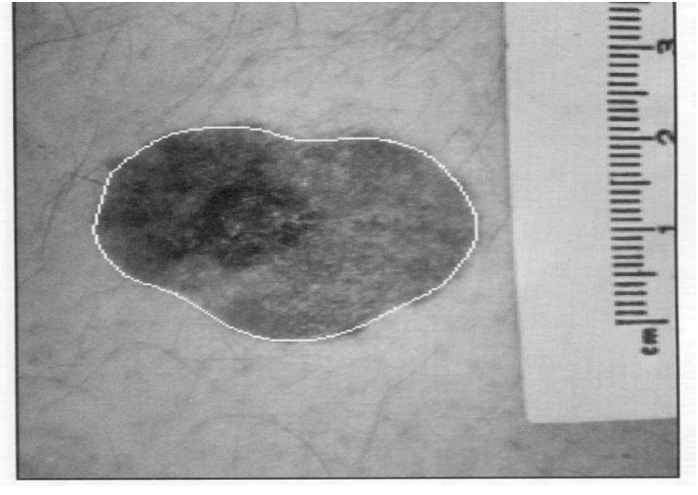


Unmanned Vehicles

Obstacle detection and avoidance Object recognition

Medical Applications

Skin Cancer Detection



Breast Cancer Detection

Land Cover Classification

(using aerial or satellite images)

Many applications including “precision” agriculture.

What is Machine Learning?

Machine learning is a field of Computer Science and Artificial Intelligence (AI) that uses data and algorithms to make the machine learn the way humans learn.

Machine learning uses algorithms and data patterns to build a model and using the model can produce a desired output.

Through the use of data and algorithms, machines can make predictions, recommendations, decisions as well as classifications.

Some common uses of Machine learning include:

Providing online recommendations,

Filtering spam emails

Giving web search results.

Classify objects and assign appropriate class labels

Machine Learning

- Machine learning uses algorithms and data to teach machines how to perform tasks performed by humans. This can help businesses reach a desired outcome either with the assistance of human labor or through machine actions.
- The classifier to be designed is built using input samples which is a mixture of all the classes.
- The classifier learns how to discriminate between samples of different classes.
- If the Learning uses Supervised method then, the classifier is first given a set of training samples and the optimal decision boundary found, and then the classification is done.
- If the learning uses unsupervised method then there is no teacher and no training samples (Unsupervised) are used for building the model. The input samples are the test samples itself. The classifier learns and classifies at the same time.

There are four categories of learning.

- **Supervised learning** makes use of a set of examples which already have the class labels assigned to them.
- **Unsupervised learning** attempts to find inherent structures in the data.
- **Reinforcement learning** uses a trial-and-error method to improve and learn from new situations.
- **Semi-supervised learning** makes use of a small number of labeled data and a large number of unlabeled data to learn the classifier.

Classification is the task of assigning a class label to an input pattern. The class label indicates one of a given set of classes. The classification is carried out with the help of a model obtained using a learning procedure. Model is built according to the type of learning used.

- Supervised learning: (classification) is seen as supervised learning from labeled examples.
  - Supervision: The data (observations, measurements, etc.) are labeled with pre-defined classes. It is like that a "teacher" gives the classes (supervision).
  - Test data are classified into different classes.

- Unsupervised learning (clustering)
  - Class labels of the data are unknown
  - Given a set of data, the task is to establish the existence of classes or clusters in the data

Supervised learning

Supervised learning is when a machine uses data and feedback about a case to help it produce the desired outcome.

For instance, a company may show the machine 500 images of a stop sign and 500 images that are not a stop sign.

In this scenario, the stop sign and not a stop sign are the outcome and become the labeled data.

Under the supervision of the labeled data, the machine learns about the relationship of the stop sign. This allows it to classify whether or not an image is a stop sign.

Common supervised learning algorithms include:

- Linear regression
- Support vector machines (SVM)
- **Decision trees**
- **Neural networks**
- **Naive Bayes**
- **Nearest neighbor**
- Gradient boosted trees
- Random forest
- Logistic regression

Unsupervised learning

In unsupervised learning, the machine lacks assistance from the user.

Instead, it finds patterns in data that humans may have missed and discovers unknown results.

Unlike supervised learning, unsupervised learning uses unlabeled data for data points. Through using these data points, the machine makes references to discover meaningful patterns and structures.

Unsupervised learning is data driven and focuses on finding clusters. Some unsupervised learning algorithms include:

- **K-means clustering**
- **Self-organizing maps**
- **Gaussian mixture models**
- **Principal component analysis (PCA)**
- **Association rule**

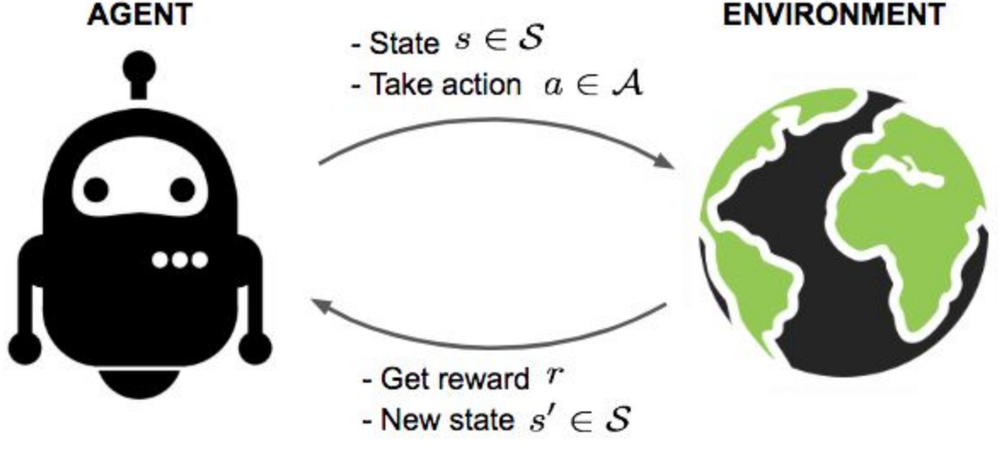
Reinforcement learning uses a trial-and-error method to improve and learn from new situations. To reinforce and maximize favorable actions, it uses a reward system that sends a positive signal for good behavior.

This type of learning is behavior driven.

In order to use reinforcement learning, **you must have an agent and an environment, with the goal being to connect the two using a feedback loop.**

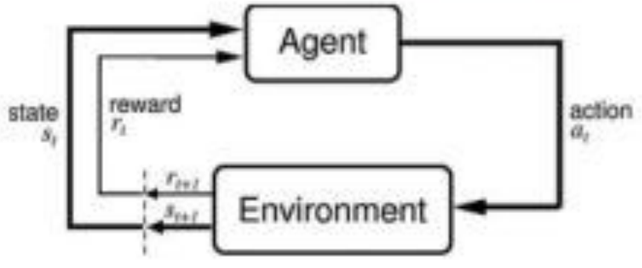
Reinforcement learning algorithms include:

- Q-learning
- Monte-Carlo tree search (MCTS)
- Temporal difference (TD)
- Asynchronous advantage actor-critic (A3C)

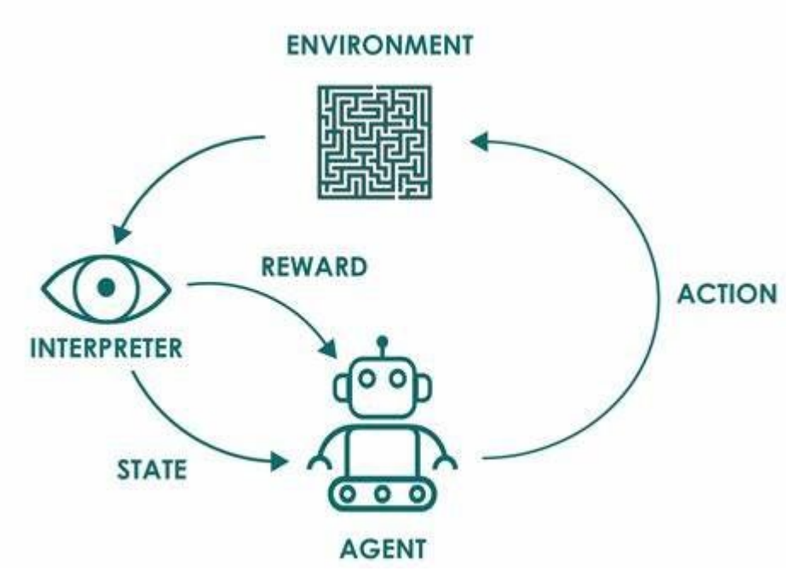


Reinforcement Learning

- Basic idea:
  - Receive feedback in the form of rewards
  - Agent's utility is defined by the reward function
  - Must learn to act so as to maximize expected rewards



This slide deck courtesy of Dan Klein at UC Berkeley



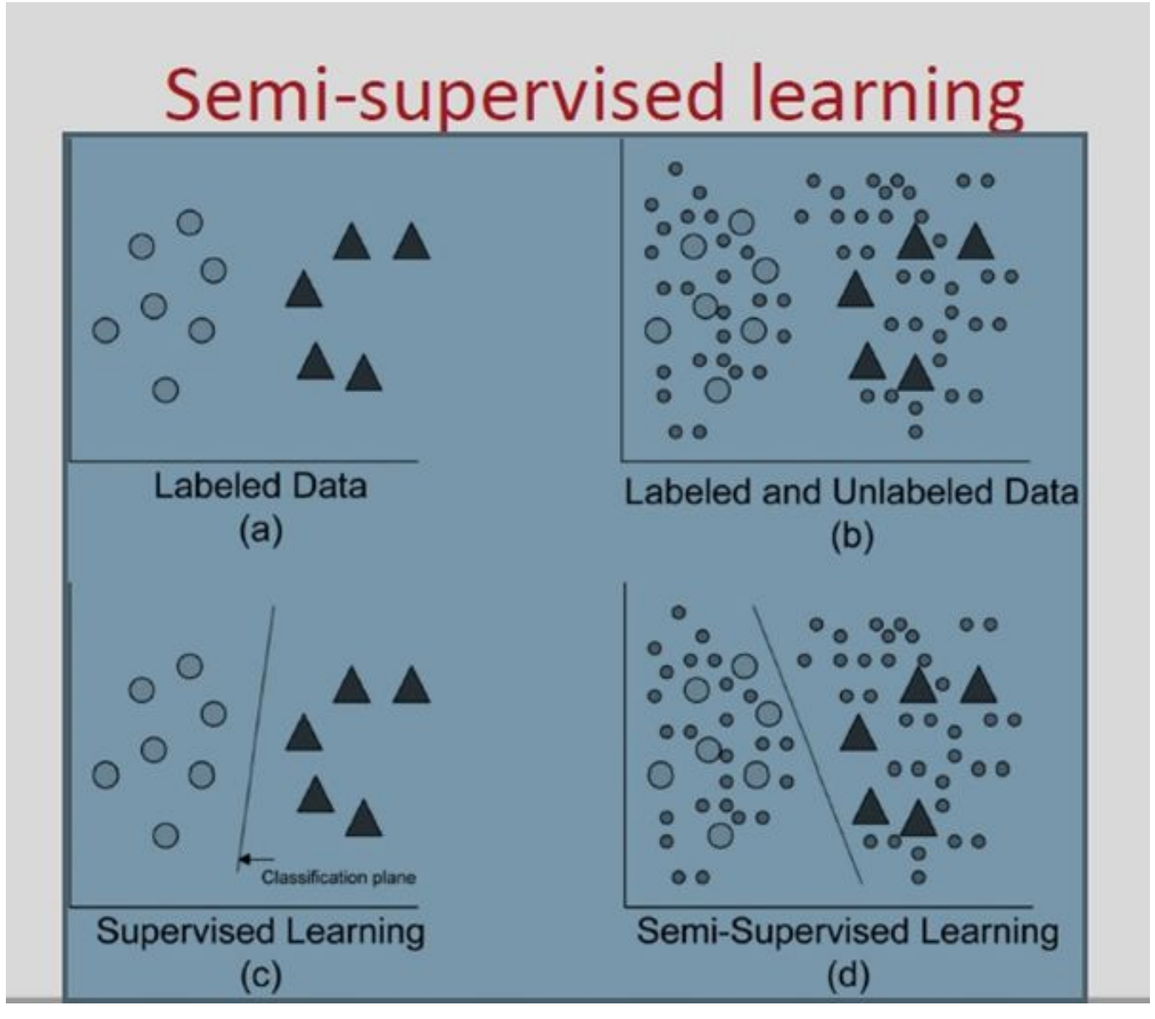
Semi-supervised learning

Semi-supervised learning uses a limited set of labeled data to train itself in shaping the requirements of an operation.

This learning combines a small amount of labeled data with a large volume of unlabeled data, using both supervised and unsupervised learning. It can be a cost-saving method since it involves only using a limited amount of labeled data.

To use this type of learning, train the machine with a small amount of labeled data. You then give it an unlabeled dataset to predict the outputs. These outputs are pseudo labels since they may be inaccurate.

Once you have your pseudo labels, link them with the labeled data. You also link the data inputs from the labeled data with the inputs in the unlabeled data. Finally, train the model with the label data to minimize errors and improve the model's accuracy.


Semi-supervised learning
Labeled Data
(a)
Labeled and Unlabeled Data
(b)
Classification plane
Supervised Learning
(c)
Semi-Supervised Learning
(d)

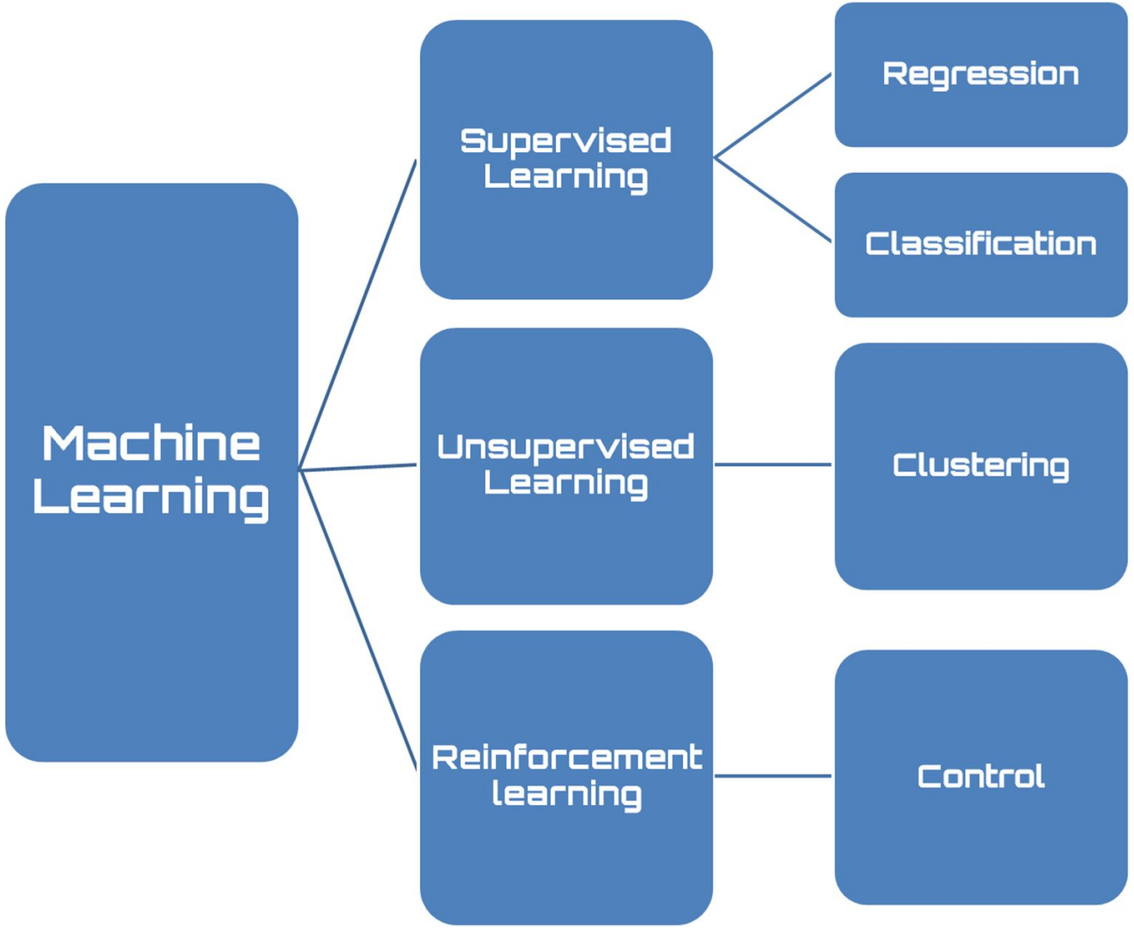
Some semi-supervised learning algorithms include:

- Self-trained Naive Bayes classifier
- Generative adversarial networks (GAN)
- General architecture for text engineering (GATE)

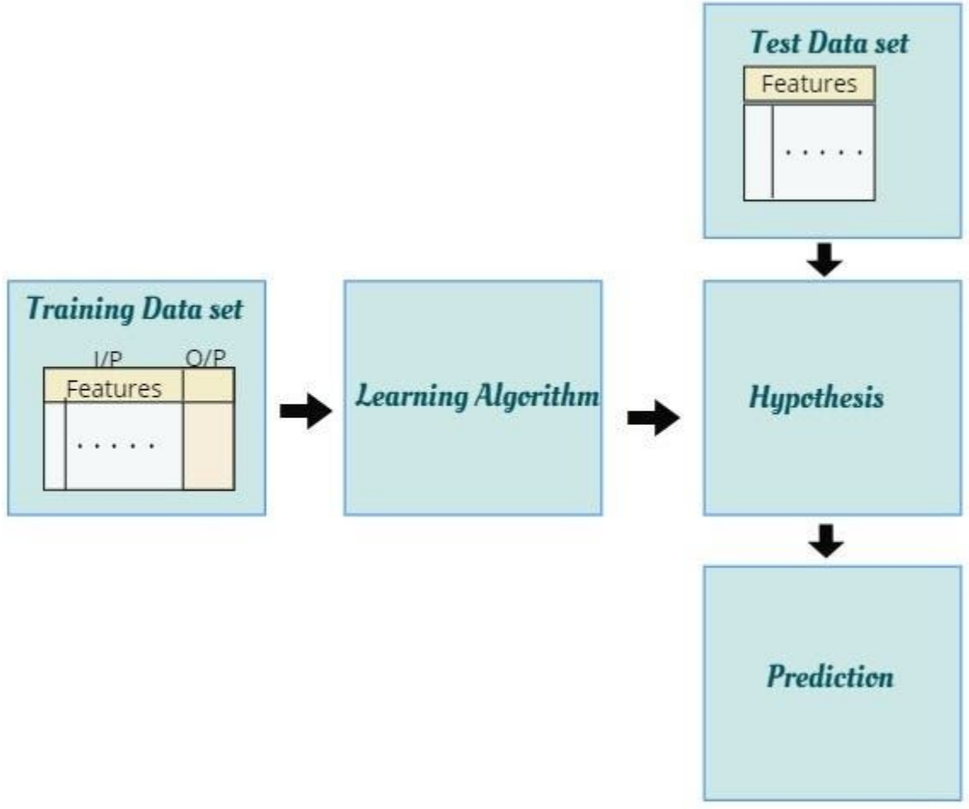
Semi-supervised learning helps businesses complete tasks such as:

- Classifying and locating a large amount of labeled text documents
- Manipulating and labeling audio and videos
- Processing natural language
- Ranking the relevance of web pages on a search engine site

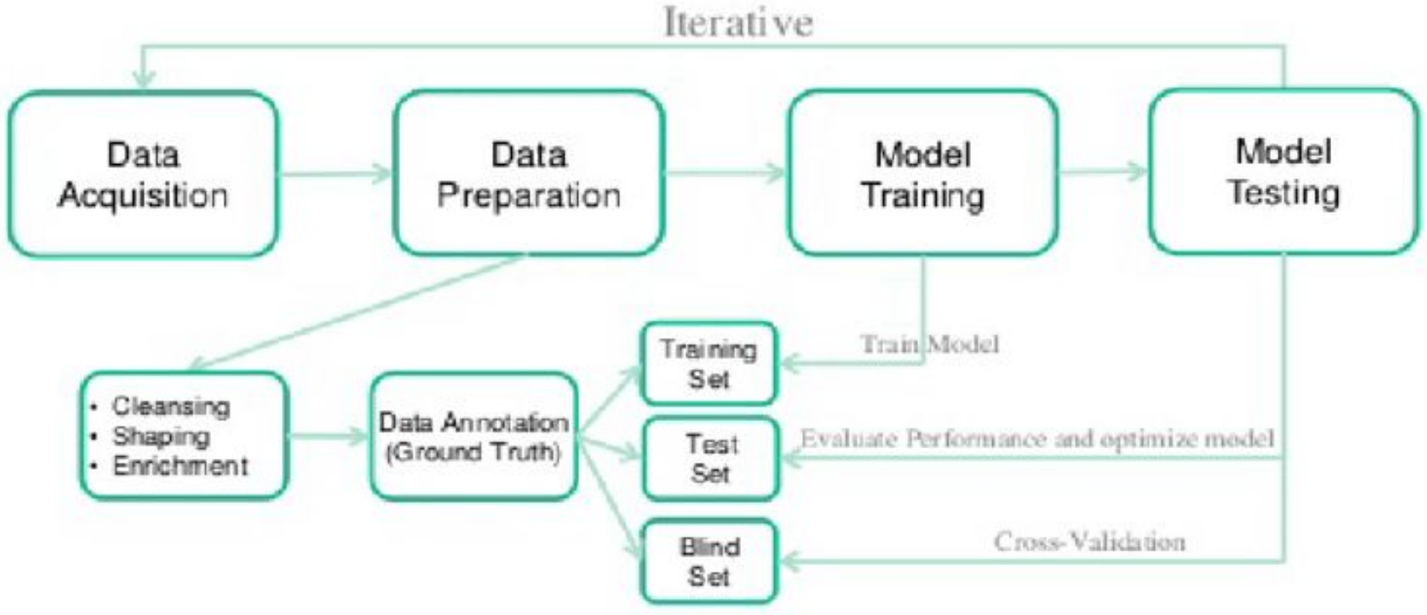
Machine Learning Categories



ML Process



Typical Machine Learning Flow diagram



Summarization of ML

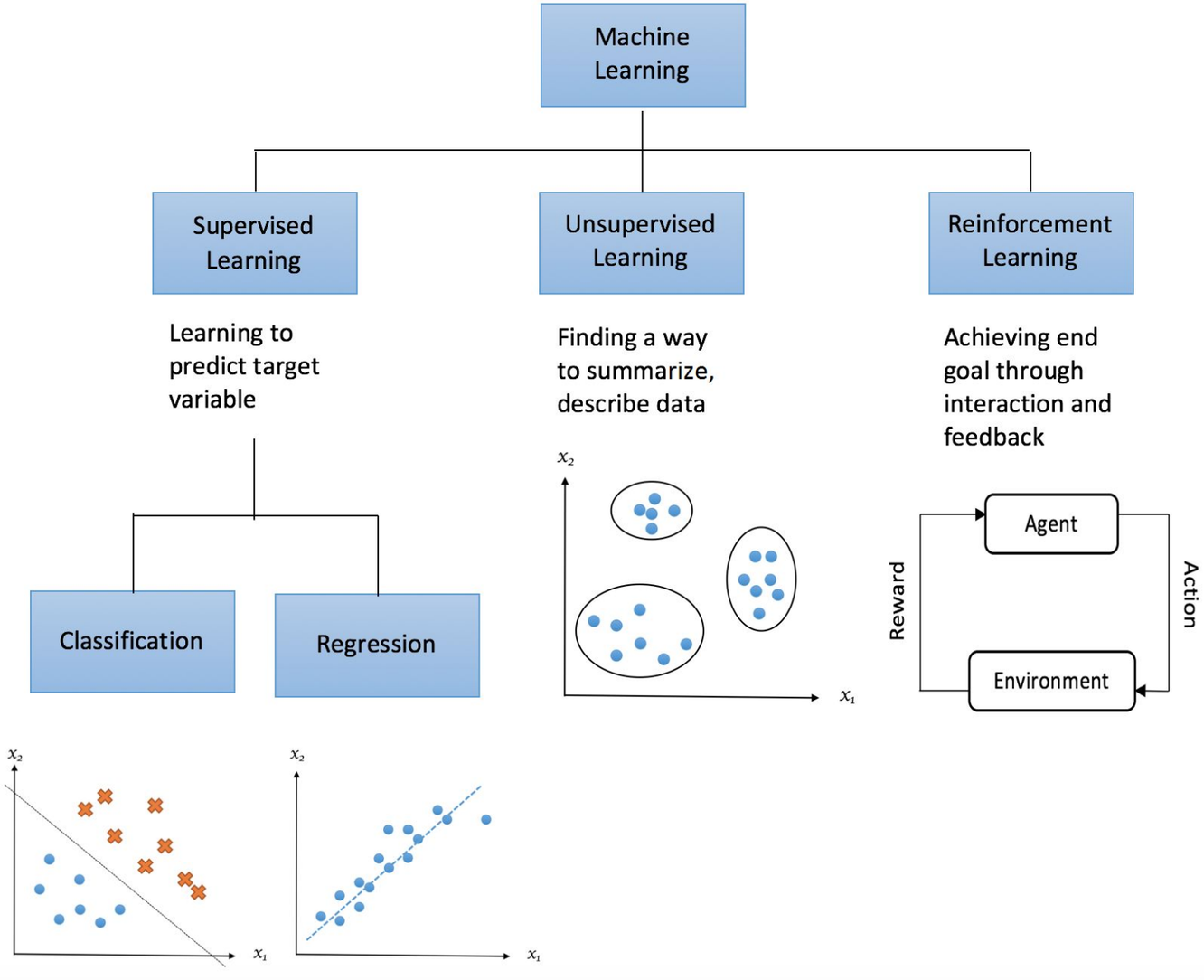
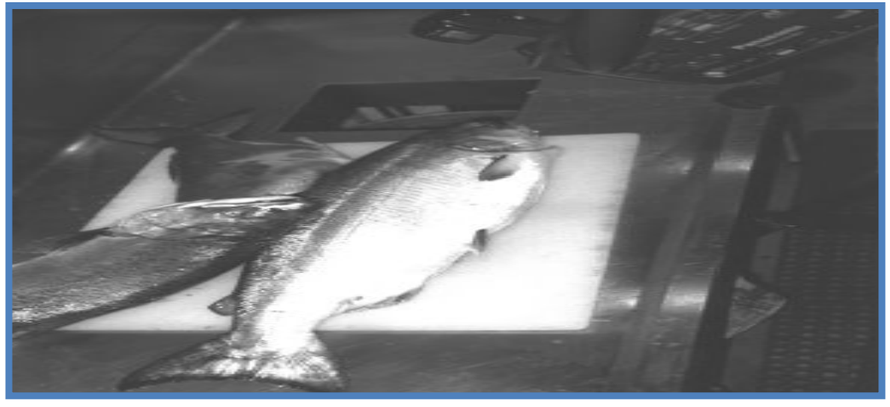
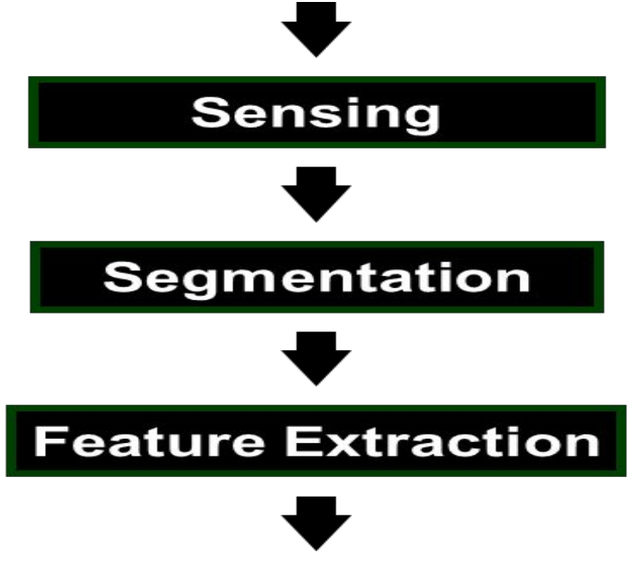


Image Processing Example

- **Sorting Fish:** incoming fish are sorted according to species using optical sensing (sea bass or salmon?)

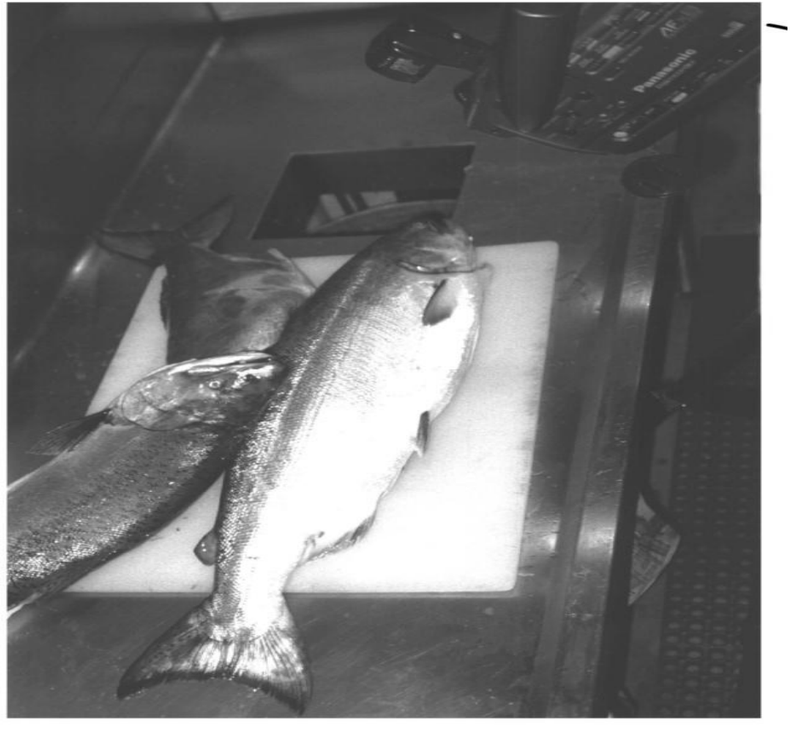
- **Problem Analysis:**
  - set up a camera and take some sample images to extract features
  - Consider features such as length, lightness, width, number and shape of fins, position of mouth, etc.





Preprocessing

A **critical** step for reliable feature extraction!

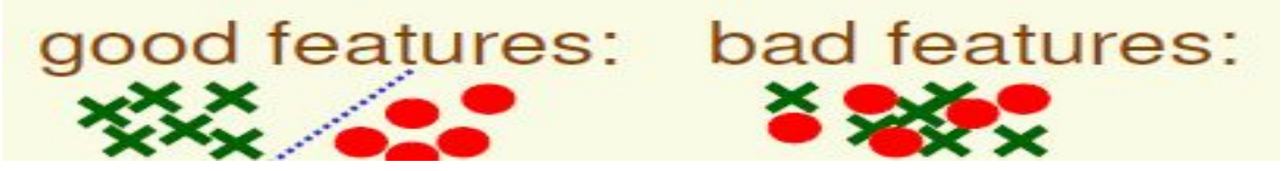


Examples:

- Noise removal
- Image enhancement
- Separate touching or occluding fish
- Extract boundary of each fish

Feature Extraction

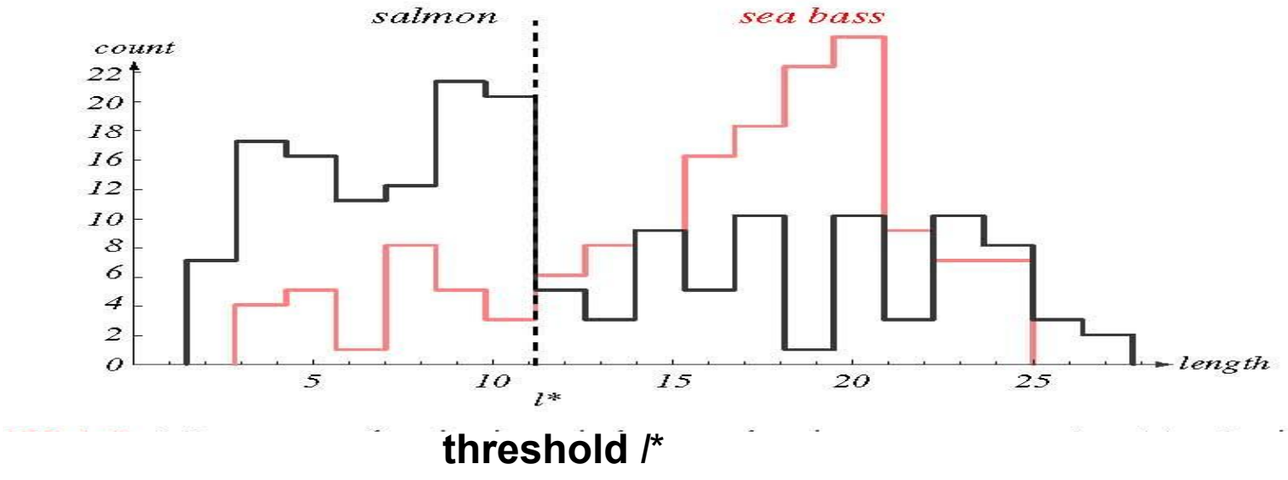
- How to choose a good set of features?
  - Discriminative features



  - Invariant features (e.g., invariant to geometric transformations such as translation, rotation and scale)
- Are there ways to automatically learn which features are best ?

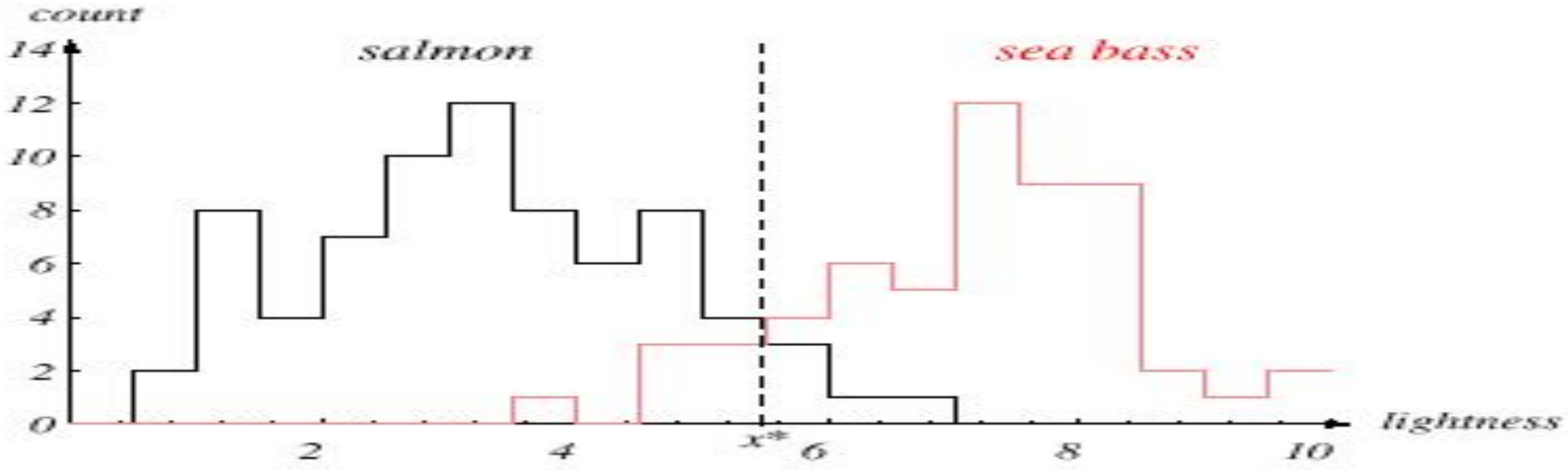
Feature Extraction (cont'd)

Histogram of "length"



- Even though sea bass is longer than salmon on the average, there are many examples of fish where this observation does not hold.

Add Another Feature



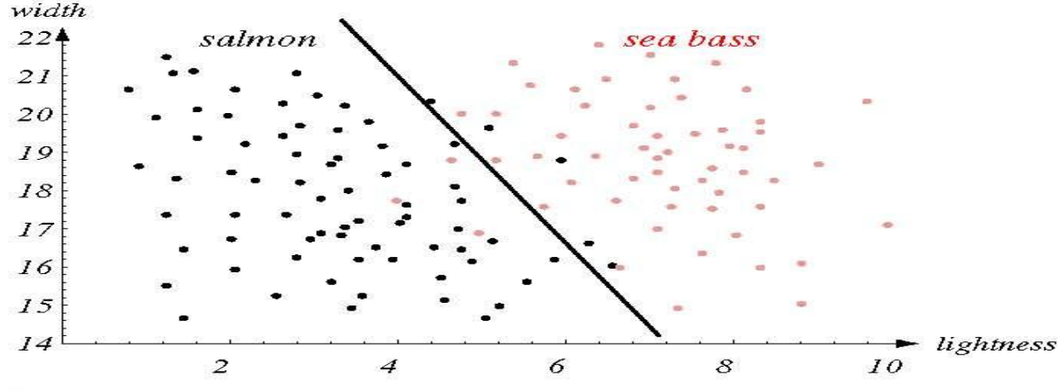
Lightness is a better feature than length because it reduces the misclassification error.
Can we combine features in such a way that we improve performance?
(Hint: correlation)

Multiple Features

- To improve recognition accuracy, we might need to use more than one features.
  - Single features might not yield the best performance.
  - Using combinations of features might yield better performance.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \begin{matrix} x_1 : lightness \\ x_2 : width \end{matrix}$$

Let us consider the scatterplot of lightness Vs Width. A graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present.



Decision region and Decision Boundary

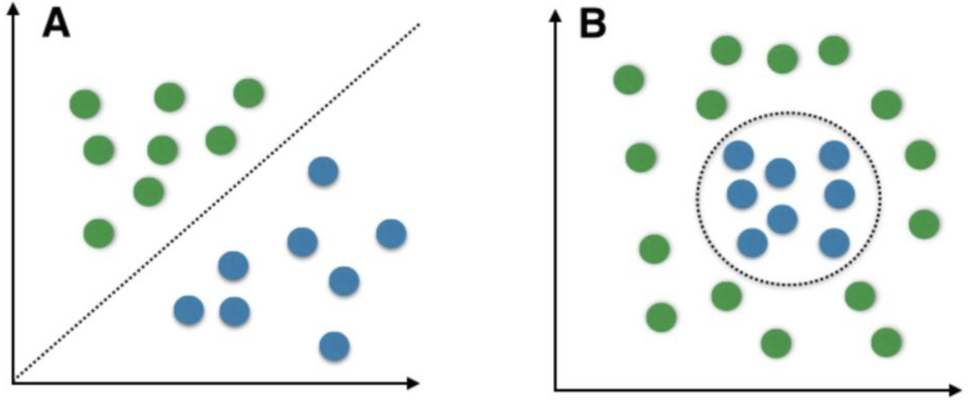
- Our goal of machine learning is to use an optimal decision rule to categorize the incoming data into their respective categories
- The decision boundary separates points belonging to one class from points of other
- The decision boundary partitions the feature space into decision regions.
- The nature of the decision boundary is decided by the discriminant function which is used for decision. It is a function of the feature vector.
- A linear relationship creates a straight line when plotted on a graph, a nonlinear relationship does not create a straight line but instead creates a curve.

Hyper planes and Hyper surfaces

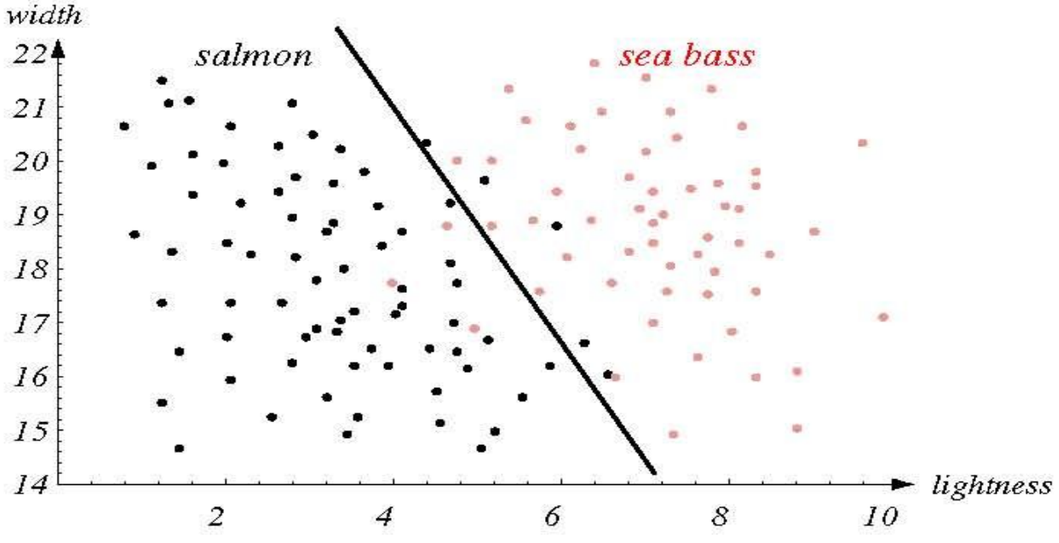
- For two category case, a positive value of discriminant function decides class 1 and a negative value decides the other.
- If the number of dimensions is three. Then the decision boundary will be a plane or a 3-D surface. The decision regions become semi-infinite volumes
- If the number of dimensions increases to more than three, then the decision boundary becomes a hyper-plane or a hyper-surface. The decision regions become semi-infinite hyperspaces.

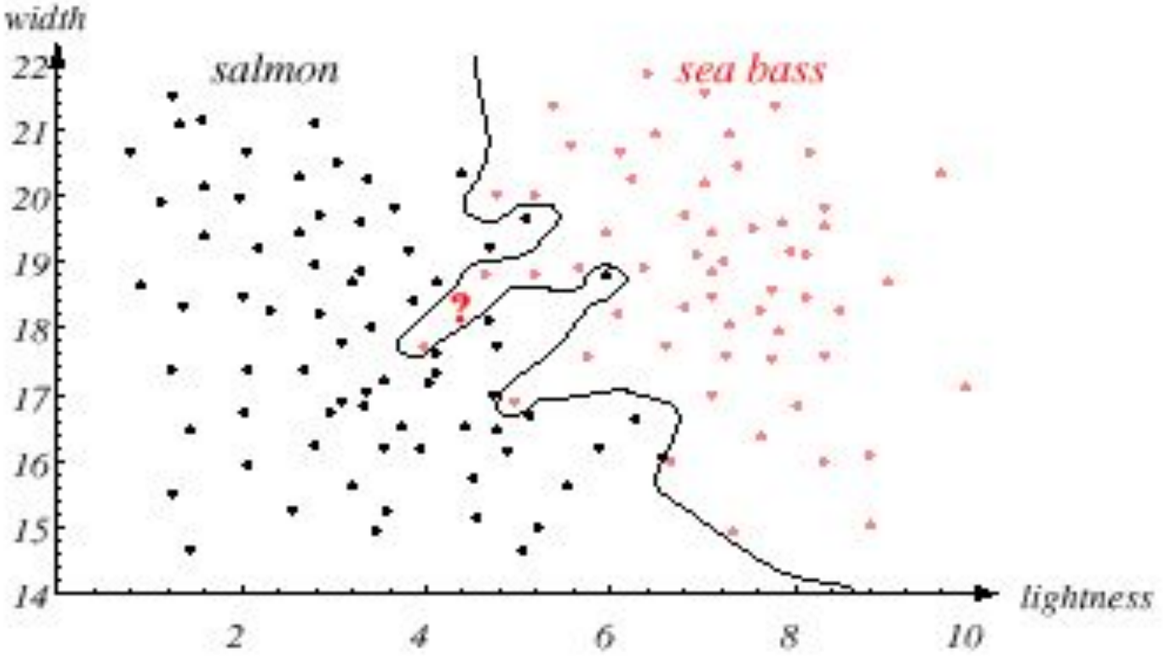
- If the feature space cannot be perfectly separated by a straight line, a more complex boundary might be used. (non-linear)
- Alternatively a simple decision boundary such as straight line might be used even if it did not perfectly separate the classes, provided that the error rates were acceptably low.

Linear Decision **Non-Linear Decision Boundaries**



Decision Boundary for Type1 Vs Type2 Fish





Example for Statistical Decision Theory

- Consider Hypothetical Basket ball Association:
- The prediction could be based on the difference between the home team's average number of points per game (apg) and the visiting team's 'apg' for previous games.
- The training set consists of scores of previously played games, with each home team is classified as winner or loser
- Now the prediction problem is : given a game to be played, predict the home team to be a winner or loser using the feature 'dapg',
- Where dapg = Home team apg – Visiting team apg

Game	*dapg*	Home Team	Game	*dapg*	Home Team
1	1.3	Won	16	-3.1	Won
2	-2.7	Lost	17	1.7	Won
3	-0.5	Won	18	2.8	Won
4	-3.2	Lost	19	4.6	Won
5	2.3	Won	20	3.0	Won
6	5.1	Won	21	0.7	Lost
7	-5.4	Lost	22	10.1	Won
8	8.2	Won	23	2.5	Won
9	-10.8	Lost	24	0.8	Won
10	-0.4	Won	25	-5.0	Lost
11	10.5	Won	26	8.1	Won
12	-1.1	Lost	27	-7.1	Lost
13	2.5	Won	28	2.7	Won
14	-4.2	Won	29	-10.0	Lost
15	-3.4	Lost	30	-6.5	Won

Data set of games showing outcomes, differences between average numbers of points scored and differences between winning percentages for the participating teams in previous games

- The figure shown in the previous slide, lists 30 games and gives the value of dapg for each game and tells whether the home team won or lost.
- Notice that in this data set the team with the **higher apg** usually wins.

- For example in the 9th game the home team on average, scored 10.8 fewer points in previous games than the visiting team, on average and also the home team lost.

- When the teams have about the same apgs, the outcome is less certain. For example, in the 10th game , the home team on average scored 0.4 fewer points than the visiting team, on average, but the home team won the match.

- Similarly 12th game, the home team had an apg 1.1. less than the visiting team on average and the team lost.

Histogram of dapg

- Histogram is a convenient way to describe the data.
- To form a histogram, the data from a single class are grouped into intervals.
- Over each interval rectangle is drawn, with height proportional to number of data points falling in that interval. In the example interval is chosen to have width of two units.
- General observation is that, the prediction is not accurate with single feature ‘dapg’

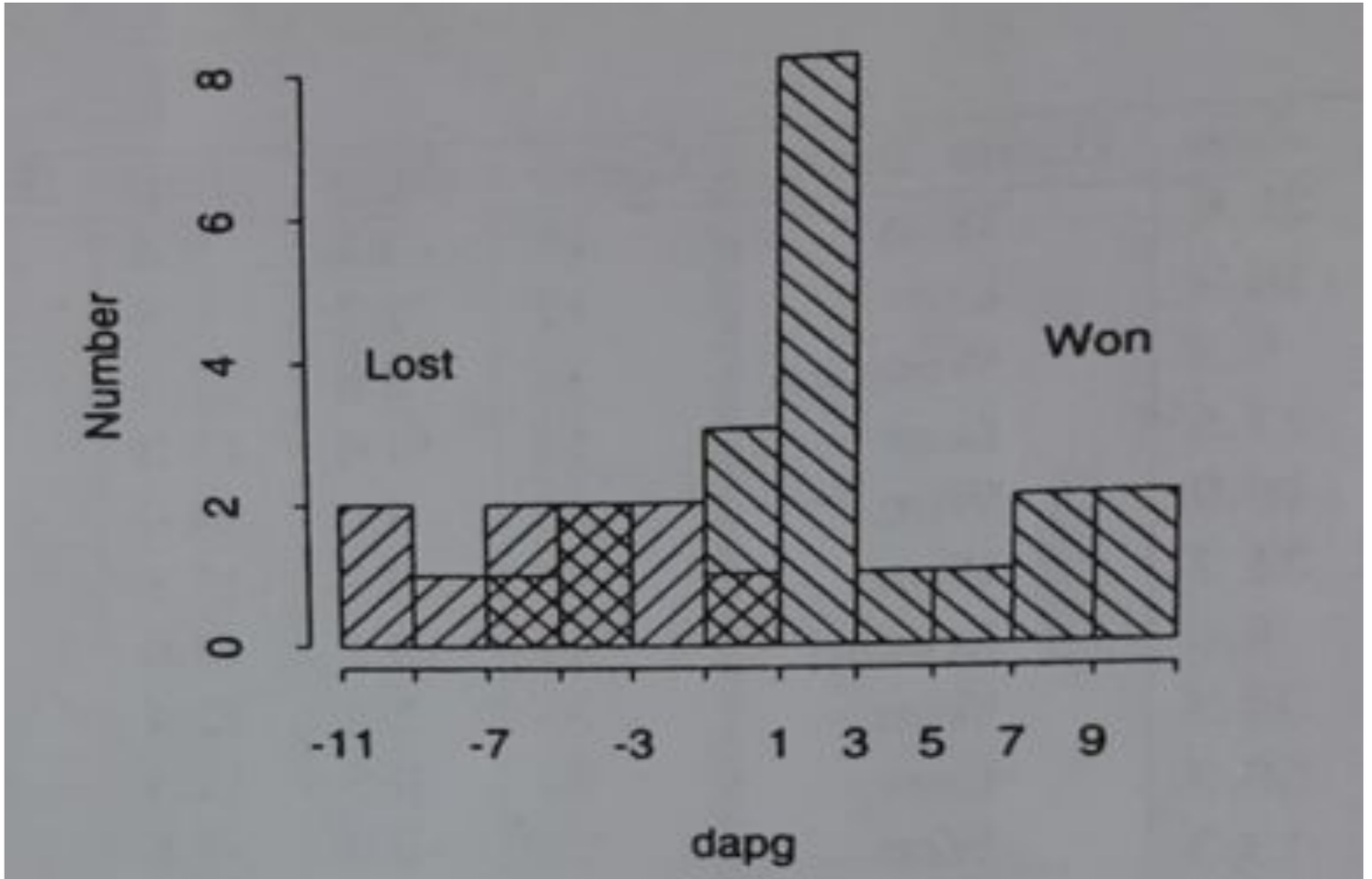


Figure 1.2: Histogram of *dapg*.

Lost

Won

Prediction

- To predict normally a threshold value T is used.
- ‘dapg’ > T consider to be win
- ‘dapg’ < T consider to be lost

- T is called decision boundary or threshold.
- If T=-1, four samples in the original data are misclassified.
  - Here 3 winners are called losers and one loser is called winner.
- If T=0.8, results in no samples from the loser class being misclassified as winner, but 5 samples from the winner class would be misclassified as loser.
- IF T=-6.5, results no samples from the winner class being misclassified as losers, but 7 samples from the loser would be misclassified as winners.
- By inspection, we see that when a decision boundary is used to classify the samples the minimum number of samples that are misclassified is four.
- In the above observations, the minimum number of samples misclassified is 4 when T=0.8

Using Additional Feature dwp

- To make it more accurate let us consider two features.
- Additional features often increases the accuracy of classification.
- Along with ‘dapg’ another feature ‘dwp’ is considered.

- wp= winning percentage of a team in previous games
- dwp = difference in winning percentage between teams
- dwp = Home team wp – visiting team wp

Game	*dapg*	*dwp*	Home Team	Game	*dapg*	*dwp*	Home Team
1	1.3	25.0	Won	16	-3.1	9.4	Won
2	-2.7	-16.9	Lost	17	.1.7	6.8	Won
3	-0.5	5.3	Won	18	2.8	17.0	Won
4	-3.2	-27.5	Lost	19	4.6	13.3	Won
5	2.3	-18.0	Won	20	3.0	-24.0	Won
6	5.1	31.2	Won	21	0.7	-17.8	Lost
7	-5.4	5.8	Lost	22	10.1	44.6	Won
8	8.2	34.3	Won	23	2.5	-22.4	Won
9	-10.8	-56.3	Lost	24	0.8	12.3	Won
10	-0.4	13.3	Won	25	-5.0	-3.8	Lost
11	10.5	16.3	Won	26	8.1	36.0	Won
12	-1.1	-17.6	Lost	27	-7.1	-20.6	Lost
13	2.5	5.7	Won	28	2.7	23.2	Won
14	-4.2	16.0	Won	29	-10.0	-46.9	Lost
15	-3.4	-3.4	Lost	30	-6.5	19.7	Won

Data set of games showing outcomes, differences between average number of points scored and differences between winning percentages for the participating teams in previous games

- Now observe the results on a scatterplot

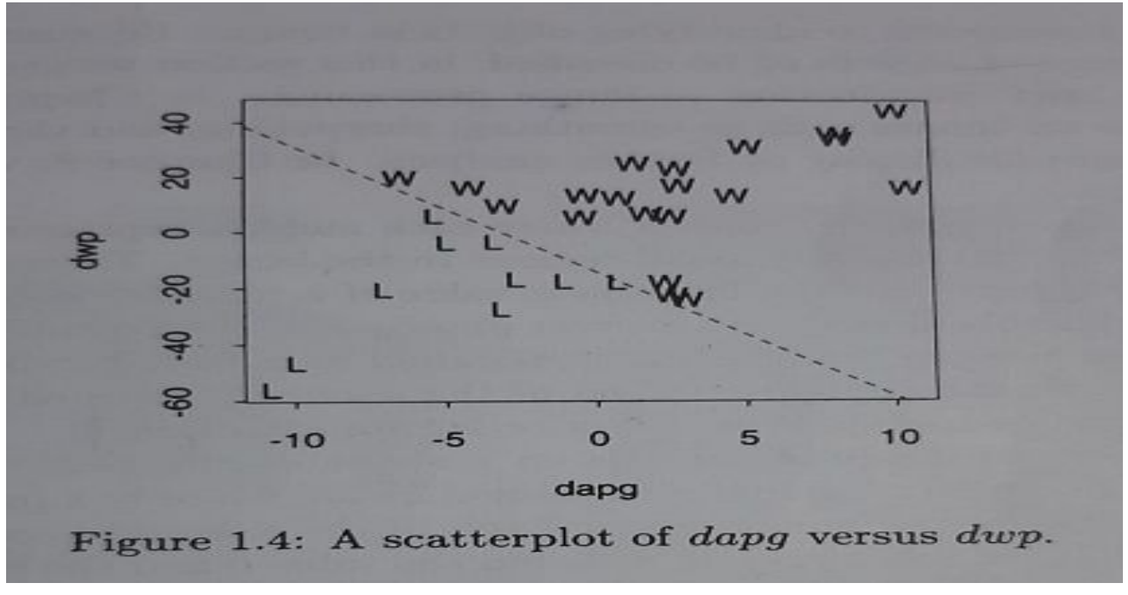
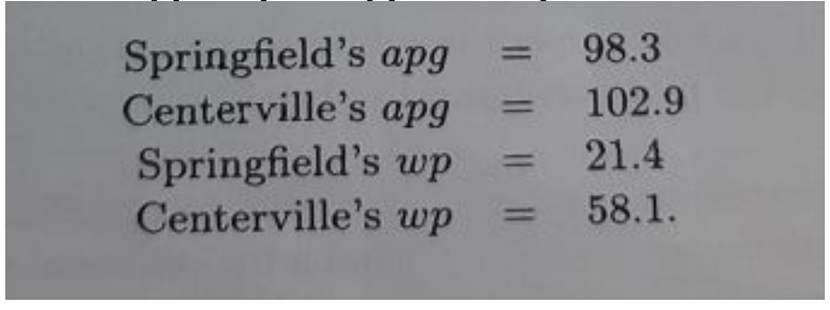


Figure 1.4: A scatterplot of *dapg* versus *dwp*.

- Each sample has a corresponding feature vector (dapg,dwp), which determines its position in the plot.
- Note that the feature space can be classified into two decision regions by a straight line, called a linear decision boundary. (refer line equation). Prediction of this line is logistic regression.
- If the sample lies above the decision boundary, the home team would be classified as the winner and it is below the decision boundary it is classified as loser.

Prediction with two parameters.

- Consider the following : springfield (Home team)



- dapg= home team apg – visiting team apg = 98.3-102.9 = -4.6
- dwp = Home team wp – visiting team wp = -21.4-58.1 = -36.7

- Since the point (dapg, dwp) = (-4.6,-36.7) lies below the decision boundary, we predict that the home team will lose the game.

Basic Concepts of Probability

- An **experiment** is the process by which an observation (or measurement) is obtained.

- An **event** is an outcome of an experiment, usually denoted by a capital letter.

  – The basic element to which probability is applied

  – When an experiment is performed, a particular event either happens, or it doesn't!

What is Probability?

- A number that reflects the chance or likelihood that a particular event will occur
- The extent to which an event is likely to occur.

We measure "how often" an event occurs using

Relative frequency = *f/n*

As *n* gets larger,

Sample → **Population**
And "How often"
= Relative frequency → **Probability**

- **Methods of obtaining probability estimates**
  1. Frequentist approach
  2. Subjective approach
- In the frequentist approach, the probability of an event is estimated by dividing the number of occurrences of an event by the number of trials.
- Frequentist approach is for repeatable events.
- Subjective approach is used for events that are not repeatable. If an estimate reflects a person's opinion or best guess whether an outcome will occur.
- The probability that BJP will win the next election can not be estimated using frequent-list approach as the event is not repeatable

Properties of Probabilities

- **Property 1.** Probabilities are always between 0 and 1
- **Property 2.** The **sample space (S)** for a random variable represents all possible outcomes and must sum to 1 exactly.
- **Property 3.** The probability of the **complement** of an event ("*NOT* the event")= 1 MINUS the probability of the event.
- **Property 4.** Probabilities of **disjoint events** can be added.

Properties of Probabilities In symbols

- **Property 1.** $0 \leq \Pr(A) \leq 1$

- **Property 2.** $\Pr(S) = 1$

- **Property 3.** $\Pr(\bar{A}) = 1 - \Pr(A)$,
  $\bar{A}$ represents the complement of A

- **Property 4.** $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$
  when A and B are disjoint

Random phenomena

- Unable to predict the outcomes, but in the long-run, the outcomes exhibit statistical regularity.

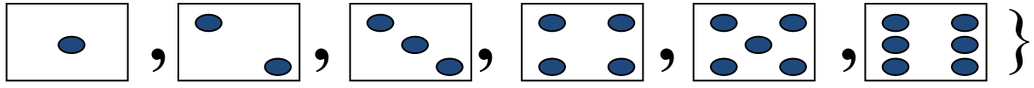
Examples

1. Tossing a coin – outcomes $S = \{\textbf{Head, Tail}\}$

Unable to predict on each toss whether is Head or Tail.

In the long run can predict that 50% of the time heads will occur and 50% of the time tails will occur

2. Rolling a die – outcomes

$S = \{$  $\}$

Unable to predict outcome but in the long run can one can determine that each outcome will occur 1/6 of the time.

Use symmetry. Each side is the same. One side should not occur more frequently than another side in the long run. If the die is not balanced this may not be true.

Basic Concepts of Probability

- Two events are **mutually exclusive** if, when one event occurs, the other cannot, and vice versa.

- **Experiment1: Toss a die**
  - A: observe an odd number
  - B: observe a number greater than 2

  This is not mutually exclusive

  Experiment2: Toss a die
  - C: observe a 6
  - D: observe a 3

  This is mutually exclusive

Basic Concepts of Probability

- An event that cannot be decomposed is called a **simple event.**
- Denoted by E with a subscript.
- Each simple event will be assigned a probability, measuring “how often” it occurs.
- The set of all simple events of an experiment is called the **sample space, S.**

Examples

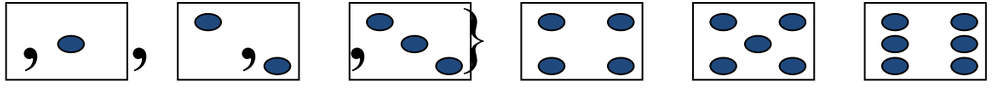
Sample Space of Tossing a coin = {H,T}

Tossing 2 Coins = {HH,HT,TH,TT}

Examples

1. Tossing a coin – outcomes $S = \{$**Head, Tail**$\}$

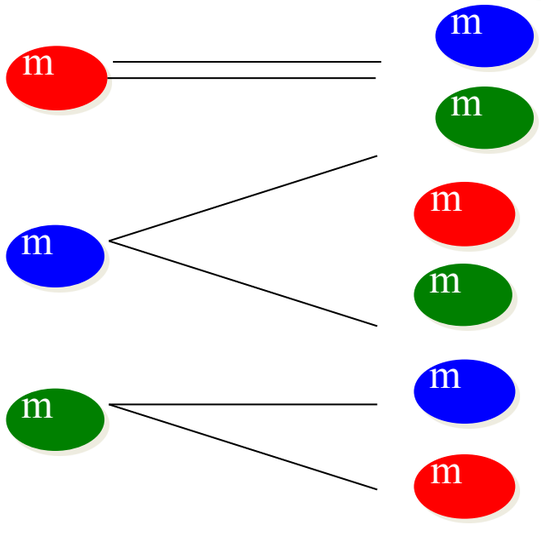
2. Rolling a die – outcomes

$S = \{$,  $\}$

$= \{1, 2, 3, 4, 5, 6\}$

Example 1

A bowl contains three M&Ms®, one red, one blue and one green. A child selects two M&Ms at random. What is the probability that at least one is red?



E_i	$P(E_i)$
RB	1/6
RG	1/6
BR	1/6
BG	1/6
GB	1/6
GR	1/6

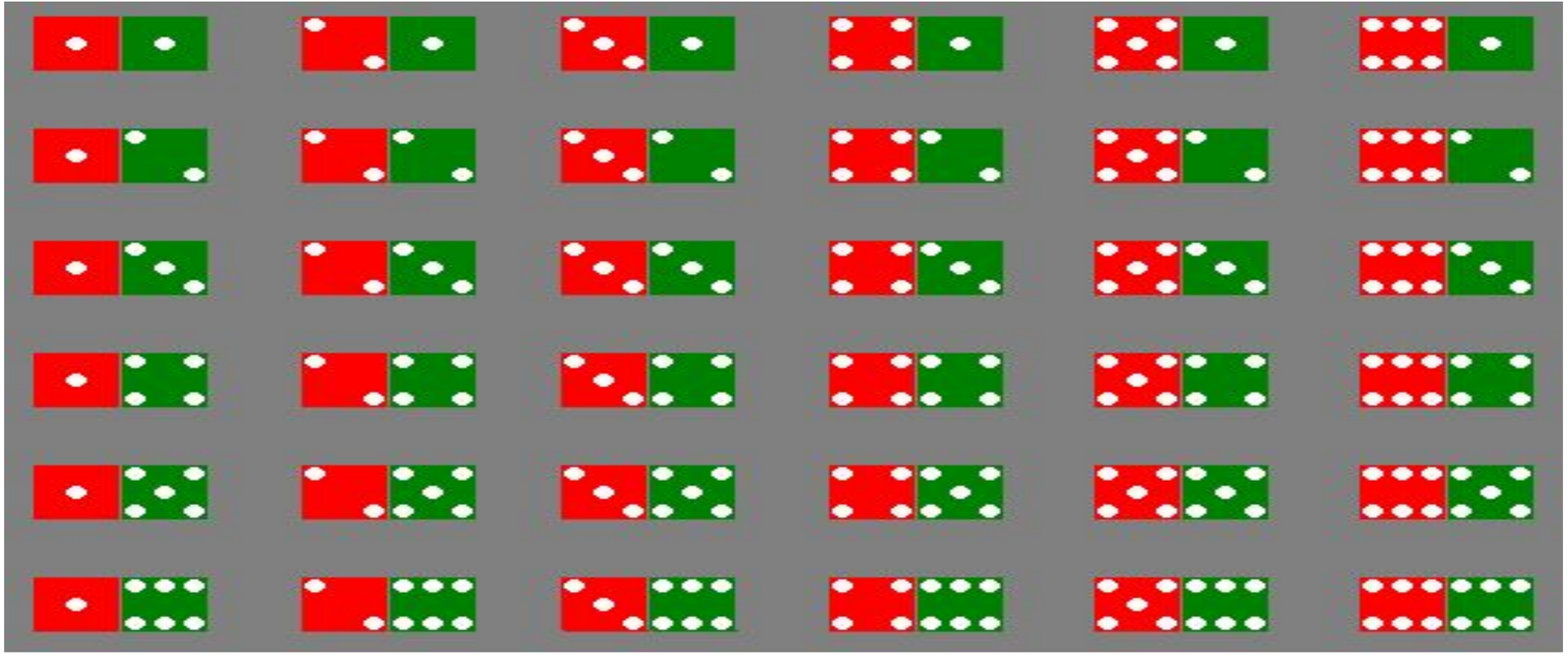
P(at least 1 red)

= P(RB) + P(BR)+ P(RG) + P(GR)

= 4/6 = 2/3

Example 2

The sample space of throwing a pair of dice is



Example 3

Event	Simple events	Probability
Dice add to 3	(1,2),(2,1)	2/36
Dice add to 6	(1,5),(2,4),(3,3), (4,2),(5,1)	5/36
Red die show 1	(1,1),(1,2),(1,3), (1,4),(1,5),(1,6)	6/36
Green die show 1	(1,1),(2,1),(3,1), (4,1),(5,1),(6,1)	6/36

Counting Rules

- Sample space of throwing 3 dice has 216 entries, sample space of throwing 4 dice has 1296 entries, …
- At some point, we have to stop listing and start thinking …
- We need some counting rules

The *mn* Rule

- If an experiment is performed in two stages, with ***m*** ways to accomplish the first stage and ***n*** ways to accomplish the second stage, then there are ***mn*** ways to accomplish the experiment.
- This rule is easily extended to ***k*** stages, with the number of ways equal to

$$n_1 n_2 n_3 \ldots n_k$$

Example: Toss two coins. The total number of simple events is:

$$2 \times 2 = 4$$

Examples

Example: Toss three coins. The total number of simple events is:

$$2 \times 2 \times 2 = 8$$

Example: Toss two dice. The total number of simple events is:

$$6 \times 6 = 36$$

Example: Toss three dice. The total number of simple events is:

$$6 \times 6 \times 6 = 216$$

Example: Two M&Ms are drawn from a dish containing two red and two blue candies. The total number of simple events is:

$$4 \times 3 = 12$$

Permutations

- The number of ways you can arrange ***n*** distinct objects, taking them ***r*** at a time is

$$P_r^n = \frac{n!}{(n-r)!}$$

$$\text{where } n! = n(n-1)(n-2)\ldots(2)(1) \text{ and } 0! \equiv 1.$$

Example: How many 3-digit lock combinations can we make from the numbers 1, 2, 3, and 4?

The order of the choice is important!

$$P_3^4 = \frac{4!}{1!} = 4(3)(2) = 24$$

Examples

Example: A lock consists of five parts and can be assembled in any order. A quality control engineer wants to test each order for efficiency of assembly. How many orders are there?

The order of the choice is important!

$$P_5^5 = \frac{5!}{0!} = 5(4)(3)(2)(1) = 120$$

Combinations

- The number of distinct combinations of ***n*** distinct objects that can be formed, taking them ***r*** at a time is

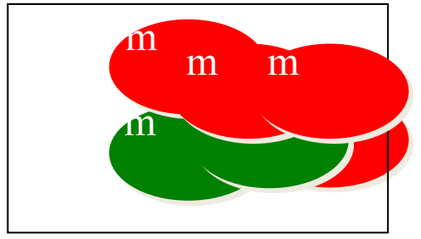
$$C_r^n = \frac{n!}{r!(n-r)!}$$

Example: Three members of a 5-person committee must be chosen to form a subcommittee. How many different subcommittees could be formed?

The order of the choice is not important!

$$C_3^5 = \frac{5!}{3!(5-3)!} = \frac{5(4)(3)(2)1}{3(2)(1)(2)1} = \frac{5(4)}{(2)1} = 10$$

Example



- A box contains six M&Ms®, four red and two green. A child selects two M&Ms at random. What is the probability that exactly one is red?

The order of the choice is not important!

$C_2^6 = \frac{6!}{2!4!} = \frac{6(5)}{2(1)} = 15$ ways to choose 2 M & Ms.

$C_1^2 = \frac{2!}{1!1!} = 2$ ways to choose 1 green M & M.

$C_1^4 = \frac{4!}{1!3!} = 4$ ways to choose 1 red M & M.

4 × 2 =8 ways to choose 1 red and 1 green M&M.

P(exactly one red) = 8/15

A committee of 3 persons is to be constituted from a group of 2 men and 3 women. In how many ways can this be done? How many of these committees would consist of 1 man and 2 women?

Solution: Given, Men = 2, Women = 3

A committee of 3 persons to be constituted.

Here, the order does not matter.

Therefore, we need to count combinations.

There will be as many committees as combinations of 5 different persons taken 3 at a time.

Hence, the required number of ways = $5C3 = 5!/(3!\ 2!) = (5 \times 4 \times 3!)/(3! \times 2) = 10$

Committees with 1 man and 2 women:

1 man can be selected from 2 men in $2C1$ ways. 2 women can be selected from 3 women in $3C2$ ways.

Therefore, the required number of committees = $2C1 \times 3C2 = 2 \times 3C2 = 2 \times 3$

$= 6$

A group consists of 4 girls and 7 boys. In how many ways can a team of 5 members be selected if the team has (i) no girls (ii) at least one boy and one girl (iii) at least three girls

Solution:

Given, Number of girls = 7, Number of boys = 7

(i) No girls

Total number of ways the team can have no girls = $^4C_0 \times {}^7C_5$

$= 1 \times 21$

$= 21$

(ii) at least one boy and one girl

1 boy and 4 girls = $^7C_1 \times {}^4C_4 = 7 \times 1 = 7$

2 boys and 3 girls = $^7C_2 \times {}^4C_3 = 21 \times 4 = 84$

3 boys and 2 girls = $^7C_3 \times {}^4C_2 = 35 \times 6 = 210$

4 boys and 1 girl = $^7C_4 \times {}^4C_1 = 35 \times 4 = 140$

Total number of ways the team can have at least one boy and one girl = 7 + 84 + 210 + 140

= 441

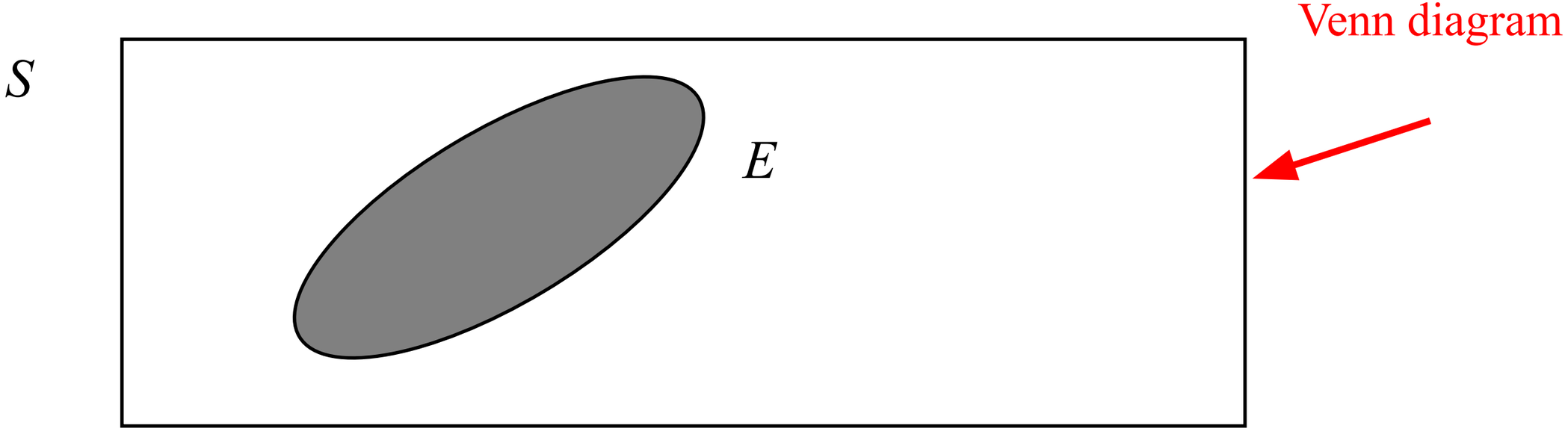
(iii) At least three girls

Total number of ways the team can have at least three girls = $^4C_3 \times {}^7C_2 + {}^4C_4 \times {}^7C_1$

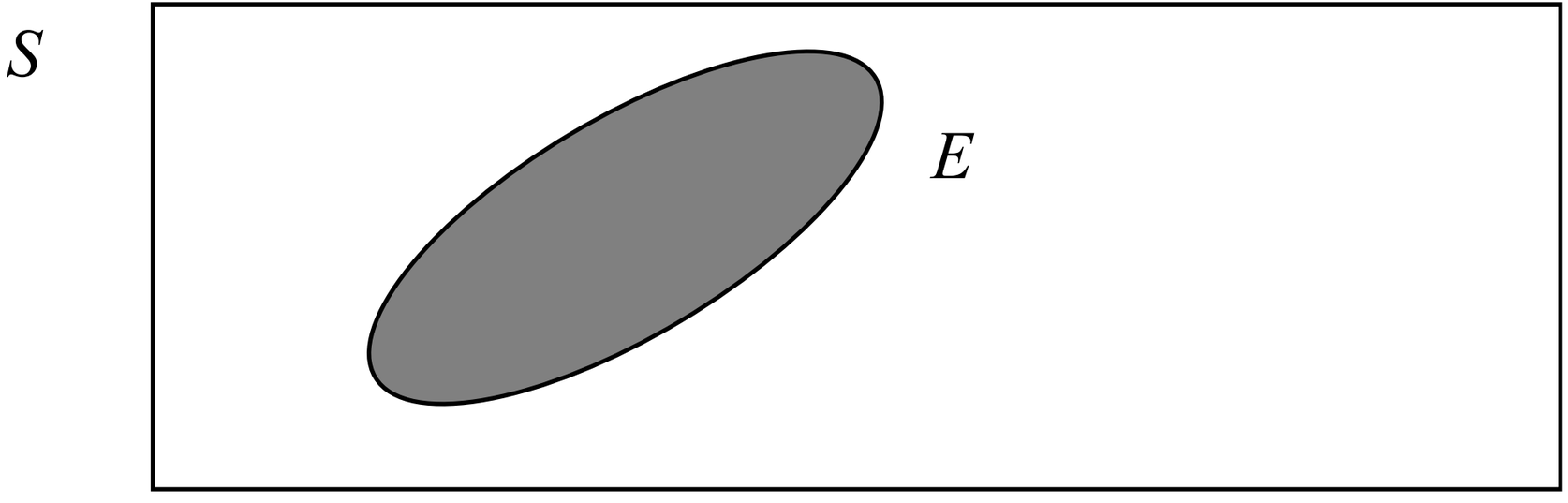
$= 4 \times 21 + 7 = 84 + 7 = 91$

An Event , *E*

The **event**, *E*, is any subset of the **sample space**, *S*. i.e. any set of outcomes (not necessarily all outcomes) of the random phenomena



The **event**, *E*, is said to **have occurred** if after the outcome has been observed the outcome lies in *E.*



Examples

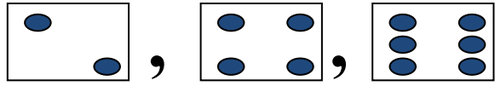
1. Rolling a die – outcomes

$S =$ {  }

$= \{1, 2, 3, 4, 5, 6\}$

$E =$ the event that an even number is rolled

$= \{2, 4, 6\}$

= {  }

Special Events

The Null Event, The empty event - φ

$\varphi = \{\ \}$ = the event that contains no outcomes

The Entire Event, The Sample Space - S

S = the event that contains all outcomes

The empty event, φ , never occurs.

The entire event, S, always occurs.

In problems you will recognize that you are working with:

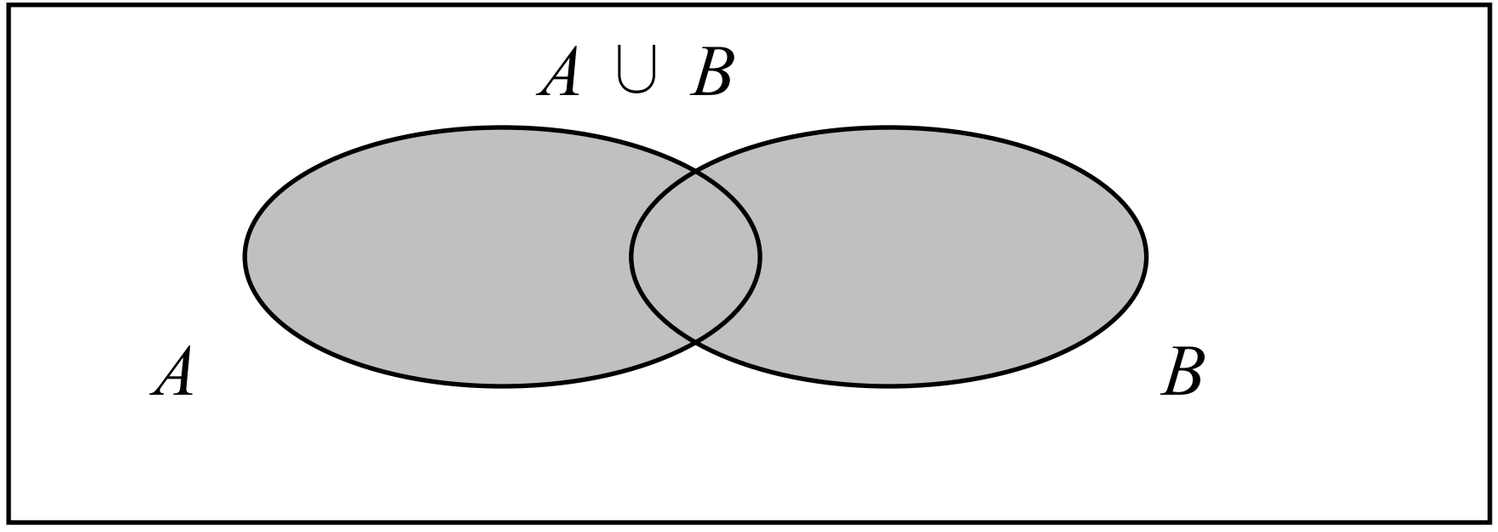
1. **Union** if you see the word **or**,
2. **Intersection** if you see the word **and**,
3. **Complement** if you see the word **not**.

Set operations on Events

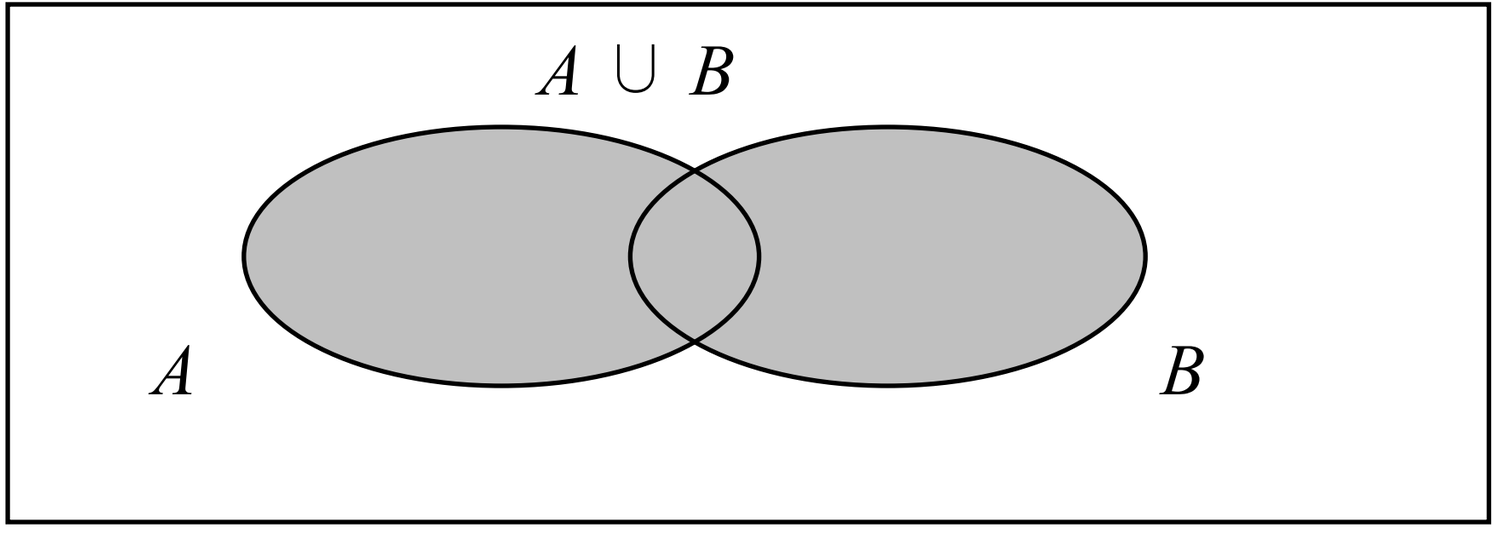
Union

Let A and B be two events, then the **union** of A and B is the event (denoted by $A \cup B$) defined by:

$$A \cup B = \{e |\ e \text{ belongs to } A \textbf{ or } e \text{ belongs to } B\}$$



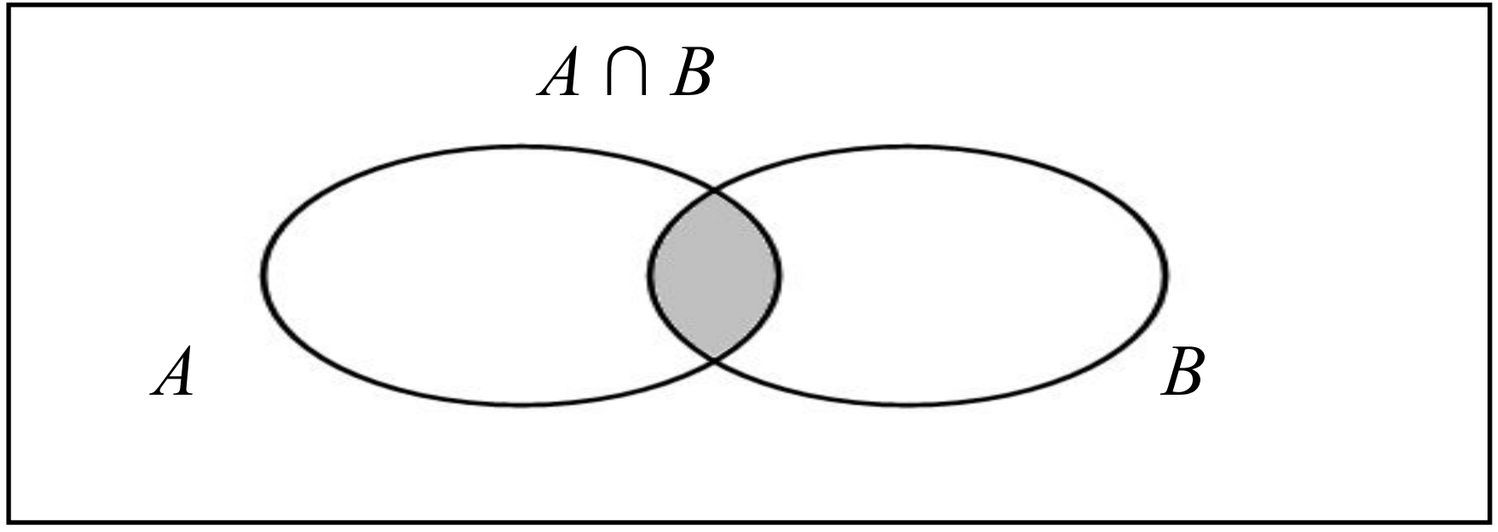
The event $A \cup B$ **occurs** if the event A **occurs or** the event and B **occurs** .



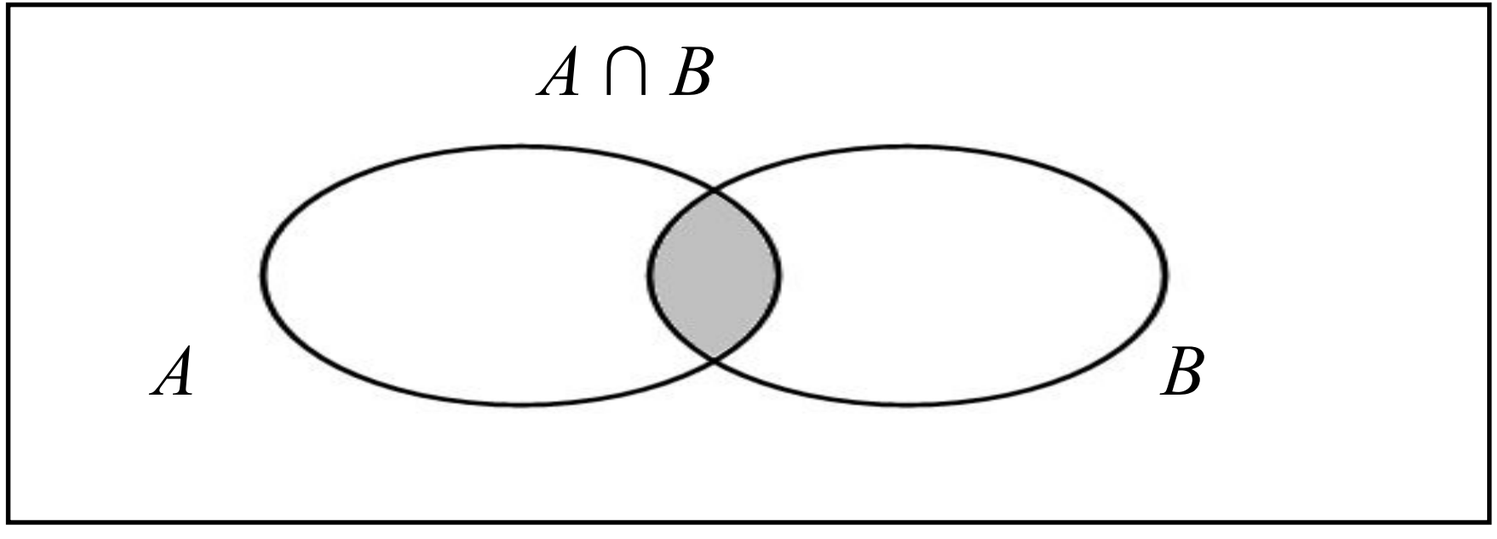
Intersection

Let A and B be two events, then the **intersection** of A and B is the event (denoted by $A \cap B$) defined by:

$$A \cap B = \{e |\ e \text{ belongs to } A \textbf{ and } e \text{ belongs to } B\}$$



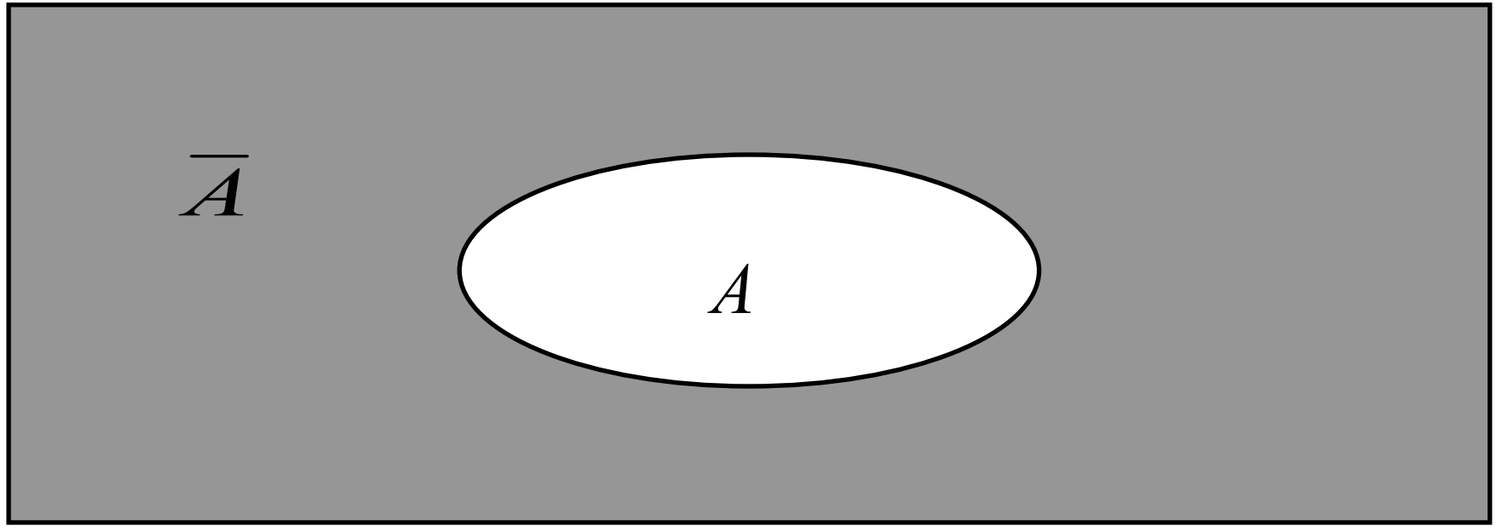
The event $A \cap B$ **occurs** if the event A **occurs and** the event and B **occurs** .



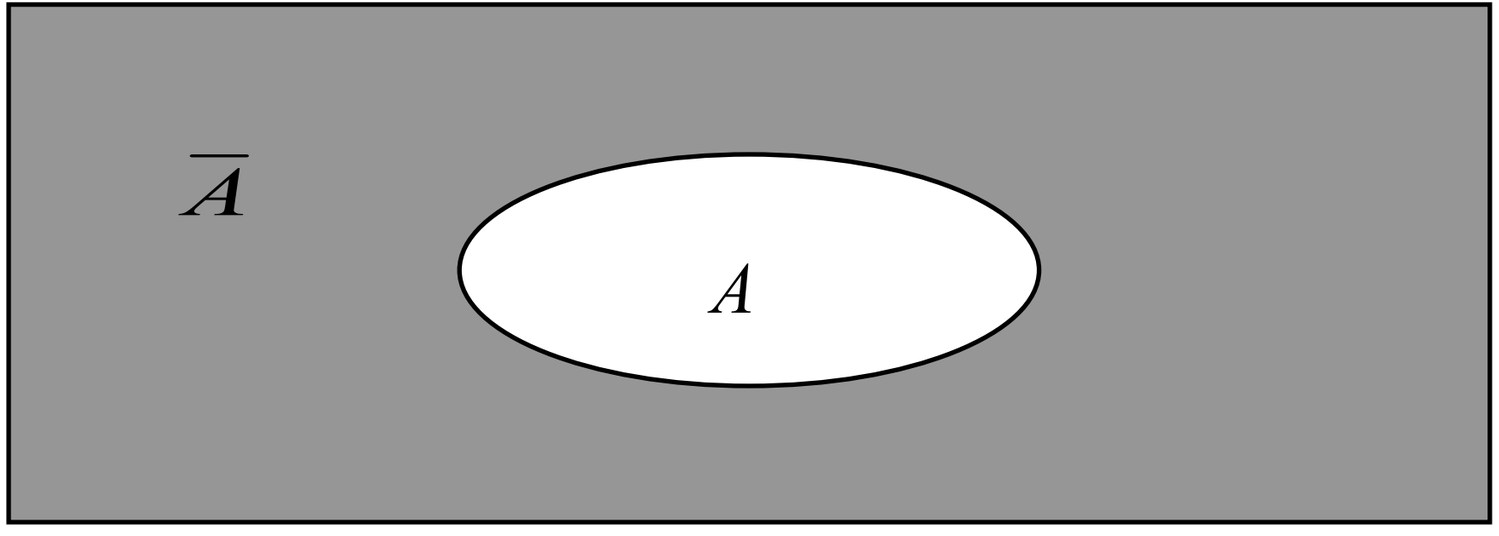
Complement

Let A be any event, then the **complement** of A (denoted by $\overline{A}$) defined by:

$\overline{A}$= {e| e **does not** belongs to A}



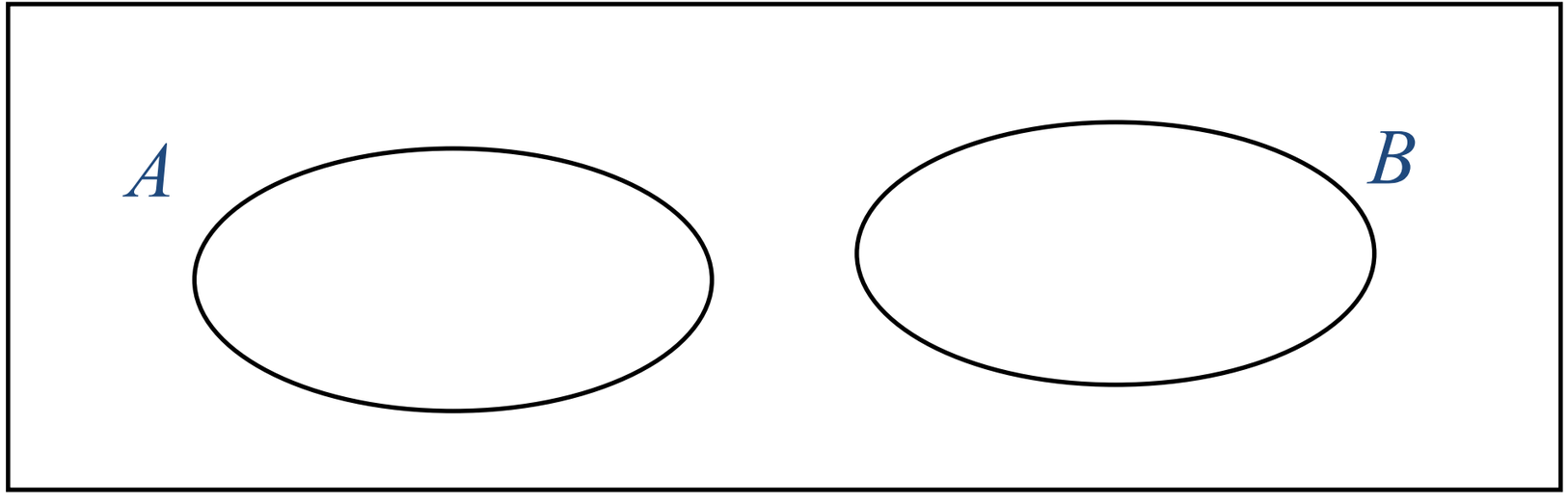
The event $\overline{A}$ **occurs** if the event A **does not occur**



Definition: mutually exclusive

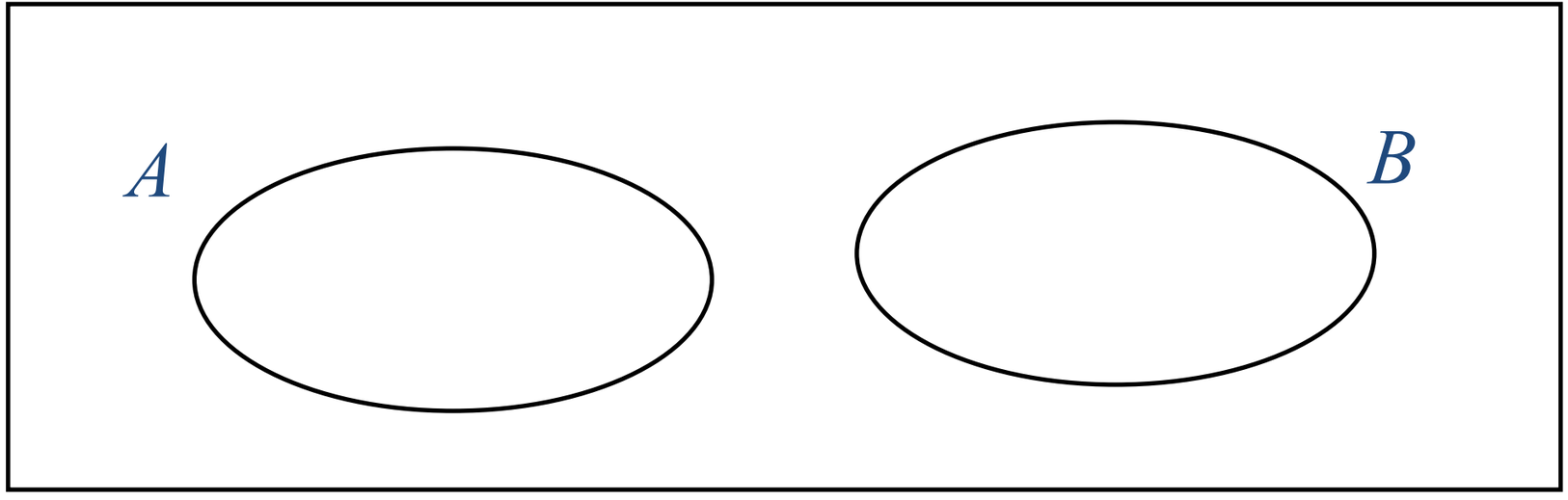
Two events A and B are called **mutually exclusive** if:

$$A \cap B = \phi$$



If two events A and B are **mutually exclusive** then:

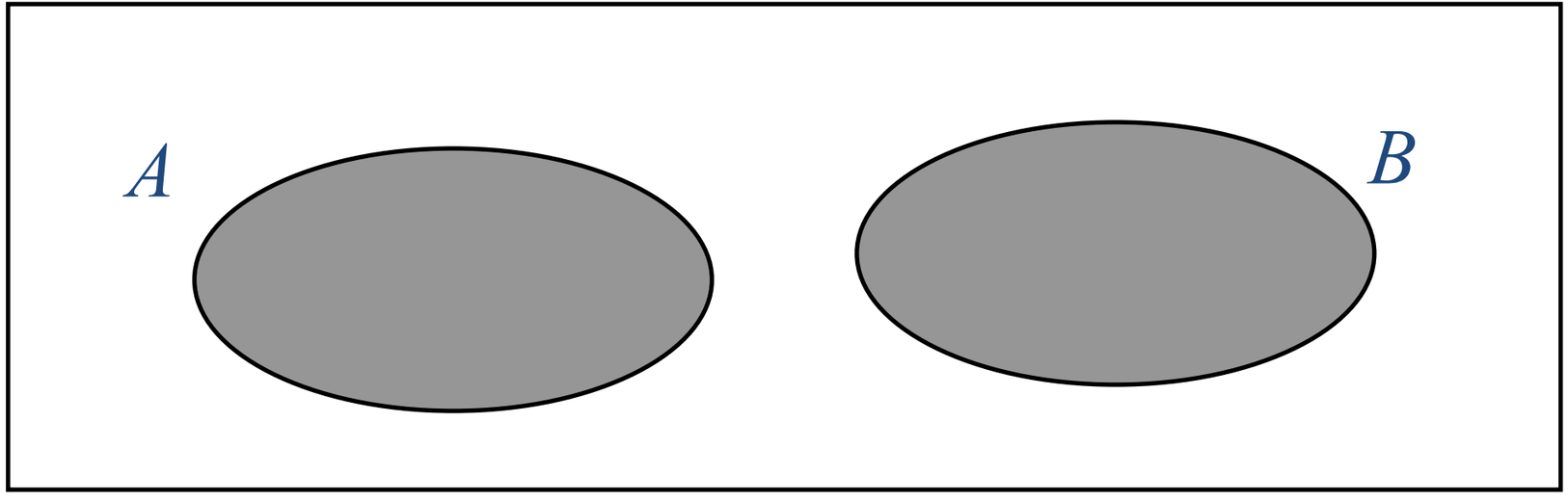
1. They have no outcomes in common.
   They can't occur at the same time. The outcome of the random experiment can not belong to both A and B.



$$P[A \cup B] = P[A] + P[B]$$

i.e.

$$P[A \text{ or } B] = P[A] + P[B]$$



Rule The additive rule
(Mutually exclusive events)

$$P[A \cup B] = P[A] + P[B]$$

i.e.

$$P[A \text{ or } B] = P[A] + P[B]$$

if $A \cap B = \varphi$

(A and B mutually exclusive)

Rule The additive rule

(In general)

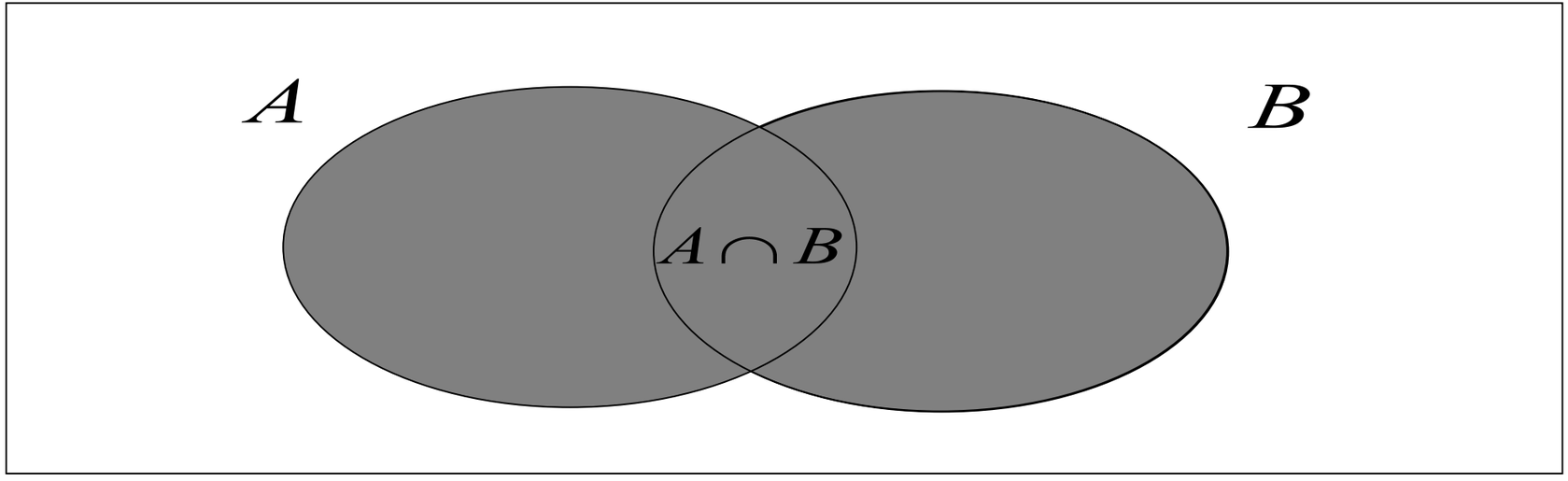
$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

or

$$P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$$

Logic

$A \cup B$



When $P[A]$ is added to $P[B]$ the outcome in $A \cap B$ are counted twice

hence

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

Example:

Bangalore and Mohali are two of the cities competing for the National university games. (There are also many others). The organizers are narrowing the competition to the **final 5 cities.**

There is a 20% chance that Bangalore will be amongst the **final 5**. There is a 35% chance that Mohali will be amongst the **final 5** and an 8% chance that both Bangalore and Mohali will be amongst the **final 5**. What is the probability that Bangalore or Mohali will be amongst the **final 5.**

Solution:

Let A = the event that Bangalore is amongst the **final 5**.

Let B = the event that Mohali is amongst the **final 5**.

Given $P[A] = 0.20$, $P[B] = 0.35$, and $P[A \cap B] = 0.08$

What is $P[A \cup B]$?

Note: "and" ≡ ∩**,** "or" ≡ ∪ .

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$
$$= 0.20 + 0.35 - 0.08 = 0.47$$

Find the probability of drawing an ace or a spade from a deck of cards.
There are 52 cards in a deck; 13 are spades, 4 are aces. Probability of a single cord being spade is: 13/52.
Probability of drawing an Ace is : 4/52.
Probability of a single card being both Spade and Ace : 1/52
Let A = Event of drawing a spade .
Let B = Event drawing Ace.
Given $P[A] = 1/4$, $P[B] = 1/13$, and $P[A \cap B] = 1/52$

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

$$P[A \cup B] = 1/4 + 1/13 - 1/52$$

Rule for complements

2. $$P\left[\overline{A}\right] = 1 - P[A]$$

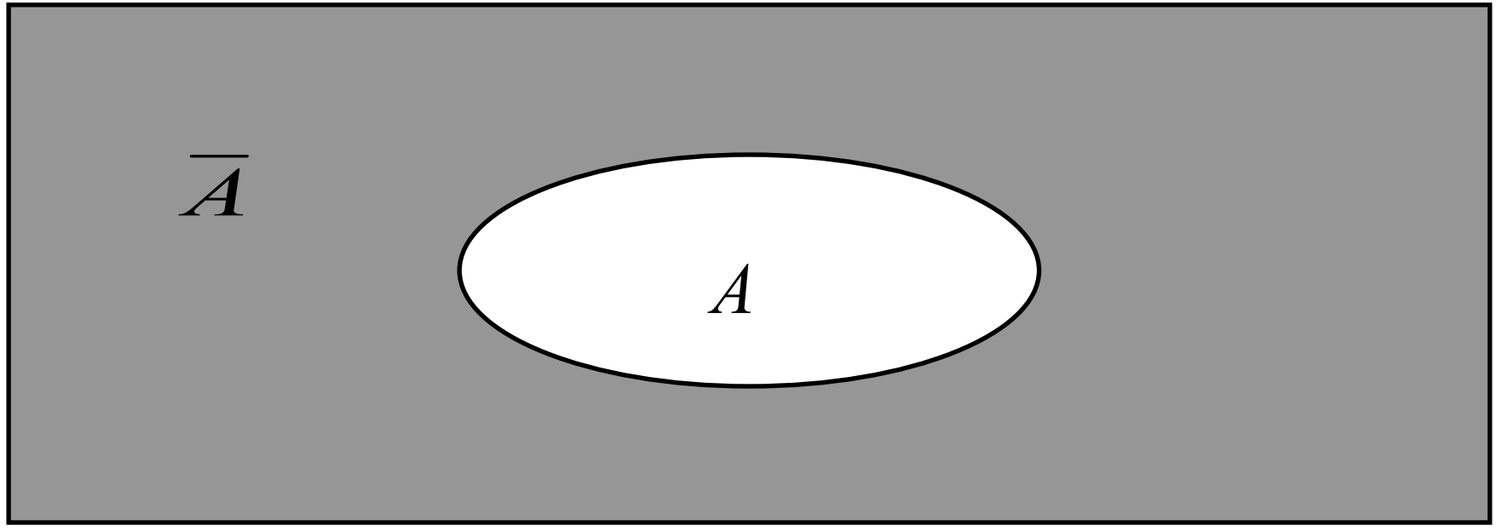
or

$$P[\text{not } A] = 1 - P[A]$$

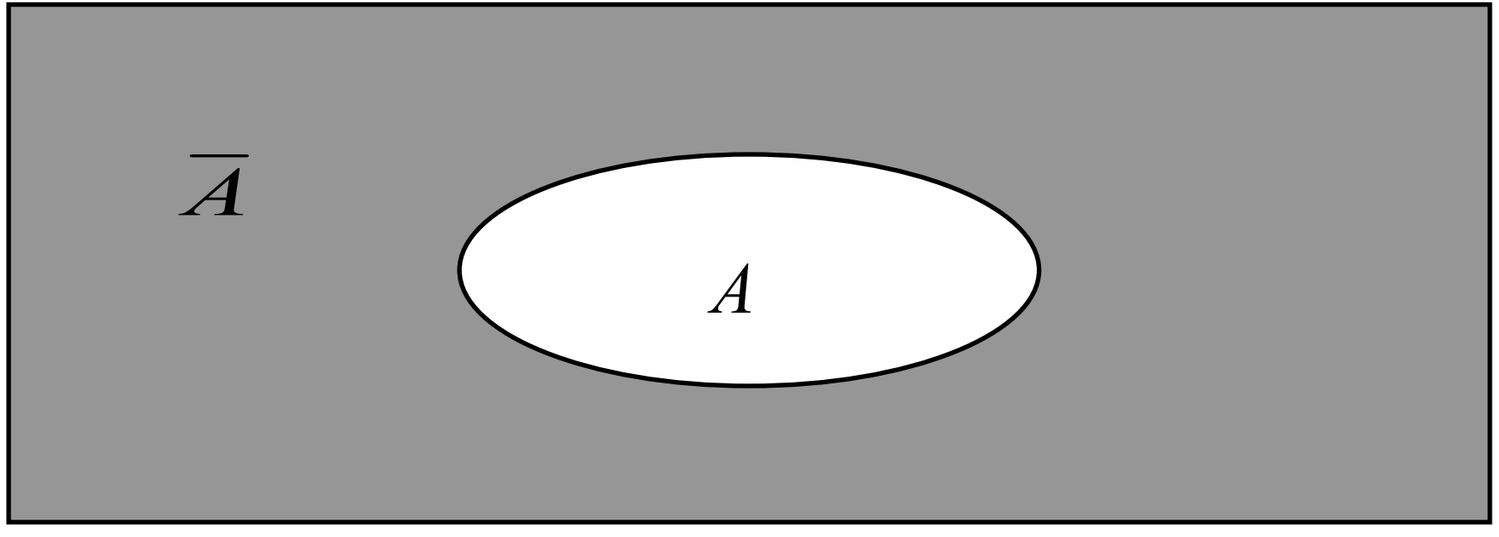
Complement

Let A be any event, then the **complement** of A (denoted by $\overline{A}$) defined by:

$\overline{A}$= {e| e **does not** belongs to A}



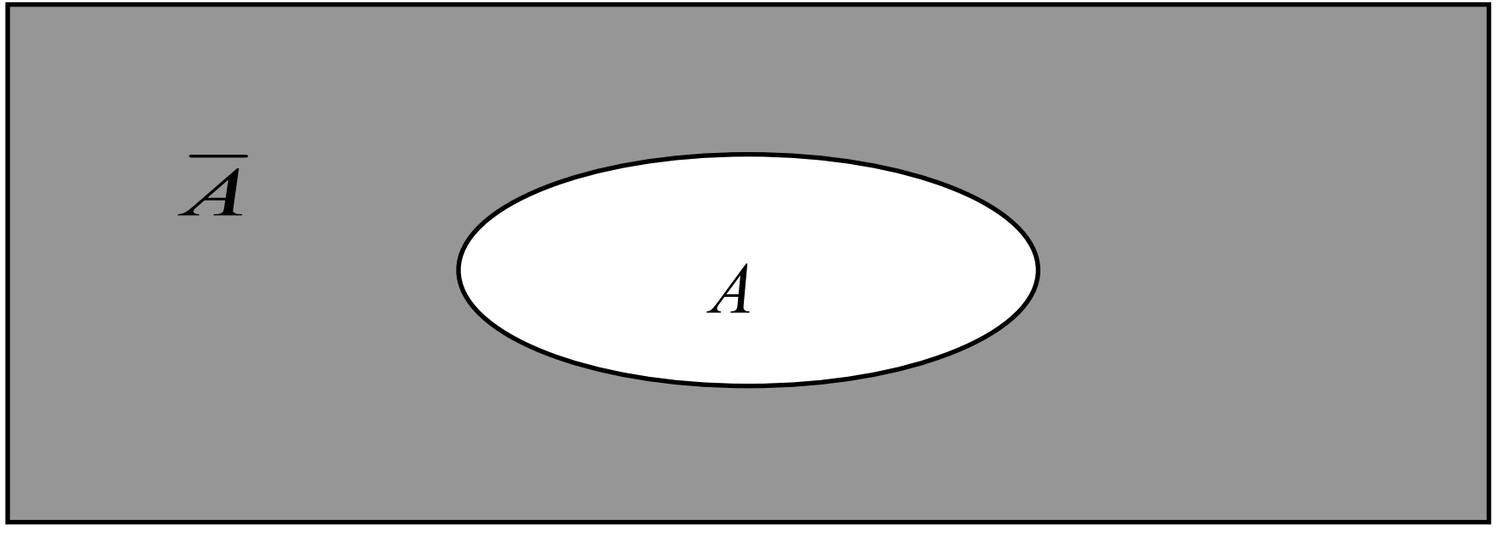
The event $\overline{A}$ **occurs** if the event A **does not occur**



Logic:

$\overline{A}$ and A are **mutually exclusive.**

and $S = A \cup \overline{A}$



thus $1 = P[S] = P[A] + P[\overline{A}]$

and $P[\overline{A}] = 1 - P[A]$

Conditional Probability

- Frequently before observing the outcome of a random experiment you are given information regarding the outcome
- How should this information be used in prediction of the outcome.
- Namely, how should probabilities be adjusted to take into account this information
- Usually the information is given in the following form: You are told that the outcome belongs to a given event. (i.e. you are told that a certain event has occurred)

Definition

Suppose that we are interested in computing the probability of event *A* and we have been told event *B* has occurred.

Then the conditional probability of *A* given *B* is defined to be:

$$P[A|B] = \frac{P[A \cap B]}{P[B]} \qquad \text{if} \quad P[B] \neq 0$$

Similarly, P[B|A] = $P[A \cap B]$ /P[A]

- From the previous two expressions

$P[A \cap B] = \mathrm{P[B].P[A|B]}$

And $P[A \cap B] = \mathrm{P[A].P[B|A]}$

Can also be used to calculate $P[A \cap B]$

The Multiplication Rule

- In many of the cases, P(A) may not depend on whether B has occurred. We say that the event A is independent of B if P(A) = P(A|B). An important consequence of the definition of independence is multiplication rule, which is obtained by substituting P(A) for P(A|B) in the above expressions
- $P[A \cap B] = \mathrm{P[A].P[B]}$ whenever A is independent of B

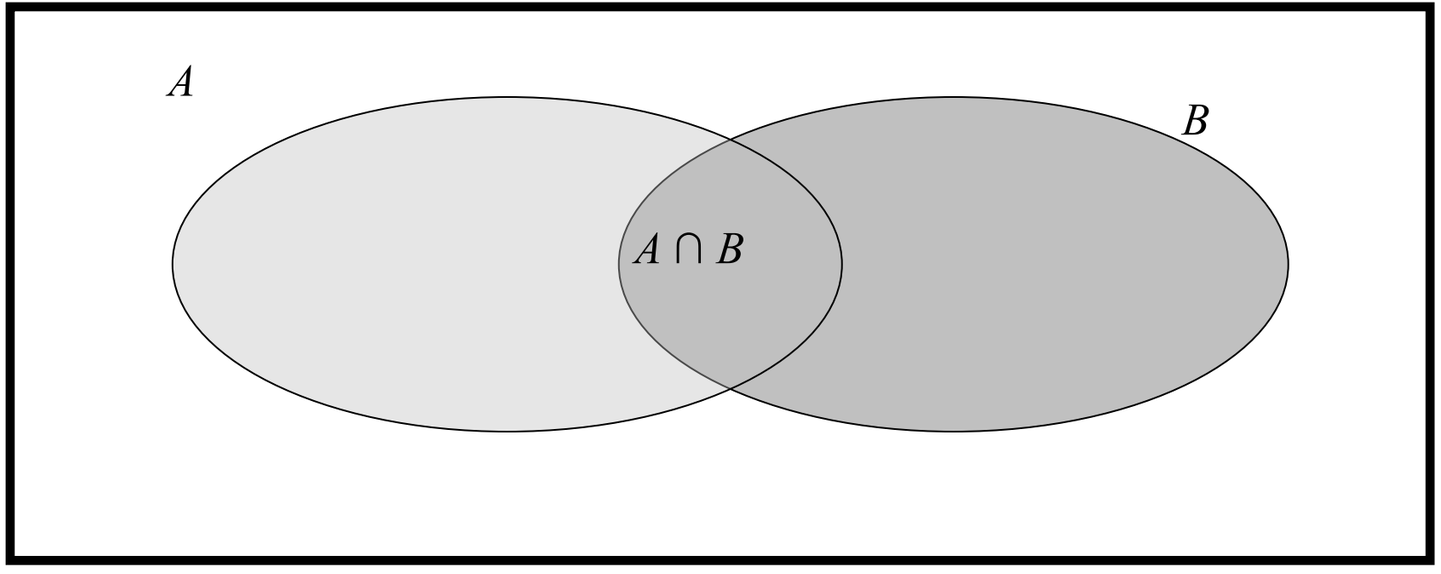
Rationale:

If we're told that event *B* has occurred then the sample space is restricted to *B*.

The probability within *B* has to be normalized, This is achieved by dividing by *P*[*B*]

The event *A* can now only occur if the outcome is in $A \cap B$. Hence the new probability of *A* is:

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$



An Example

The academy awards is soon to be shown.

For a specific married couple the probability that the husband watches the show is 80%, the probability that his wife watches the show is 65%, while the probability that they both watch the show is 60%.

If the husband is watching the show, what is the probability that his wife is also watching the show

Solution:

The academy awards is soon to be shown.

Let B = the event that the husband watches the show

$P[B]= 0.80$

Let A = the event that his wife watches the show

$P[A]= 0.65$ and $P[A \cap B]= 0.60$

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{0.60}{0.80} = 0.75$$

Example : Calculating the conditional probability of rain given that the barometric pressure is high.

Weather record shows that high barometric pressure (defined as being over 760 mm of mercury) occurred on 160 of the 200 days in a data set, and it rained on 20 of the 160 days with high barometric pressure.

If we let R denote the event “rain occurred” and H the event “ High barometric pressure occurred” and use the frequentist approach to define probabilities.

$P(H) = 160/200 = 0.8$

and $P(R \text{ and } H) = 20/200 = 0.10$

We can obtain the probability of rain given high pressure, directly from the data.

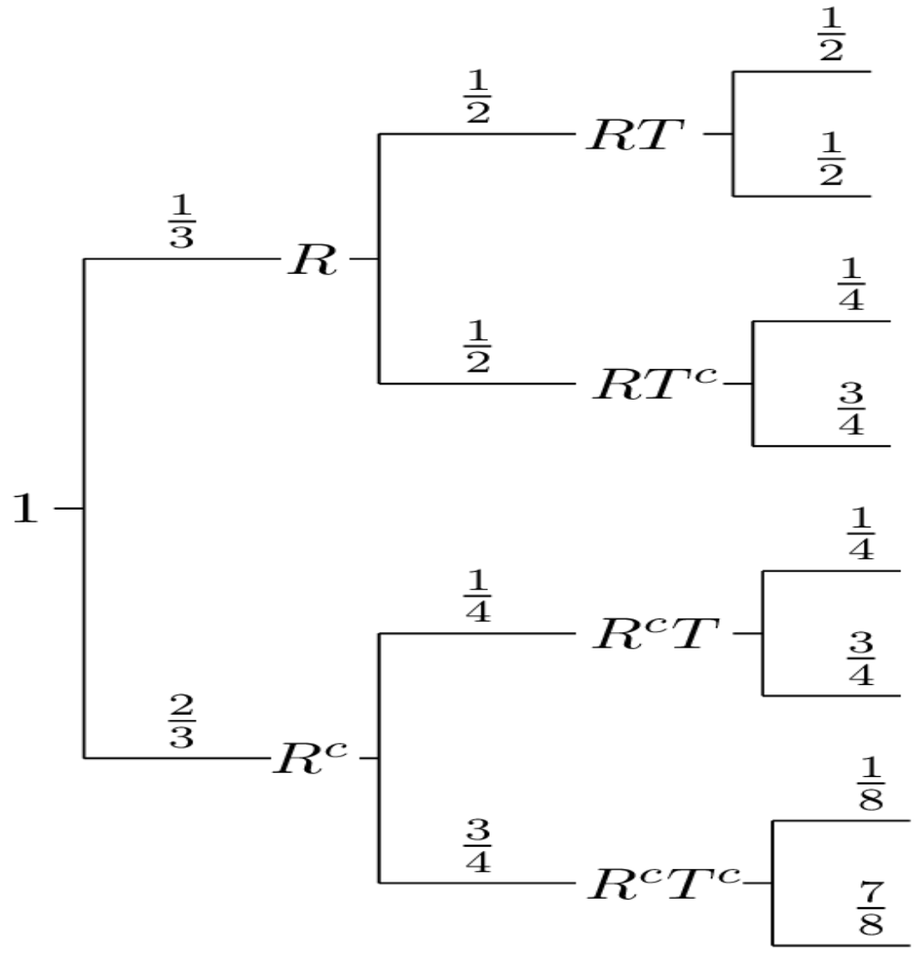
$P(R|H) = 20/160 = 0.125$

Using conditional probability

$P(R|H) = P(R \text{ and } H)/P(H) = 0.10/0.8 = 0.125.$

In my town, it's rainy one third of the days. Given that it is rainy, there will be heavy traffic with probability 1/2, and given that it is not rainy, there will be heavy traffic with probability 1/4. If it's rainy and there is heavy traffic, I arrive late for work with probability 1/2. On the other hand, the probability of being late is reduced to 1/8 if it is not rainy and there is no heavy traffic. In other situations (rainy and no traffic, not rainy and traffic) the probability of being late is 0.25. You pick a random day.

- What is the probability that it's not raining and there is heavy traffic and I am not late?
- What is the probability that I am late?
- Given that I arrived late at work, what is the probability that it rained that day?



$$RTL \to P(RTL) = \frac{1}{3} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{12}$$

$$RTL^c \to P(RTL^c) = \frac{1}{3} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{12}$$

$$RT^cL \to P(RT^cL) = \frac{1}{3} \times \frac{1}{2} \times \frac{1}{4} = \frac{1}{24}$$

$$RT^cL^c \to P(RT^cL^c) = \frac{1}{3} \times \frac{1}{2} \times \frac{3}{4} = \frac{3}{24} = \frac{1}{8}$$

$$R^cTL \to P(R^cTL) = \frac{2}{3} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{24}$$

$$R^cTL^c \to P(R^cTL^c) = \frac{2}{3} \times \frac{1}{4} \times \frac{3}{4} = \frac{1}{8}$$

$$R^cT^cL \to P(R^cT^cL) = \frac{2}{3} \times \frac{3}{4} \times \frac{1}{8} = \frac{1}{16}$$

$$R^cT^cL^c \to P(R^cT^cL^c) = \frac{2}{3} \times \frac{3}{4} \times \frac{7}{8} = \frac{7}{16}$$

Let **R** be the event that it's rainy, **T** be the event that there is heavy traffic, and **L** be the event that I am late for work. As it is seen from the problem statement, we are given conditional probabilities in a chain format. Thus, it is useful to draw a tree diagram for this problem. In this figure, each leaf in the tree corresponds to a single outcome in the sample space. We can calculate the probabilities of each outcome in the sample space by multiplying the probabilities on the edges of the tree that lead to the corresponding outcome

a. The probability that it's not raining and there is heavy traffic and I am not late can be found using the tree diagram which is in fact applying the chain rule:

$$P(R^c \cap T \cap L^c) = P(R^c)P(T|R^c)P(L^c|R^c \cap T)$$

$$= 2/3 \cdot 1/4 \cdot 3/4$$

$$= 1/8.$$

b. The probability that I am late can be found from the tree. All we need to do is sum the probabilities of the outcomes that correspond to me being late. In fact, we are using the law of total probability here.

P(L) =P(R and T and L)+P(R and Tc and L) + P(Rc and T and L) + P(Rc and Tc and L)

=1/12+1/24+1/24+1/16

=11/48.

c. We can find P(R|L) using

P(R|L)=P(R∩L)/P(L)

We have already found P(L)=11/48 and we can find P(R∩L) similarly by adding the probabilities of the outcomes that belong to R∩L.

In particular,

$P(R \cap L) = P(R,T,L) + P(R,Tc,L)$

$= 1/12 + 1/24$

$= 1/8$

Thus we obtain

$P(R|L) = P(R \cap L)/P(L)$

$= (1/8)/(11/48)$

$= 6/11.$

Random Variables

Real value generated in a random experiment is called Random value and this will be held by random variable.

A random variable is a rule that assigns a numerical value to an outcome of interest.

Example: Let us consider an experiment of tossing two coins.

Then sample space is S= { HH, HT, TH, TT}

Given X as random variable with condition number of heads.

X(HH) =2

X(HT) =1

X(TH) =1

X(TT) = 0

- Two types of random variables
  - **Discrete random variables** (countable set of possible outcomes)
  - **Continuous random variable** (unbroken chain of possible outcomes)

  A discrete random variable is described by its distribution function which lists for each outcome x the probability P(x) of x.
- Discrete random variables are understood in terms of their **probability mass function (*pmf*)**
- ***pmf*** ≡ a mathematical function that assigns probabilities to all possible outcomes for a discrete random variable.

Discrete random variables

- If the variable value is finite or infinite but countable, then it is called discrete random variable.

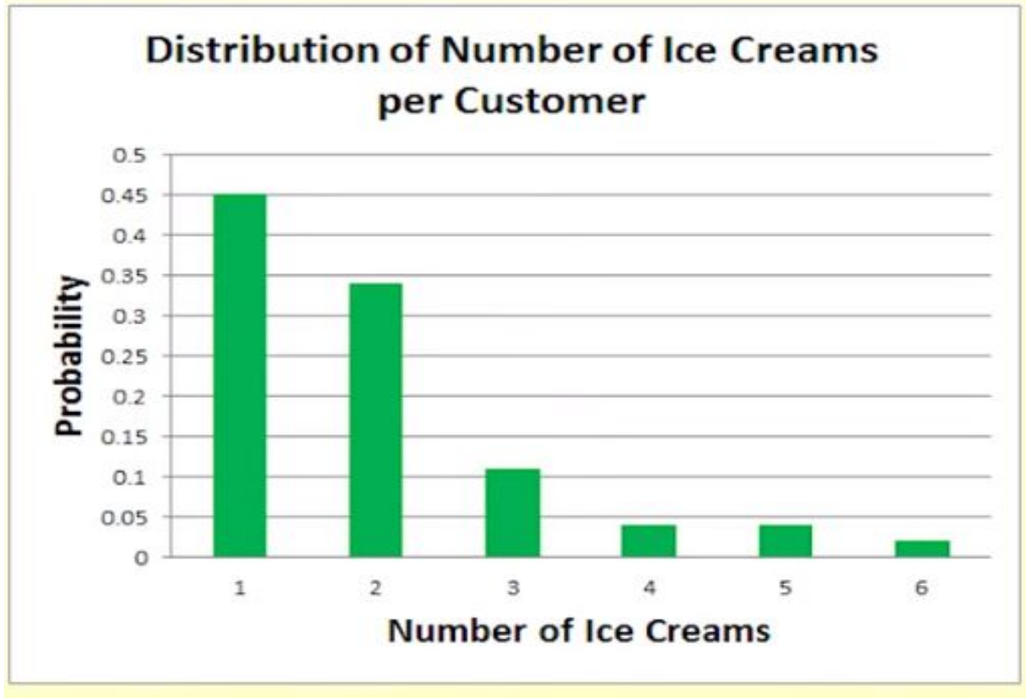
- Example : Tossing two coins or ice cream purchase in the following slides are examples for discrete random variable.

Example: Random variable

Number of ice creams	Customers
1	225
2	170
3	55
4	20
5	20
6	10

$X =$ number of ice creams a customer orders

Number of ice creams (x)	Customers	P(X= x)
1	225	0.45
2	170	0.34
3	55	0.11
4	20	0.04
5	20	0.04
6	10	0.02
Total	500	1



Probability of customers purchasing more than 3 ice creams:

- Consider P(X), where X>3

- Then it will be sum of P(X=4) + P(X=5) + P(X=6)= 0.04+0.04+0.02=0.10
- 10 percent of the customers will purchase more than 3 ice creams.

- However Summation of i=1 to n P(Xi)=1

Continuous Random Variable

- If the random variable values lies between two certain fixed numbers then it is called continuous random variable. The result can be finite or infinite.
- If X is the random value and it's values lies between a and b then,

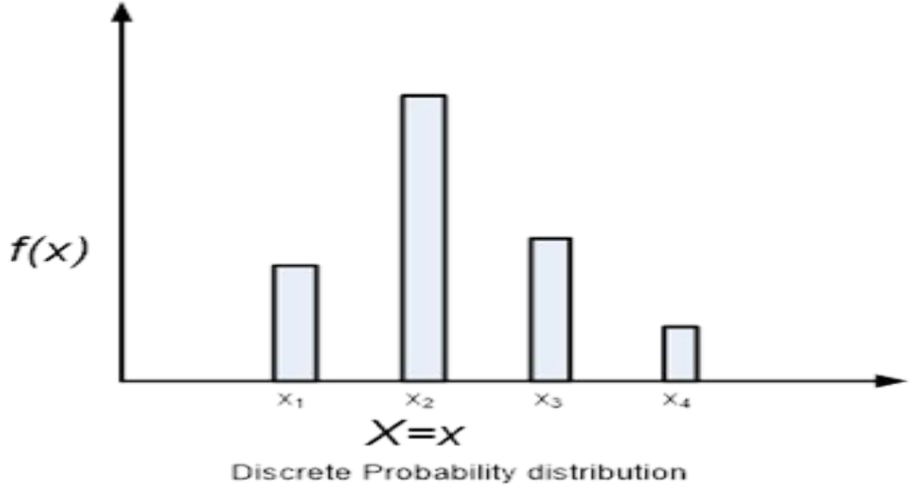
  It is represented by : a <= X <= b

  Example: Temperature, age, weight, height…etc. ranges between specific range.

- Continuous random variable differs from discrete random variable. Discrete random variables can take on only a finite number of values or at most a countably infinite values.
- A continuous random variable is described by Probability density function. This function is used to obtain he probability that the value of a continuous random variable is **in the given interval**.

- $\sum_{i=1}^{n} P(xi) = 1$

$\int_{-\infty}^{\infty} f(x)dx = 1$



Discrete Probability distribution

Continuous Probability Distribution

Probability distribution

- Frequency distribution is a listing of the observed frequencies of all the output of an experiment that actually occurred when experiment was done.

- Where as a probability distribution is a listing of the probabilities of all possible outcomes that could result if the experiment were done. (distribution with expectations).

Broad classification of Probability distribution

- Discrete probability distribution
  - Binomial distribution
  - Poisson distribution

- Continuous Probability distribution
  - Normal distribution

Example: Let us consider an experiment of tossing two coins.

Then sample space is S= { HH, HT, TH, TT}

- The distribution function for the number of heads from two flips of a coin.
- The random variable x is defined to be the total number of heads that occur when a fair coin is flipped two times.
- This random variable can have only 3 values 0,1,2, so it is discrete.
- Distribution function is (T, T), (T, H), (H, T), (H, H)

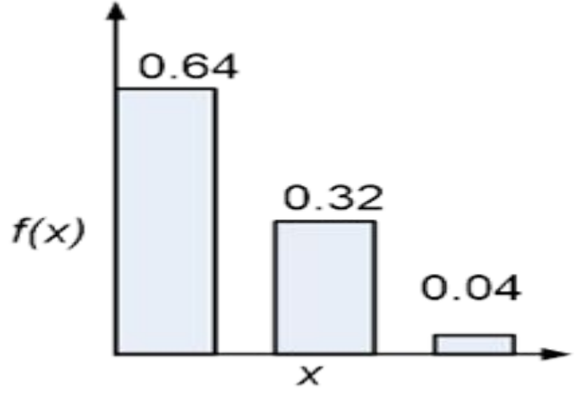
x	P(k)
0	1/4
1	2/4
2	1/4

- In general, a probability distribution function takes the following form.

x	x_1	$x_2 \dots \dots \dots \dots \dots x_n$
$f(x) = P(X = x)$	$f(x_1)$	$f(x_2) \dots \dots \dots f(x_n)$

- The table shows the *pmf* of a dataset. **Areas under** *pmf* graphs correspond to probability
- For example:
  Pr(X = 2)
  = shaded rectangle
  = height × base

X	Probability
0	0.64
1	0.32
2	0.04



Binomial Distribution

- A binomial distribution can be thought of as simply **the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times**. (When we have only two possible outcomes)

- Example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

Binomial Probability Distribution

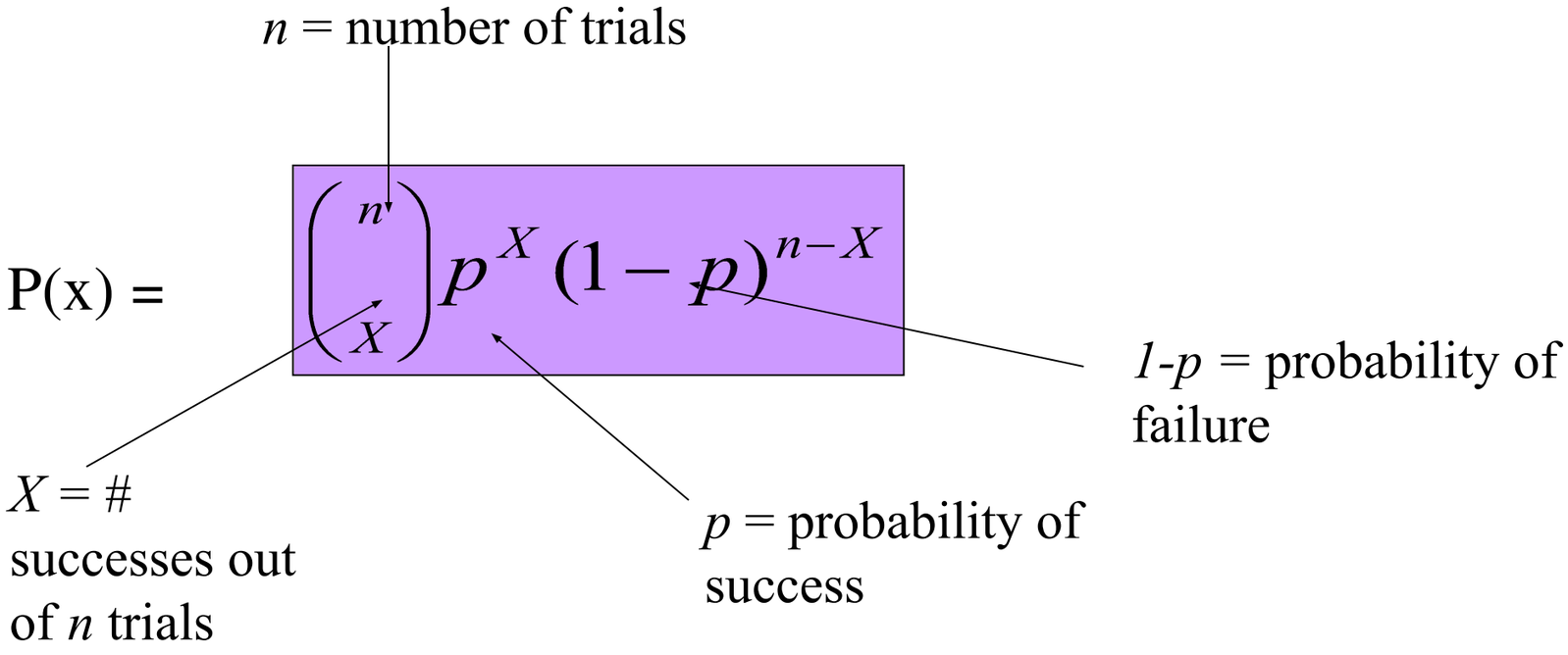
- A fixed number of observations (trials), n
  - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- A binary outcome
  - e.g., head or tail in each toss of a coin; disease or no disease
  - Generally called “success” and “failure”
  - Probability of success is p, probability of failure is 1 – p
- Constant probability for each observation
  - e.g., Probability of getting a tail is the same each time we toss the coin

Binomial distribution, generally

Note the general pattern emerging if you have only two possible outcomes (call them 1/0 or yes/no or success/failure) in *n* independent trials.

Total number of successes X obtained in trials is called a binomial random variable

then the probability of exactly *X* "successes" P(x) is as follows



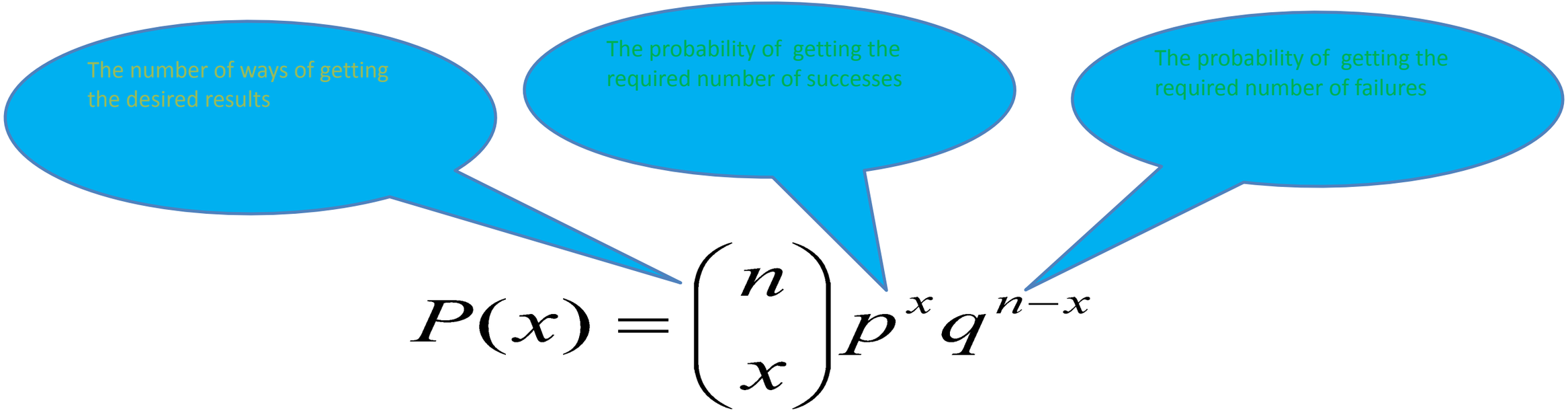
The Binomial Distribution

- The Binomial Probability Distribution
  - p = P(**S**) on a single trial
  - $q = 1 - p$
  - n = number of trials
  - x = number of successes

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

The Binomial Distribution

- The Binomial Probability Distribution



Binomial distribution: example

- If I toss a coin 20 times, what's the probability of getting exactly 10 heads?

$$\binom{20}{10}(.5)^{10}(.5)^{10} = .176$$

Binomial distribution: example

- If I toss a coin 20 times, what's the probability of getting of getting 2 or fewer heads?

$$\binom{20}{0}(.5)^0(.5)^{20} = \frac{20!}{20!0!}(.5)^{20} = 9.5x10^{-7} +$$
$$\binom{20}{1}(.5)^1(.5)^{19} = \frac{20!}{19!1!}(.5)^{20} = 20x9.5x10^{-7} = 1.9x10^{-5} +$$
$$\binom{20}{2}(.5)^2(.5)^{18} = \frac{20!}{18!2!}(.5)^{20} = 190x9.5x10^{-7} = 1.8x10^{-4}$$
$$= 1.8x10^{-4}$$

The Binomial Distribution

- Say 40% of the class is female.
- What is the probability that 6 of the first 10 students walking in will be female?

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

$$= \binom{10}{6} (.4^6)(.6^{10-6})$$

$$= 210(.004096)(.1296)$$

$$= .1115$$

Binomial Distribution: Illustration with example

- Consider a pen manufacturing company
- 10% of the pens are defective

- (i)Find the probability that exactly 2 pens are defective in a box of 12
- So n=12,
- p=10% = 10/100 = 1/10
- q= (1-q) =90/100 = 9/10
- X=2

$P(X=r) = {}^{n}C_{r}\, p^{r}\, q^{n-r}$

$P[X=2] = {}^{n}C_{x}\, p^{x}\, q^{n-x}$

$= {}^{12}C_{2}\,(1/10)^{2}\,(9/10)^{10}$

- Consider a pen manufacturing company
- 10% of the pens are defective

- (i)Find the probability that at least 2 pens are defective in a box of 12
- So n=12,
- p=10% = 10/100 = 1/10
- q= (1-q) =90/100 = 9/10
- X>=2
- P(X>=2) = 1- [P(X<2)]
- = 1-[P(X=0) +P(X=1)]

Continuous Probability Distributions

- When the random variable of interest can take any value in an interval, it is called continuous random variable.
  - Every continuous random variable has an infinite, uncountable number of possible values (i.e., any value in an interval).

- **Examples** Temperature on a given day, Length, height, intensity of light falling on a given region.
- The length of time it takes a truck driver to go from New York City to Miami.
- The depth of drilling to find oil.
- The weight of a truck in a truck-weighing station.
- The amount of water in a 12-ounce bottle.

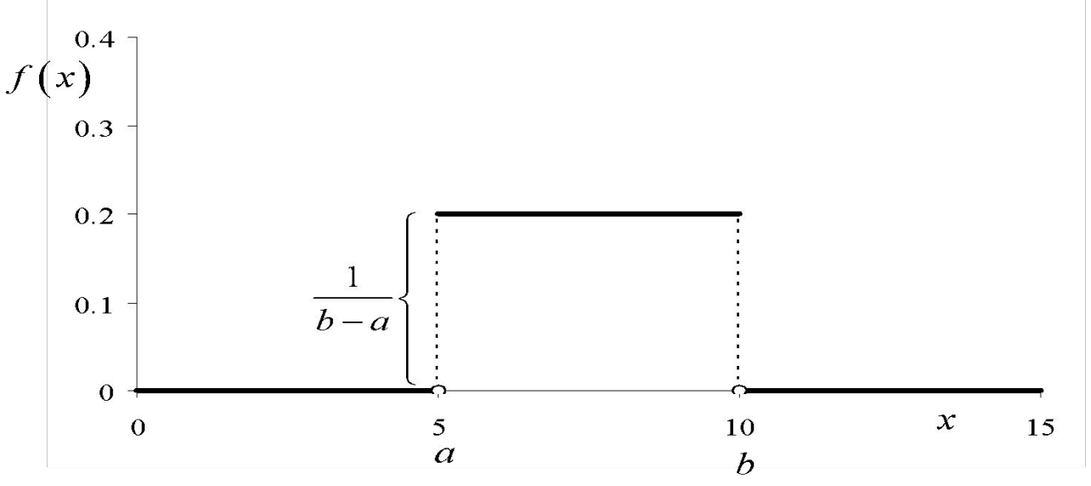
For each of these, if the variable is X, then $x>0$ and less than some maximum value possible, but it can take on any value within this range

Continuous Distributions: One of the simplest continuous distribution in all of statistics is the continuous uniform distribution.

The **Uniform distribution** from *a* to *b* : All values are equally likely to occur in the interval [a, b].

The denominator (b-a) makes the total area 1.

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



NORMAL DISTRIBUTION

- The most often used continuous probability distribution is the normal distribution; it is also known as **Gaussian** distribution.
- A continuous random variable X having the bell-shaped distribution is called a normal random variable.
- Its graph called the normal curve is the bell-shaped curve.
- Such a curve approximately describes many phenomenon occur in nature, industry and research.
  - Physical measurement in areas such as
  - Meteorological experiments,
    Rainfall studies and
    Measurement of manufacturing parts

NORMAL DISTRIBUTION

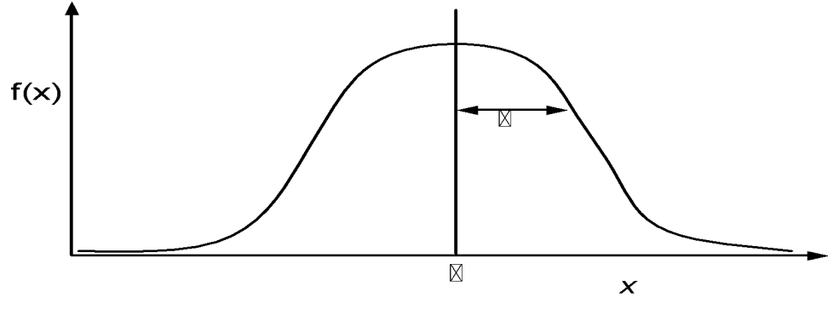
The normal (or Gaussian) distribution, is a very commonly used (occurring) function in the fields of probability theory, and has wide applications in the fields of:

- Pattern Recognition;
- Machine Learning;
- Artificial Neural Networks and Soft computing;
- Digital Signal (image, sound , video etc.) processing
- Vibrations, Graphics etc.

- The probability distribution of the normal variable depends upon the two parameters μ and σ
  - **The parameter μ is called the mean or expectation of the distribution.**
  - **The parameter σ is the standard deviation; and variance is thus σ^2.**
  - **standard deviation** is a measure of the amount of variation or dispersion of a set of values.
  - A low standard deviation indicates that the values tend to be close to the mean (expected value) of the set,
  - a high standard deviation indicates that the values are spread out over a wider range.
- The density of the normal variable x with mean μ and variance σ^2 is

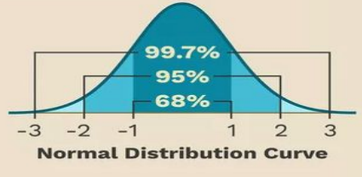
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

where $\pi = 3.14159$ ... and $e = 2.71828$ ….., the Naperian constant



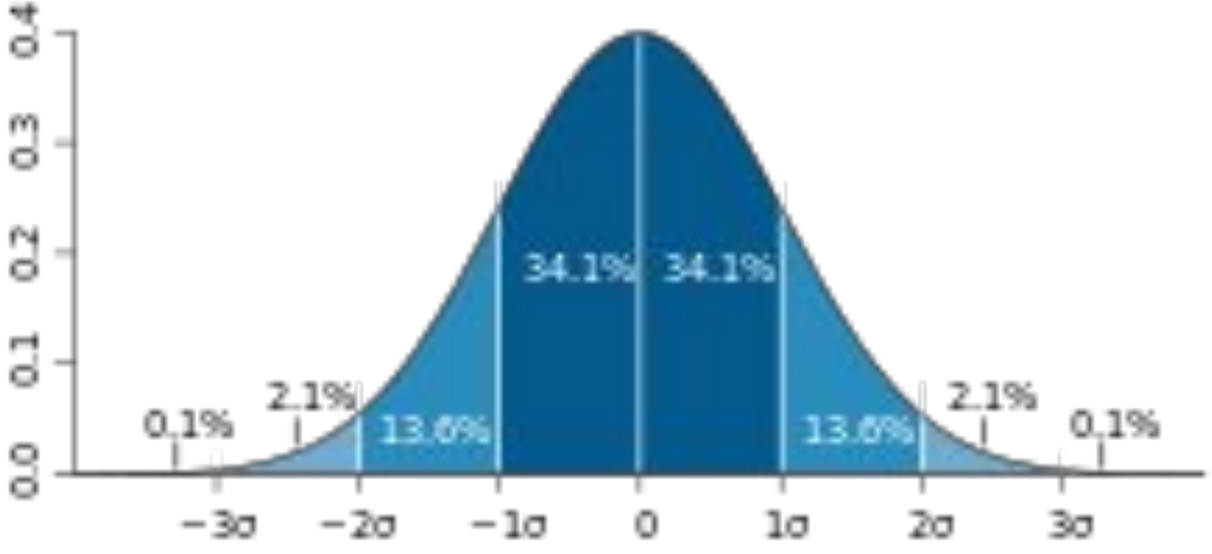
$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

n = The number of data points
x_i = Each of the values of the data
$\bar{x}$ = The mean of x_i



The Normal distribution
(mean μ, standard deviation σ)

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



A plot of normal distribution (or bell-shaped curve) where each band has a width of 1 standard deviation – See also: 68–95–99.7 rule.

Standard Normal Distribution : The distribution of a normal random variable with mean 0 and variance 1 is called a standard normal distribution.

Properties All forms of (normal) distribution share the following characteristics:

1. It is symmetric

A normal distribution comes with a perfectly symmetrical shape. This means that the distribution curve can be divided in the middle to produce two equal halves. The symmetric shape occurs when one-half of the observations fall on each side of the curve.

2. The mean(average), median(mid point), and mode(max repeated term) are equal

The middle point of a normal distribution is the point with the maximum frequency, which means that it possesses the most observations of the variable. The midpoint is also the point where these three measures fall. The measures are usually equal in a perfectly (normal) distribution.

3. Empirical rule

In normally distributed data, there is a constant proportion of distance lying under the curve between the mean and specific number of standard deviations from the mean. For example, 68.25% of all cases fall within +/- one standard deviation from the mean. 95% of all cases fall within +/- two standard deviations from the mean, while 99% of all cases fall within +/- three standard deviations from the mean.

For Standard Normal distribution:

- For standard normal distribution, the area under the given range is given by:

$$\int_a^b p(x)\,dx = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz$$

$$= C\left(\frac{b-\mu}{\sigma}\right) - C\left(\frac{a-\mu}{\sigma}\right).$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986
3.0	0.998650								
3.5	0.9998674								
4.0	0.99996833								
4.5	0.999996602								
5.0	0.999999997133								
5.5	0.999999998101								
6.0	0.999999999013								
6.5	0.999999999999588								
7.0	0.999999999999872								

$$C(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

Figure 2.11: Areas under the standard normal curve

z	0.00	−0.01	−0.02	−0.03	−0.04	−0.05	−0.06	−0.07	−0.08	−0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014

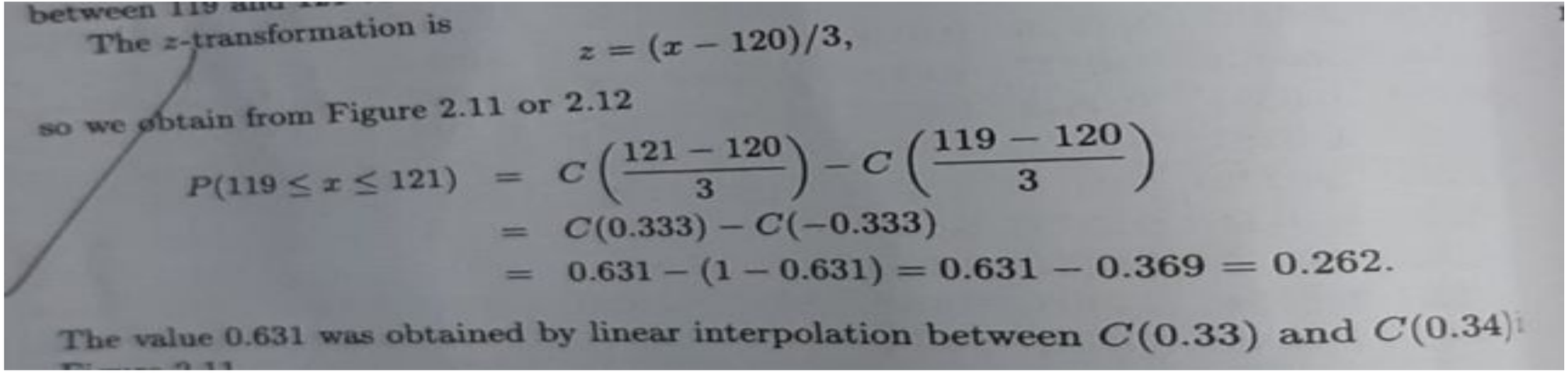
z	
−3.0	1.350×10^{-3}
−3.5	2.326×10^{-4}
−4.0	3.167×10^{-5}
−4.5	3.398×10^{-6}
−5.0	2.867×10^{-7}
−5.5	1.800×10^{-8}
−6.0	9.866×10^{-10}
−6.5	4.016×10^{-11}
−7.0	1.280×10^{-12}

$$C(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

igure 2.12: Areas under the standard normal curve from $-\infty$ to z, where $z < 0$. For

Problem: Normal distribution

- Consider an electrical circuit in which the voltage is normally distributed with mean 120 and standard deviation of 3. What is the probability that the next reading will be between 119 and 121 volts?
-



between 119 and

The z-transformation is

$$z = (x - 120)/3,$$

so we obtain from Figure 2.11 or 2.12

$$\begin{aligned} P(119 \le x \le 121) &= C\left(\frac{121-120}{3}\right) - C\left(\frac{119-120}{3}\right) \\ &= C(0.333) - C(-0.333) \\ &= 0.631 - (1 - 0.631) = 0.631 - 0.369 = 0.262. \end{aligned}$$

The value 0.631 was obtained by linear interpolation between $C(0.33)$ and $C(0.34)$

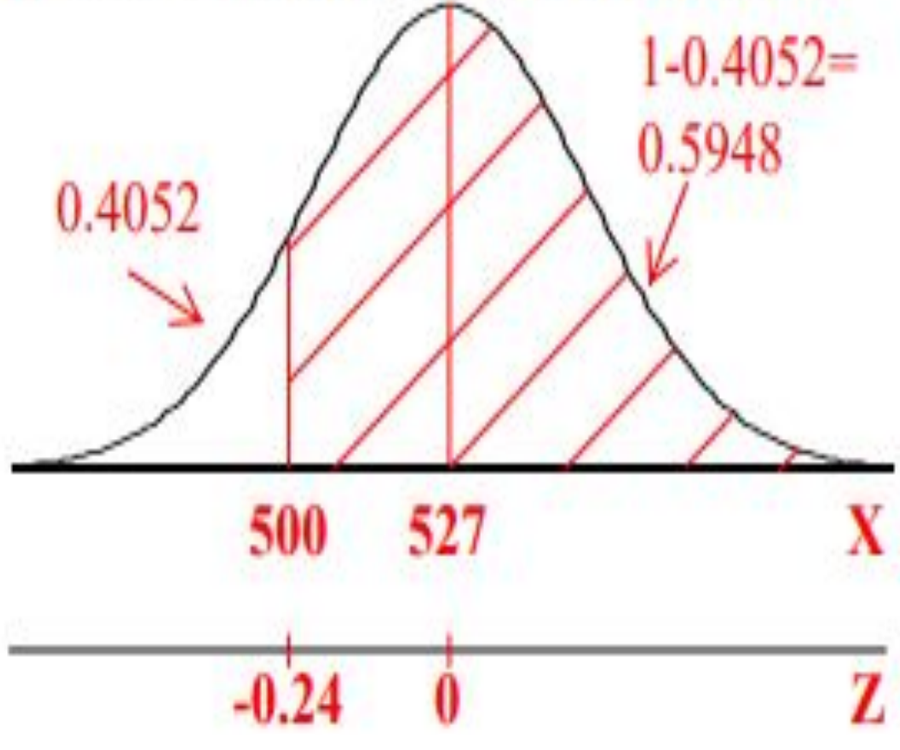
1. Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination. Scores on the GMAT are roughly normally distributed with a mean of 527 and a standard deviation of 112. What is the probability of an individual scoring above 500 on the GMAT?

Normal Distribution $$Z = \frac{500 - 527}{112} = -0.24107$$

$\mu = 527$

$\sigma = 112$

$\Pr\{X > 500\} = \Pr\{Z > -0.24\} = 1 - 0.4052 = \boxed{0.5948}$



Another problem

Normal Probabilities Practice Problems Solution

5. The average number of acres burned by forest and range fires in a large New Mexico county is 4,300 acres per year, with a standard deviation of 750 acres. The distribution of the number of acres burned is normal. What is the probability that between 2,500 and 4,200 acres will be burned in any given year?

Normal Distribution

$$Z = \frac{2500 - 4300}{750} = -2.40$$

$\mu = 4300$

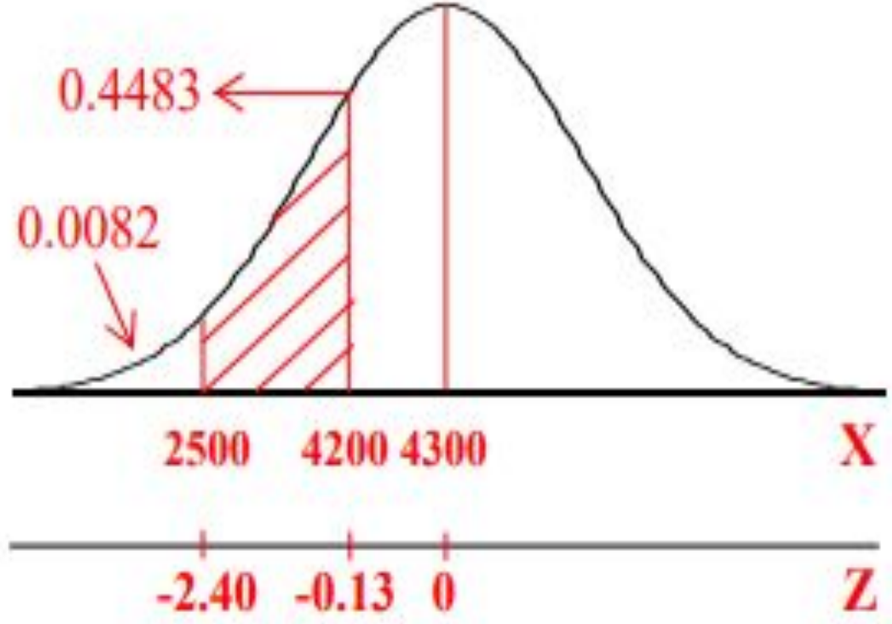
$$Z = \frac{4200 - 4300}{750} = -0.13333$$

$\sigma = 750$

$P(2500 < X < 4200) = P(-2.40 < Z < -0.13)$

$P(-2.40 < Z < -0.13) = P(Z < -0.13) - P(Z < -2.40)$

$P(-2.40 < Z < -0.13) = 0.4483 - 0.0082 = \boxed{0.4401}$



4. A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is travelling at more than 100 km/hr?

Let x be the random variable that represents the speed of cars. x has $\mu = 90$ and $\sigma = 10$.

We have to find the probability that x is higher than 100 or $P(x > 100)$
For $x = 100$, $z = (100 - 90) / 10 = 1$
$P(x > 90) = P(z > 1) =$ [total area] - [area to the left of $z = 1$]
$= 1 - 0.8413 = 0.1587$
The probability that a car selected at a random has a speed greater than 100 km/hr is equal to 0.1587

5. For a certain type of computers, the length of time bewteen charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. John owns one of these computers and wants to know the probability that the length of time will be between 50 and 70 hours.

Let x be the random variable that represents the length of time. It has a mean of 50 and a standard deviation of 15.

We have to find the probability that x is between 50 and 70 or P(50< x < 70)
For x = 50 , z = (50 - 50) / 15 = 0
For x = 70 , z = (70 - 50) / 15 = 1.33 (rounded to 2 decimal places)
P(50< x < 70) = P(0< z < 1.33) = [area to the left of z = 1.33] - [area to the left of z = 0]
= 0.9082 - 0.5 = 0.4082
The probability that John's computer has a length of time between 50 and 70 hours is equal to 0.4082.

6. Entry to a certain University is determined by a national test. The scores on this test are normally distributed with a mean of 500 and a standard deviation of 100. Tom wants to be admitted to this university and he knows that he must score better than at least 70% of the students who took the test. Tom takes the test and scores 585. Will he be admitted to this university?

Let x be the random variable that represents the scores. x is normally distributed with a mean of 500 and a standard deviation of 100.

The total area under the normal curve represents the total number of students who took the test. If we multiply the values of the areas under the curve by 100, we obtain percentages.
For x = 585 , z = (585 - 500) / 100 = 0.85
The proportion P of students who scored below 585 is given by
P = [area to the left of z = 0.85] = 0.8023 = 80.23%
Tom scored better than 80.23% of the students who took the test and he will be admitted to this University.

Joint Distributions and Densities

If we have two random variables X and Y, and we would like to study them jointly, we define the **joint probability mass function** as follows:

- If random variables X and Y are discrete, the joint density function of the joint random variable (x, y) is **the probability P(x, y) that both x and y occur.**
- Thus the joint probability of every possible combination of outcomes of the random variables that make up the joint random variable.

Ex: **The joint distribution function of the outcomes of flipping biased coins.**

A biased coin A has P(head) = 0.6 and B has P(head) = 0.3. Suppose that we flip coin A, and if the outcome is head, we flip coin B, but if the result of the first flip is tail, we flip A again. Since the probability of head on the second flip depends on the outcome of the first flip, two events are not independent. Let x be the outcome of the first flip, and let y be the outcome of the second flip. Write a table indicating probability distribution.

Joint distribution in continuous random variable

- If x and y are continuous, then the probability density function is used over the region R, where x and y is applied is used.
- It is given by:

$$P((x,y) \text{ is in } R) = \iint_R p(x,y)\,dx\,dy,$$

- Where the integral is taken over the region R. This integral represents a volume in the xyp-space.
- A continuous one dimensional random variable can take any value in the given interval, but a continuous two-dimensional random variable can take on any value in a two dimensional region.

Probability distributions can be used to **describe the population**, just as we described samples .

- **Shape:** Symmetric, skewed, mound-shaped…
- **Outliers:** unusual or unlikely measurements
- **Center and spread:** mean and standard deviation. A population mean is called **μ** and a population standard deviation is called **σ.**

Let *x* be a discrete random variable with probability distribution *p(x)*. Then the mean, variance and standard deviation of *x* are given as

$$\text{Mean}: \mu = \sum xp(x)$$

$$\text{Variance}: \sigma^2 = \sum (x-\mu)^2 p(x)$$

$$\text{Standard deviation}: \sigma = \sqrt{\sigma^2}$$

Variance, continuous

Discrete case:

$$Var(X) = \sum_{\text{all x}} (x_i - \mu)^2 \, p(x_i)$$

Continuous case:

$$Var(X) = \int_{\text{all x}} (x_i - \mu)^2 \, p(x_i) dx$$

Moments are very useful in statistics because they tell us much about our data.

- In mathematics, the **moments** of a function are **quantitative measures related to the shape of the function's graphs.**
- If the function represents mass, then the first moment is the center of the mass, and the second moment is the rotational inertia The mathematical concept is closely related to the concept of moment in physics.
- If the function is a probability distribution, then there are four commonly used moments in statistics

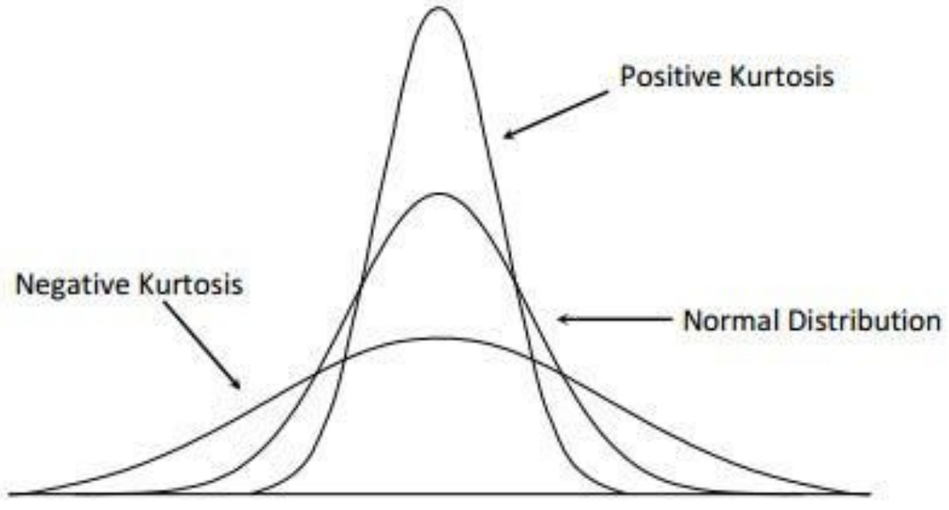
The first moment is the expected value - measure of center of the data

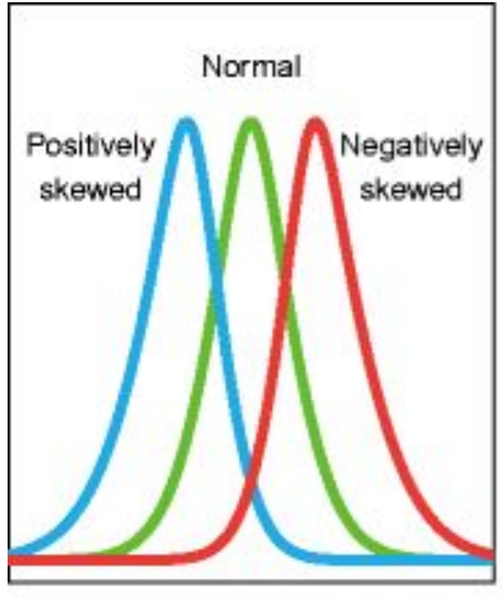
The second is the variance - spread of our data about the mean

The third standardized moment is the skewness - the shape of the distribution. **skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative or undefined.

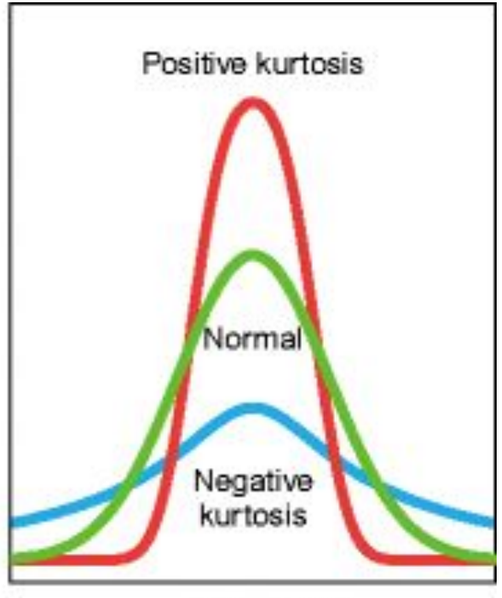
The fourth standardized moment is the kurtosis - measures the peakedness or flatness of the distribution.

Skewness gives you information about a distribution's "shift", or lack of symmetry.
Distributions with a left skew have long left tails; Distributions with a right skew have long right tails.





Skewness



Kurtosis

- **Positive kurtosis = a lot of data in the tails.**
- **Negative kurtosis = not much data in the tails.**

The “moments” of a random variable (or of its distribution) are **expected values of powers** or related functions of the random variable.

In particular, the first moment is the mean, μX = E(X). The mean is a measure of the “center” or “location” of a distribution.

The k^{th} moment of X.

$$\mu_k = E\left(X^k\right)$$

$$= \begin{cases} \sum_x x^k p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^k f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

Let X be a discrete random variable having support Rx = <1, 2> and the pmf is

$$p_X(x) = \begin{cases} 3/4 & \text{if } x = 1 \\ 1/4 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

using this compute

Solution:

$$\begin{aligned} \mu_X(1) &= \mathrm{E}[X] \\ &= \sum_{x \in R_X} p_X(x)x \\ &= \frac{3}{4} \cdot 1 + \frac{1}{4} \cdot 2 \\ &= \frac{5}{4} \end{aligned}$$

A central moment is a moment of a probability distribution of random variables about the random variables Mean (Expected Value).

$$\mu_k^{\mathrm{o}} = E\left[(X - \mu)^k\right]$$

The k^{th} **central moment** of X

$$= \begin{cases} \sum_x (x-\mu)^k \, p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^k \, f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

Expected Values of Discrete Random Variables

- The **variance** of a **discrete random variable** *x* is

$$\sigma^2 = E[(x-\mu)^2] = \sum (x-\mu)^2 \, p(x).$$

- The **standard deviation** of a **discrete random variable** *x* is

$$\sqrt{\sigma^2} = \sqrt{E[(x-\mu)^2]} = \sqrt{\sum (x-\mu)^2 \, p(x)}.$$

- Example : Let X be a discrete random variable having support Rx = <1, 2, 3> and pmf is as listed below

$$p_X(x) = \begin{cases} 1/2 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 1/6 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

- The third moment of can be computed as shown:

$$\begin{aligned} \mu_X(3) &= \mathrm{E}\left[X^3\right] \\ &= \sum_{x \in R_X} p_X(x) x^3 \\ &= \frac{1}{2} \cdot 1^3 + \frac{1}{3} \cdot 2^3 + \frac{1}{6} \cdot 3^3 \\ &= \frac{1}{2} + \frac{8}{3} + \frac{27}{6} \\ &= \frac{3 + 16 + 27}{6} = \frac{46}{6} = \frac{23}{3} \end{aligned}$$

- The third central moment of can be computed as follows:

$$
\begin{aligned}
\mu_X(3) &= \mathrm{E}\left[(X-\mathrm{E}[X])^3\right] \\
&= \sum_{x\in R_X} p_X(x)\left(x-\frac{5}{4}\right)^3 \\
&= \frac{3}{4}\cdot\left(1-\frac{5}{4}\right)^3+\frac{1}{4}\cdot\left(2-\frac{5}{4}\right)^3 \\
&= \frac{3}{4}\cdot\left(-\frac{1}{4}\right)^3+\frac{1}{4}\cdot\left(\frac{3}{4}\right)^3 \\
&= -\frac{3}{4^4}+\frac{27}{4^4} \\
&= \frac{24}{256}=\frac{3}{32}
\end{aligned}
$$

Fo the values of x listed in the table compute

1. 1st, 2nd, 3rd and 4th order raw moments
2. 1st, 2nd, 3rd and 4th order central moments

X	P(x)
0	1/6
1	2/6
2	2/6
3	1/6

Estimation of Parameters from samples: Method of moments

To estimate parameters using methods of moments, 'n' independent samples or patterns $x_1, x_2, x_3 \ldots x_n$ are collected from random variable x, which may be continuous or discrete.

Randomly choose one of these samples in the data set, let its value to be a new discrete random variable x' called the empirical random variable, which takes on of these values $x_1, x_2, x_3 \ldots x_n$.

Each with a probability 1/n.

Compute the sample mean and sample variance using the formula

$$\mu = \sum x_i p(x_i)$$

$$\sigma^2 = \sum (x_i - \mu)^2 \, p(x_i)$$

The method of moments can also be used to compute covariance :

covariance is a measure of the relationship between two random variables.

$$\mathrm{Cov}\,(X, Y) = \frac{\sum (X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables.

Maximum Likelihood Estimates:

- Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed.
- To compute maximum likelihood estimate, choose parameter (or set of parameters) that maximises the joint distribution function or multivariate density function for the entire data set when it is evaluated at the sample points $x_1, x_2, x_3 \ldots x_n$.