

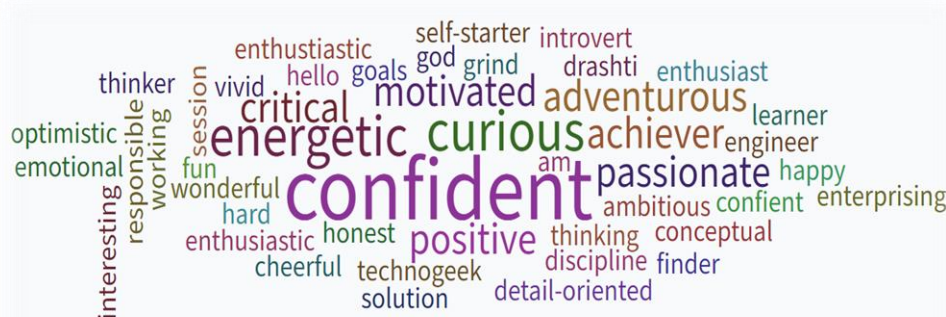


Northeastern University  
Toronto

# ALY 6000

## Introduction to Data Analytics

Mohammad (**Shafiqu**) Islam, PhD., P.Eng.  
Email: [m.islam@northeastern.edu](mailto:m.islam@northeastern.edu)



Northeastern  
University

# Agenda

---

- Module 2 Review
- Module 3
  - Probability
  - Probability distribution
- Module Project
- Summary



- I will use this deck as a reference for this class

# Pulse Check

---

## Join by Web



- 1 Go to **PollEv.com**
- 2 Enter **MISLAM933**
- 3 Respond to activity

## Join by Text



- 1 Text **MISLAM933** to **37607**
- 2 Text in your message

*Your participation and attendance matter*



# Module 2 Review

---




## Key Issues:

- R Programming:
  - New to Programming
  - Plot vs ggplot2 and other functions
- Reporting:
  - What to write?
  - How to write APA references?
  - Where to put my code/repo in the report?
- Grading
  - Grade Improvement



# Report Grading

Grading

Module 1 Assignment Rubric					
Criteria	Ratings				Pts
Summary and Report Format [M1L1] <a href="#">view longer description</a>	35 to >31.15 pts Exceeds Standard	31.15 to >28 pts Meets Standard	28 to >24.15 pts Approaching Standard	24.15 to >0 pts Below Standard	<input type="text"/> / 35 pts 
	Exceeds with exceptional professional layout and presentation skills.	Report provides a concisely written paragraph summarizing the key points of your data analysis and draws accurate takeaways from the dataset. Report is without proofreading errors. References are cited using correct format.	Report provides a paragraph summarizing the key points of your data analysis, but may include irrelevant information, not enough meaningful detail, or takeaways from the dataset are unclear. Report may have multiple proofreading errors or references are cited incorrectly.	Does not provide a summary or the summary does not reflect the key points of your data analysis or takeaways from the dataset. Report has significant proofreading errors and/or no citations for references used.	
Data Analysis [M1L4] <a href="#">view longer description</a>	35 to >31.15 pts Exceeds Standard	31.15 to >28 pts Meets Standard	28 to >24.15 pts Approaching Standard	24.15 to >0 pts Below Standard	<input type="text"/> / 35 pts 
	Exceeds with insightful analysis that goes beyond an accurate understanding of data types, descriptive statistics, and uses R creatively to support these insights.	Provides analysis that reflects an accurate understanding of data types, descriptive statistics, and uses appropriate R commands and parameters for statistical computing and graphics. Includes R console screenshots in report.	Provides analysis that reflects a general understanding of data types and descriptive statistics. There may be errors in statistical computing or using R commands and parameters. Includes R console screenshots in report.	Provides an analysis that reflects a lack of understanding of data types, descriptive statistics, or use of R commands and parameters for statistical computing and graphics.	
Data Visualization [M1L2] <a href="#">view longer description</a>	30 to >26.7 pts Exceeds Standard	26.7 to >23.7 pts Meets Standard	23.7 to >20.7 pts Approaching Standard	20.7 to >0 pts Below Standards	<input type="text"/> / 30 pts 
	Exceeds with unexpected plots or technical content or unusual artistry.	Provides all required data visualizations. Visualizations support key findings, includes descriptive statistics, and any provided text is meaningfully connected to the visual results.	Provides all required data visualizations. Visualizations that support each key finding may be missing or meaningful text connected to the visual results may be lacking.	Does not provide all required data visualizations	

Total Points: 0 out of 100



# Example Work



Toronto, Canada

A Report on  
**Executive Summary of module 1**

Under the subject of:  
Introduction to Data Analytics  
ALY 6000 (Module 1)

Guided by:  
Prof. Mohammad Shafiqul Islam

Submitted By:

Name of Student	NUID	Date of submission
		16 <sup>th</sup> January, 2021

## Objective:

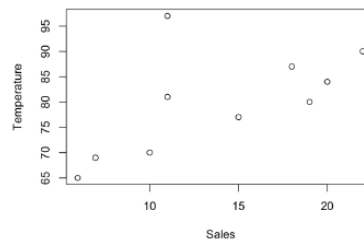
To create an executive summary based on a set of data and instructions provided, by writing and executing an R script in order to gather the required information.

## Introduction:

This module mainly encompasses Descriptive and Inferential statistics, how to collect a sample using different types of sampling methods from a given set of populations and how R can be utilized in analysing as well as visualizing statistical information. Based on these learnings the following summary report highlights the observations and findings after running the R script on the given instruction set.

## Findings & Observations:

### 1. Plotting Sales ~ Temperature scatter plot:



To plot a Scatter Plot in R, we use "plot()" functionality to visualize raw data. In the above graph we can correlate that the sales and temperature are dependent on each other, and the total number of sales increased consistently with gradual increase in temperature.

### 2. Mean of the Temperature

```
> mean(Temperature)
[1] 80
```

Using mean() function in R, It can be found that the mean of the Temperature from the data provided is 80.

### 3. Manipulating the vectors in R<sup>2</sup>

In R each element in a vector has an index value on which they are positioned. To add or remove an element in a vector one can manipulate using the following representation `vector_name[]`

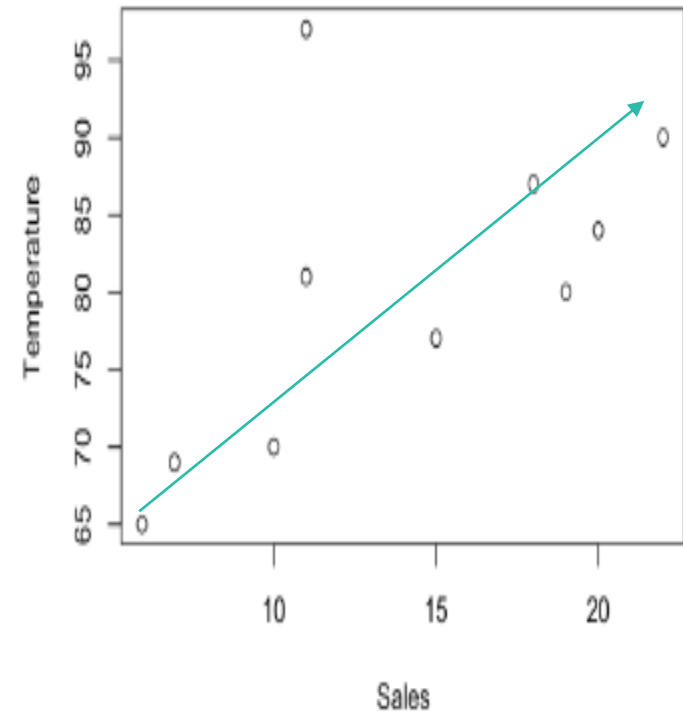
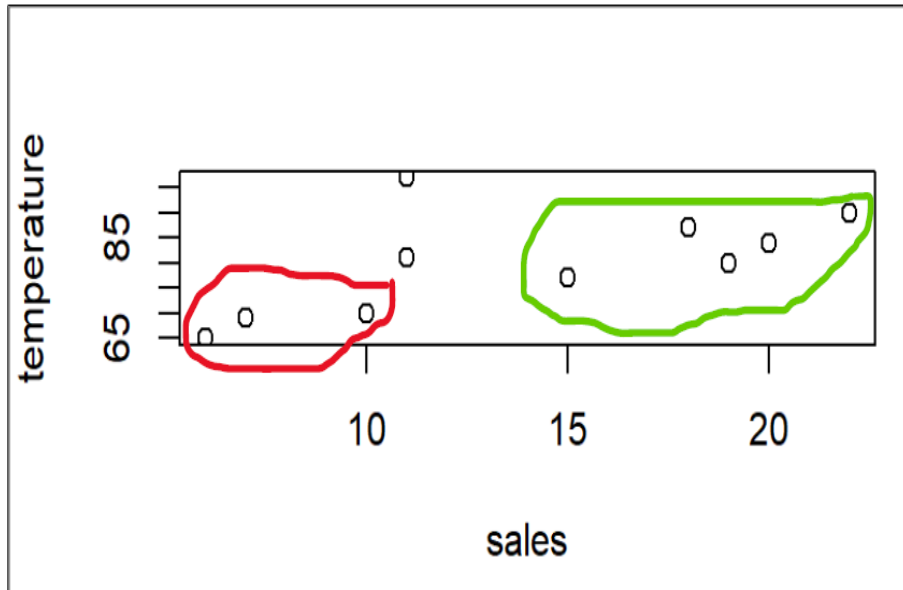
<sup>1</sup> Followed page 9 from the Textbook "R in Action" by Robert I. Kabacoff referred in bibliography [1]

<sup>2</sup> Learnt how to insert an element into R from geeksforgeeks referred in bibliography [4]



# Example Work

---



# Common Comments

---

- Cover page could be nicer
- An Introduction is missing...
- Please write references using APA format.
- Please incorporate repo as discussed in Module 2 class
- I can't find your repo. Please make sure it is public not private.
- Put your github repo in the reference section.
- Please give some basic description about this repo in readme.md file in your repo.
- Write more comments on your code.

## BIBLIOGRAPHY

- <https://youtu.be/VmOIVFXBsyY>
- <https://www.datamentor.io/r-programming/matrix/>
- [https://www.tutorialspoint.com/r/r\\_data\\_frames.htm](https://www.tutorialspoint.com/r/r_data_frames.htm)
- <https://www.r-bloggers.com/2009/11/r-tutorial-series-summary-and-descriptive->



Mohammad Islam

Please write references using APA format.



# Module 3: Probability and Counting

---

# Module Overview

---

- Data scientists need to be familiar with probability to answer business questions, which may influence strategic decisions managers make.
- This module provides an explanation of probability for processes with a finite number of possible outcomes. It explains the meaning of probability, as well as how to calculate probabilities. It also examines the relationship between disjoint and independent events.
- First, we will deal with the probability of a single event. We will look at the equation for probability, which is used to calculate the probabilities of various events.
- Finally, we will introduce some combinatorial methods, or to put it simply, ways of counting things.



# Learning Objectives

---

By the end of this module, you should be able to:

- Represent the probability of events using mathematical notation
- Solve probability problems using addition and multiplication rules of probability
- Calculate counts and probabilities based on categorical data
- Graph probability distributions
- Interpret data displayed in graphs
- Use R to manipulate datasets



# Task List

---

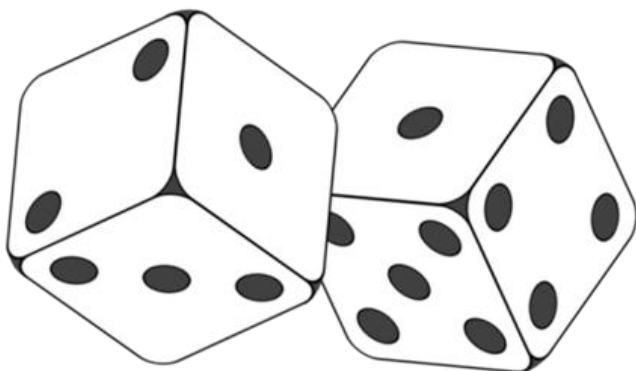
- Statistics (view the lessons in Canvas, do assigned reading and ungraded practice problems)
- Bluman textbook - Chapters 4 and 5
- Complete primary Discussion post by Friday
- Learning to use R Practice (do assigned reading, watch instructor videos, complete R practice tasks)
- Kabacoff textbook - Chapter 4
- Read the Module 3 Project assignment



# Probability Related Basic Definitions

---

Suppose a coin is tossed and the up face is recorded.



The two possible outcomes of this experiment are:  
Observe a tail (T), Observe a head (H).



# Probability Related Basic Definitions

---

- The result is called an **observation**, and the process of making an observation is called an **experiment**.
- An experiment is any process of observation with an uncertain outcome
- The possible **outcomes** for an experiment are called the experimental outcomes.
- Each one of the above possible outcomes is called an **outcome**, or a simple **event**, or a **sample point**.
- A sample point is the most basic outcome of the experiment.
- The **sample space** of an experiment is the collection of all its sample points. In our example, the sample space, denoted by  $S$ , is:  $S = \{T, H\}$ .



# Probability related Basic Definition

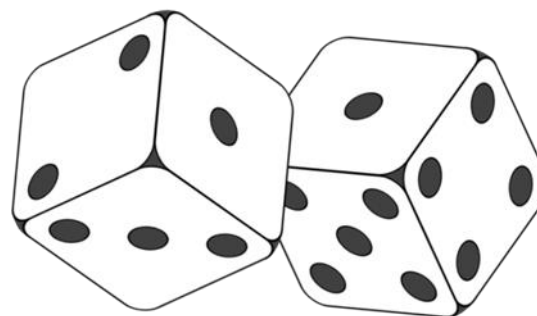
---

**Probability** is a measure of the chance that an experimental outcome will occur when an experiment is carried out

The probability of a single event  $P(A)$  is:

$$\text{Probability of an outcome} = \frac{\text{The number of time the outcome is observed}}{\text{The total number observable outcomes}}$$

What if we tossed a coin one? What is the sample space of this experiment? What is the probability of head ? tail ? head or tail? head and tail?



# Probability related Basic Definition

---

**What if we tossed a coin twice? What is the sample space of this experiment?**

- Sample Space  $S = \{TT, TH, HT, HH\}$ .
- The sample space of an experiment is the set of all possible experimental outcomes
- The experimental outcomes in the sample space are called sample space outcomes
- There are four possible outcomes, and the sample space is the collection of all above sample points:
- Now we may define an event  $A$  as: “Observing at least one tail.” In this case,  $A = \{TH, HT, TT\}$ .

$$P(A) = \frac{\text{The number of times the outcome is observed}}{\text{The total number of observable outcomes}} = \frac{3}{4}$$

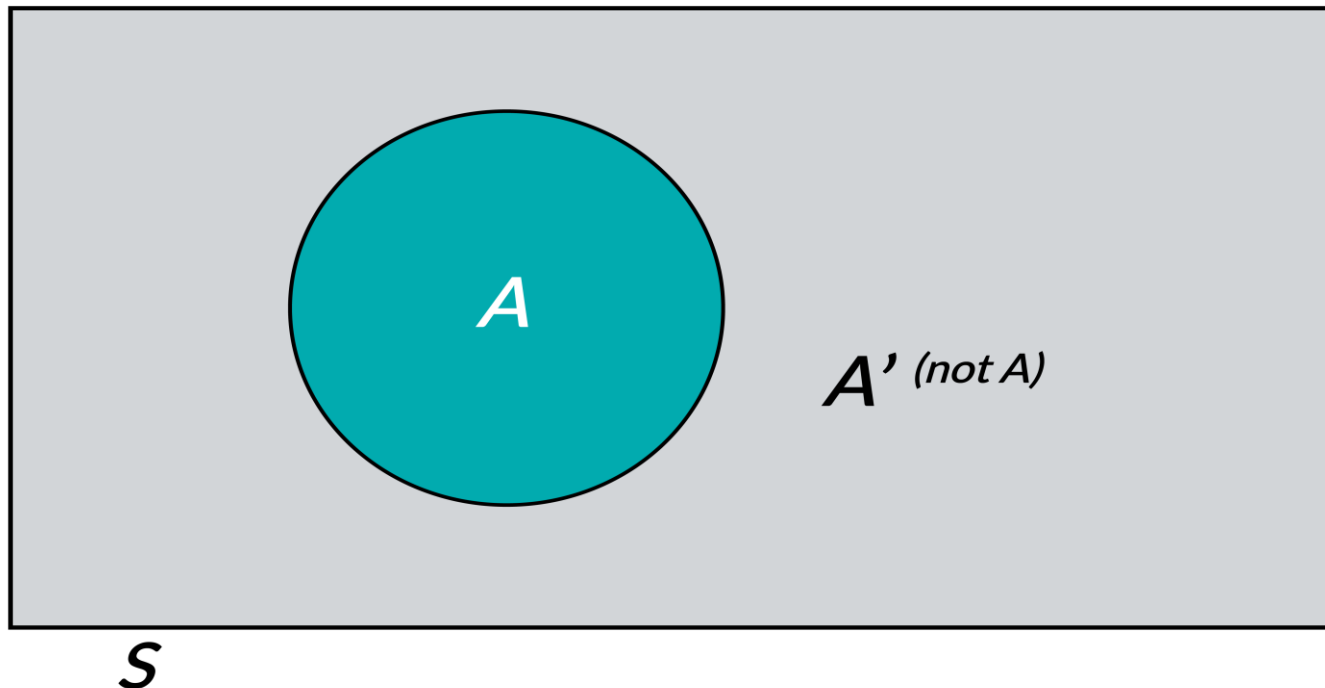




# Venn Diagrams

---

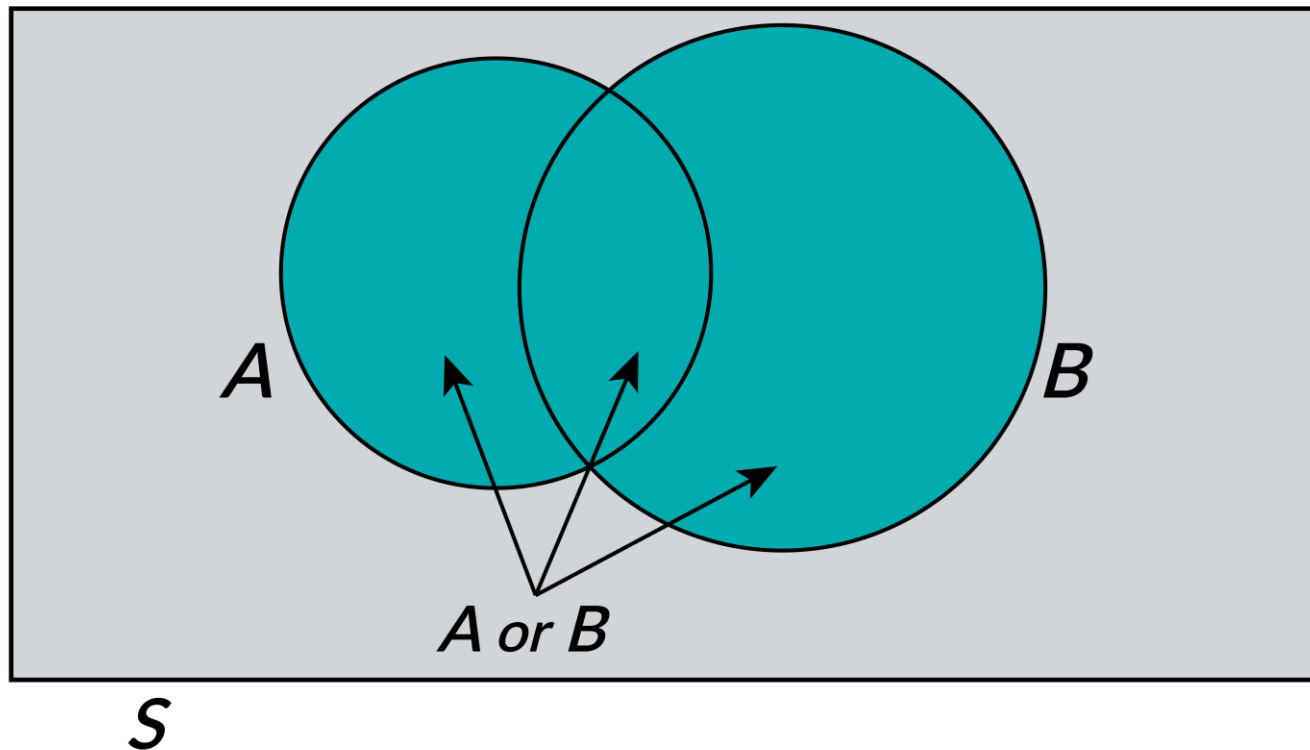
- The sample space  $S$  is shown as a closed figure, labeled  $S$ , and containing all possible sample points. Such graphical representation is called a Venn diagram.
- An event  $A$  belonging to a sample space  $S$  is shown as a round closed figure inside  $S$ .



# Union

---

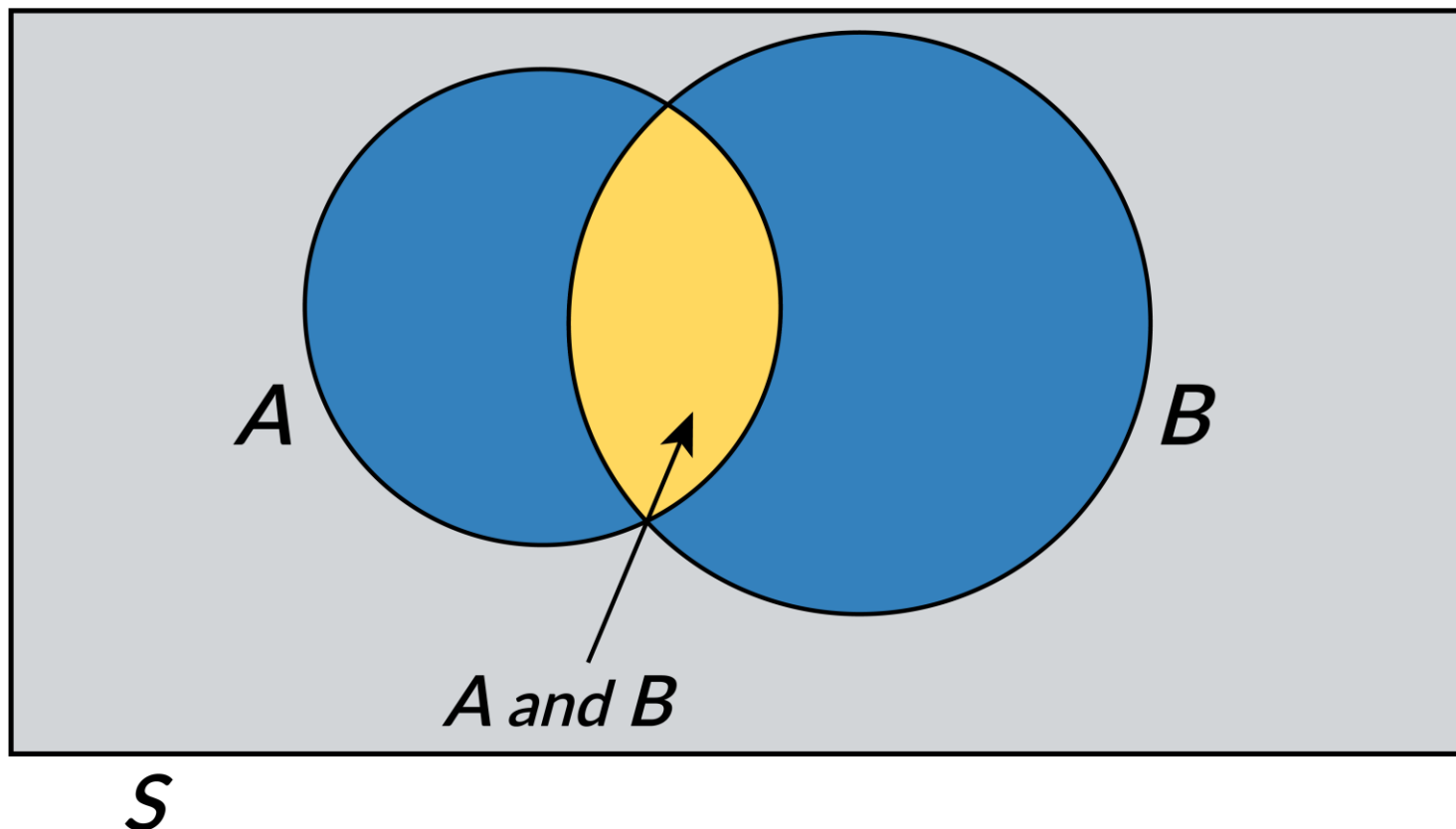
The union of events  $A$  and  $B$  occurs when  $A$  or  $B$  or both occur. The union of the two events is usually denoted by  $A \cup B$  or  $(A \text{ or } B)$ , and consists of sample points that belong to  $A$  or  $B$  or both.



# Intersection

---

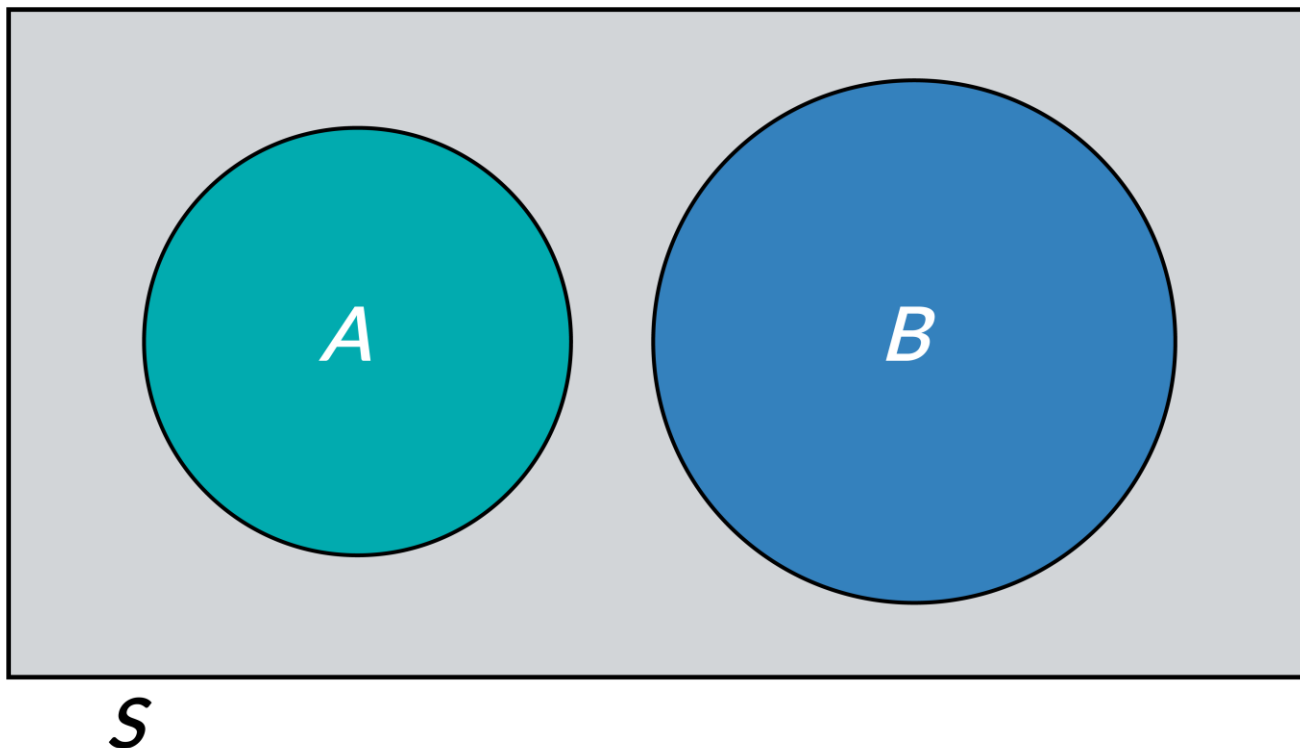
The intersection of two events  $A$  and  $B$ , denoted by  $A \cap B$  or ( $A$  and  $B$ ), occurs if both  $A$  and  $B$  occur simultaneously. ( $A$  and  $B$ ) consists of all sample points belonging to both  $A$  and  $B$ .



# Mutually Exclusive or Disjoint Events

---

Two events  $A$  and  $B$  are said to be mutually exclusive or disjoint if their intersection, contains no sample points, that is, if  $A$  and  $B$  have no sample points in common.



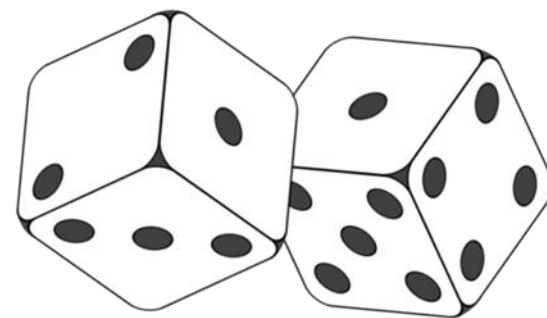
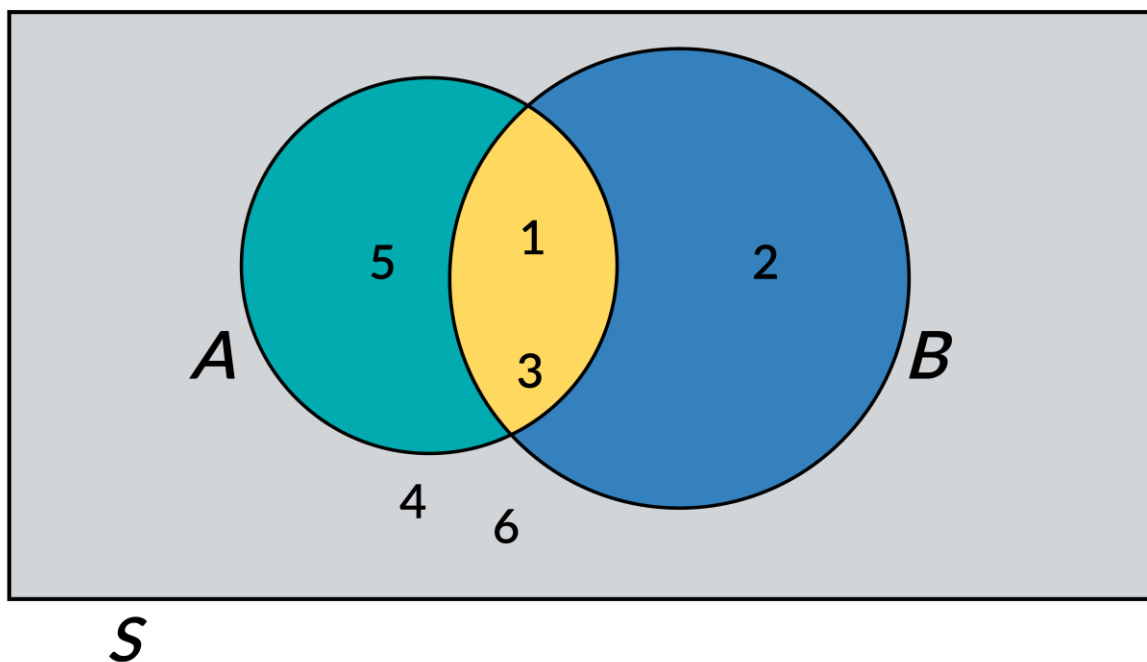
# Example 1

---

Consider a die-toss experiment. Define the following events:

A: Toss an odd number

B: Toss a number less than or equal to 3



# Example 1

---

Describe  $A \cup B$ ,  $A \cap B$ ,  $A'$  and  $B'$ , and find the probability of each one.

Solution:

$$S = \{1, 2, 3, 4, 5, 6\}, A = \{1, 3, 5\} \text{ and } B = \{1, 2, 3\}$$

The union of A and B, is the event that occurs if we observe either an odd number or a number less than or equal to 3 or both on a single throw of the die. We find:

$$A \cup B = \{1, 2, 3, 5\} \longrightarrow P(A \cup B) = P(1) + P(2) + P(3) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{2}{3}$$

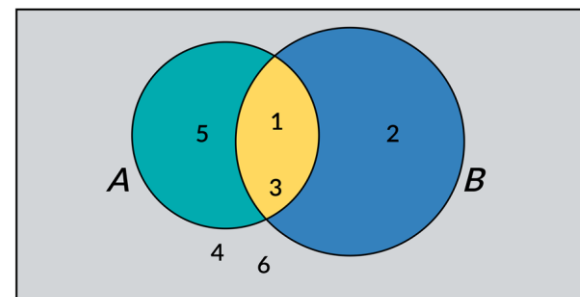
The intersection of A and B, is the event that occurs if we observe both an odd number and a number less than or equal to 3 on a single throw of the die:

$$A \cap B = \{1, 3\} \longrightarrow P(A \cap B) = P(1) + P(3) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

The complements of A and B are given below:

$$A' = \{2, 4, 6\} \longrightarrow P(A') = P(2) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

$$B' = \{4, 5, 6\} \longrightarrow P(B') = P(5) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$



S

Notice that events 4 and 6 are both outside of A and outside of B, thus part of both  $A'$  and  $B'$ .

# Rules of Probability

---

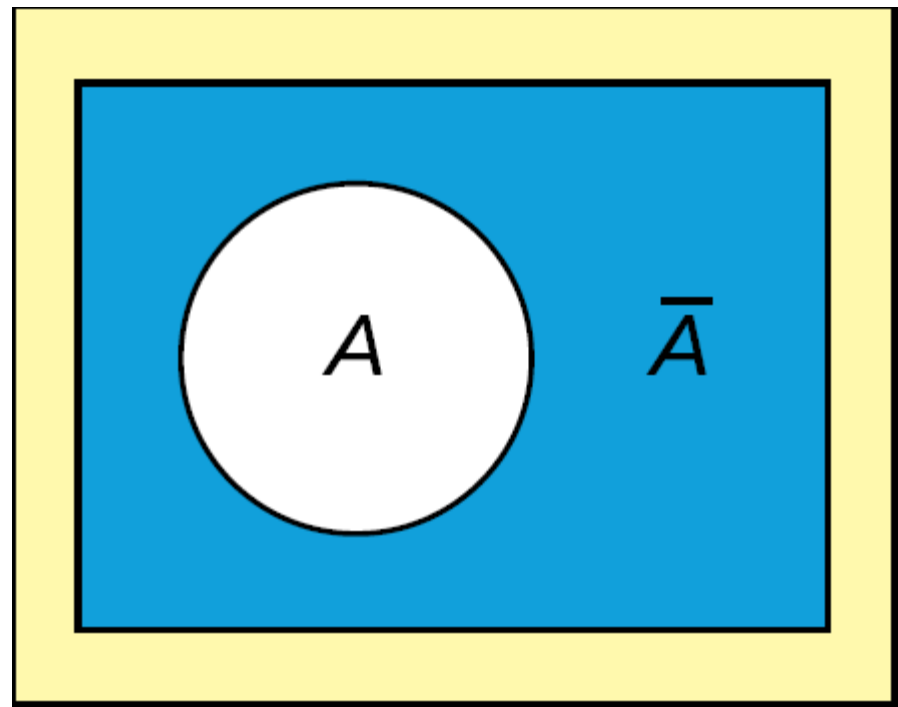
# The Complementary Rule of Probability

---

The sum of the probabilities of complementary events equals 1:  
that is,

$$\Rightarrow P(A) + P(A') = 1$$

$$\Rightarrow P(A) + P(\text{not } A) = 1$$





# The Complementary Rule of Probability

---

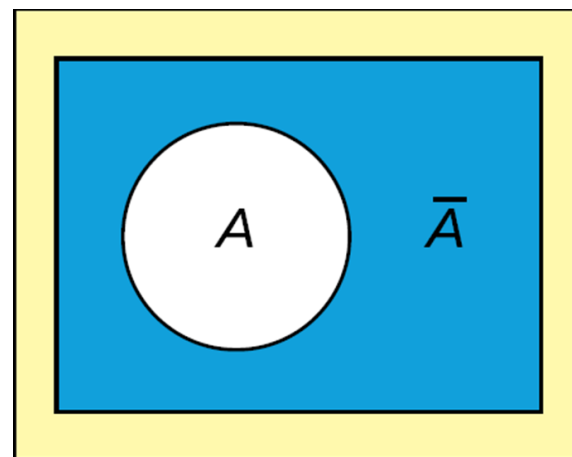
Consider the experiment of tossing a fair coin twice. Use the complementary rule to calculate the probability of event,  $A$ : {observing at least one tail}

Solution:

The complement of  $A$  is defined as the event that occurs when  $A$  does not occur.

Therefore,

- $A' = \{\text{observing no tails}\} = \{HH\} = 1/4$
- $P(A) = 1 - P(A') = 1 - 1/4 = 3/4$



# The Addition Rule of Probability

---

The probability of the union of events  $A$  and  $B$  is the sum of the probability of events  $A$  and  $B$  minus the probability of the intersection of events  $A$  and  $B$ , that is,

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

If two events are mutually exclusive, the probability of their union equals the sum of their respective probabilities:

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

Note: Because  $A$  and  $B$  cannot happen simultaneously,  $P(A \cap B)$  is 0.



# Example 2

---

Consider the experiment of tossing a coin twice. Suppose the coin is **not** balanced and the probabilities of sample points are given in the table below.

Outcome	Probability
HH	0.16
HT	0.24
TH	0.24
TT	0.36

Consider the events

**A:** {Observe exactly one tail}

**B:** {Observe at least one tail}

Calculate the probability of **A**, and the probability of **B**.



# Example 2

---

Solution:

Event **A** contains the sample points *HT* and *TH*.

We can calculate the probability of event **A** by summing the probabilities of its two sample points:

$$P(A) = P(HT) + P(TH) = 0.24 + 0.24 = 0.48$$

Similarly, since **B** contains the sample points *HT*, *TH*, and *TT*, and

$$P(B) = P(HT) + P(TH) + P(TT) = 0.24 + 0.24 + 0.36 = 0.84$$



# Example 3

---

The following table describes the income of the adult population of a small suburb of a southern city:

Income				
	<\$25,000	\$25,000-\$50,000	>\$50,000	
AGE				Total
<25	200	400	300	900
25-45	300	100	100	500
>45	100	500	600	1200
Total	600	1000	1000	2600

Consider the following events:

**A** : The person is between 25 and 45

**B** : The person has income less than \$25,000

Calculate the probability of **A**. Then calculate the probability of **B**.



# Example 3

---

## Solution:

The suburb has a total adult population of 2600. There are 500 adults who are between 25 and 45.

Therefore  $P(A) = 500/2600 = 0.192$ .

Similarly, there are 600 adults whose income is less than \$25000.

Therefore,  $P(B) = 600/2600 = 0.231$ .



# Example 4

---

- Hospital records show that 15% of all female patients are admitted for surgical treatment, 25% are admitted for obstetrics, and 5% receive both obstetrics and surgical treatments. If a new female patient is admitted to the hospital, what is the probability that the patient will be admitted either for surgery, obstetrics, or both?

**Solution:**

A: {A female patient admitted to the hospital receives surgical treatment}

B: {A female patient admitted to the hospital receives obstetrics treatment}

Then, from the given information,  $P(A) = 0.15$ ,  $P(B) = 0.25$ , and  $P(A \cap B) = 0.05$

Therefore,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.15 + 0.25 - 0.05 = 0.35$$

Thus, 35% of all female patients admitted to the hospital receive either surgical treatment, obstetrics treatment, or both.



# Conditional Probability

---

The event probabilities we have been discussing so far are often called unconditional probabilities since no special conditions other than those that define the experiment are assumed. Sometimes, on the other hand, we may have additional knowledge that might alter the probability of an event. A probability that reflects such additional knowledge is called the conditional probability of the event.

**We represent the probability of event A, given that event B occurs by the symbol  $P(A | B)$  (it reads: the probability of A condition B) and is given by:**

- $P(A|B)=(P(A \cap B))/(P(B))$  (we call this formula 1)
- $P(B|A)=(P(A \cap B))/(P(A))$  (we call this formula 2)





# The Multiplication Rule of Probability

---

Formulas (1) and (2), after cross multiplication, can be written as

$$P(A \cap B) = P(A | B) P(B) \quad (\text{we call this formula 3})$$

$$P(A \cap B) = P(B | A) P(A) \quad (\text{we call this formula 4})$$

# Independent Events

---

- Two events A and B are said to be independent if the outcome of one does not influence the outcome of the other. Mathematically, events A and B are independent if and only if
  - $P(A|B)=P(A)$
- This criterion means that the occurrence of event B does not affect the occurrence of event A.
- The "if and only if" statement implies that if events A and B are independent, then  $P(A|B)=P(A)$  ; and
  - conversely, if ,  $P(A|B)=P(A)$  then events A and B are independent.
- Events that are not independent are said to be dependent.



# Example 5

---

A manufacturer of an electromechanical kitchen utensil conducted an analysis of a large number of consumer complaints and found that they fell into six categories shown in the table below:

	The Reason of Complaint		
The Time of Complaint	Electrical	Mechanical	Appearance
During guarantee period	20%	15%	30%
After guarantee period	10%	15%	10%

Define the following events:

A: {Cause of complaint is product appearance},

B: {Complaint occurred during the guarantee period}.

Are A and B independent events?

# Example 5

---

## Solution:

Events A and B will be independent when  $P(A \cap B) = P(A) P(B)$ . Otherwise, they will be dependent events.

$$P(A \cap B) = 0.30,$$

$$P(A) = 0.30 + 0.10 = 0.40$$

$$P(B) = 0.20 + 0.15 + 0.30 = 0.65 \longrightarrow P(A) P(B) = (0.40)(0.65) = 0.26$$

Since  $(A \cap B) \neq P(A) P(B)$ , we conclude the A and B are dependent events.



# Random Variables

---

- **Random variable** is a variable whose value is determined as a consequence of a random experiment.
- Suppose that a banker is interested in knowing the number of customers using an ATM machine in a given day. Since this number may vary from one day to another, and it isn't possible to say with certainty what it will be, then it's a random variable
- There are two types of random variables:
  - A **discrete random variable** is a random variable that can assume only certain and clearly separated (discrete) values.
  - A **continuous random variable**, on the other hand, may assume any value in a given range or interval, and there is a continuity of the different possible values it may take.



# Random Variables

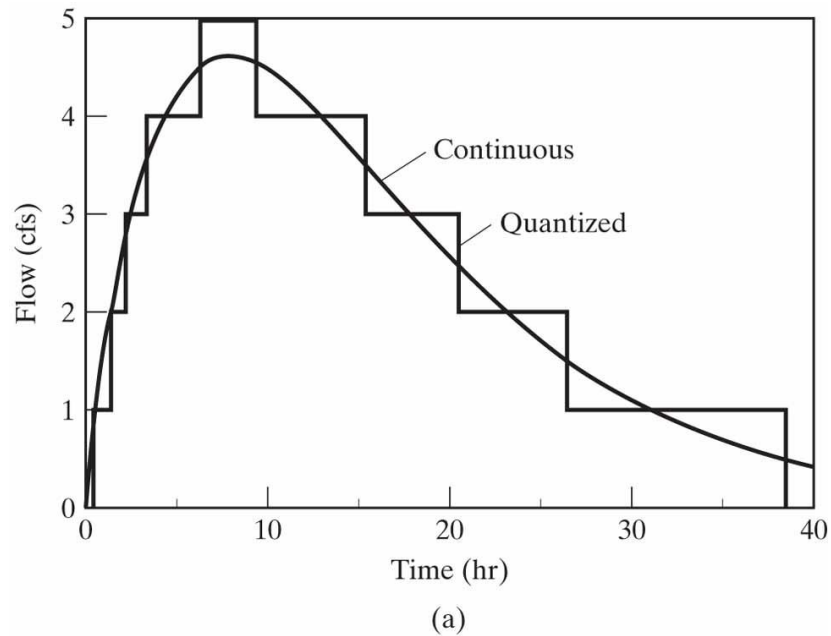
---

- **Random variable** is a variable whose value is determined as a consequence of a random experiment.
- Suppose that a banker is interested in knowing the number of customers using an ATM machine in a given day. Since this number may vary from one day to another, and it isn't possible to say with certainty what it will be, then it's a random variable
- There are two types of random variables:
  - A **discrete random variable** is a random variable that can assume only certain and clearly separated (discrete) values.
  - A **continuous random variable**, on the other hand, may assume any value in a given range or interval, and there is a continuity of the different possible values it may take.

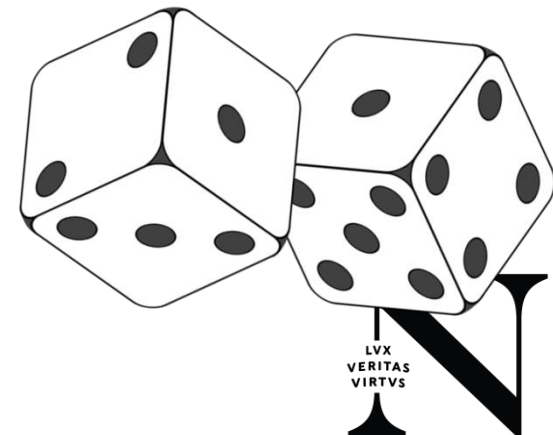
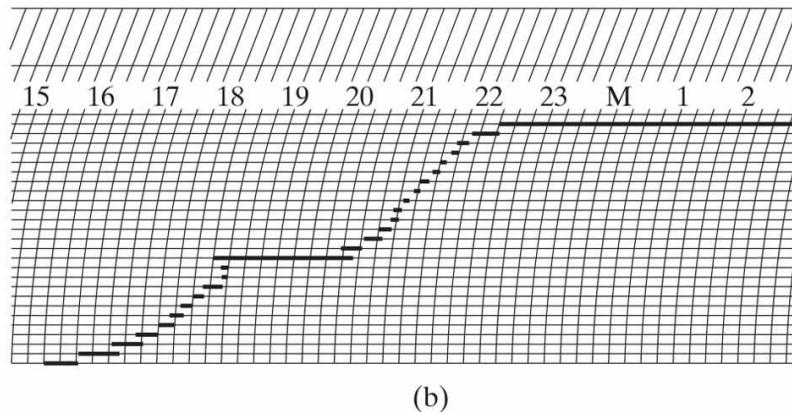


# Random Variables

Continuous and quantized data



Discrete data



# Random Variables

---

## ▪ Discrete Variable Examples:

- The score of a baseball team in a game.
- The number of tails observed when a coin is tossed twice.
- The average of grades in a test.
- The number of bank customers using an ATM machine in a given day.

## Continuous Variable Examples:

- The amount of soda in a randomly chosen 12-ounce can of a particular brand.
- The amount of time it takes a randomly chosen runner to run a mile.
- The area of a randomly drawn circle.
- The amount of rainfall in a randomly chosen summer day in New Orleans.



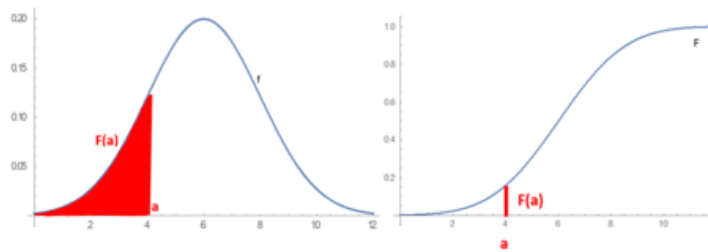


# Probability Distributions

---

*In probability theory and statistics, a probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment-Wikipedia*

- A distribution reflects the values in a data set and how often those values occur.
- Distributions can be used to either describe or generate data.
- This allows us to describe the data. For example, we can say: this data follows a normal distribution with specific parameters, in this case, mean and standard deviation.



# Probability Distributions

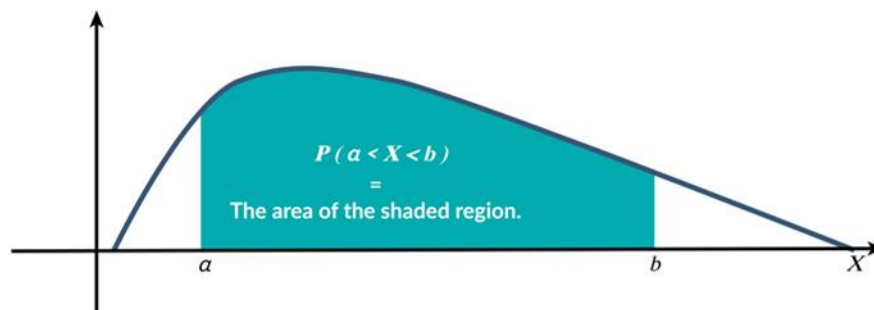
---

- **Discrete Probability Distributions:** It describes the probability of occurrence of each value of a discrete random variable.
- Also known as Probability mass function (PMF)
- Example Distributions:
  - Binomial Distribution
  - Poisson Distribution
  - Bernoulli distributions

## Continuous Probability Distributions:

A probability distribution in which the random variable  $X$  can take on any value (is continuous).

- Also known as probability density function (PDF)
- Example Distributions:
  - Normal Distribution
  - Gamma distribution
  - Logistic distribution
  - Weibull distribution

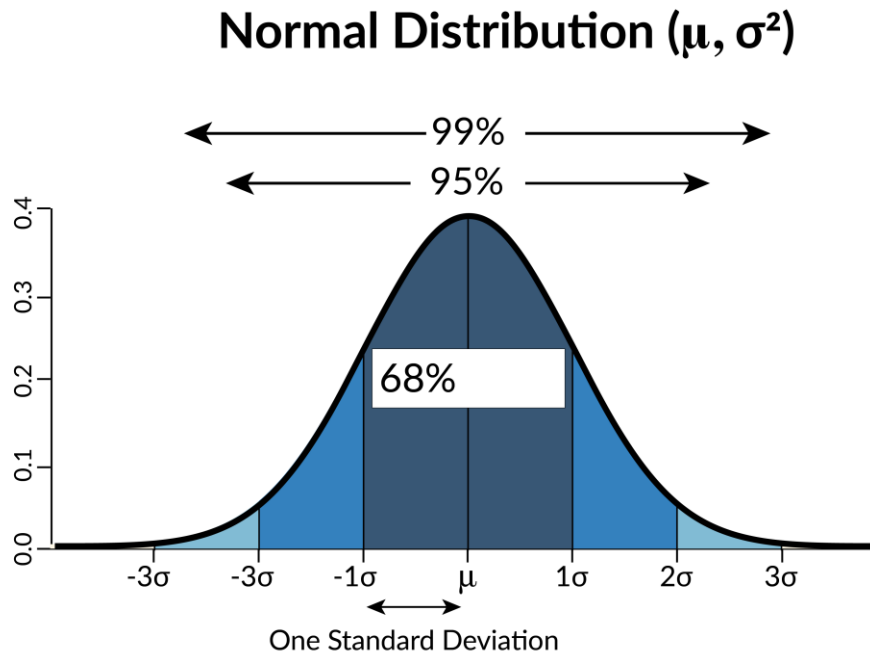


# Normal Probability Distribution

---

- A special type of a continuous probability distribution is the normal probability distribution.

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



# Normal Probability Distribution

---

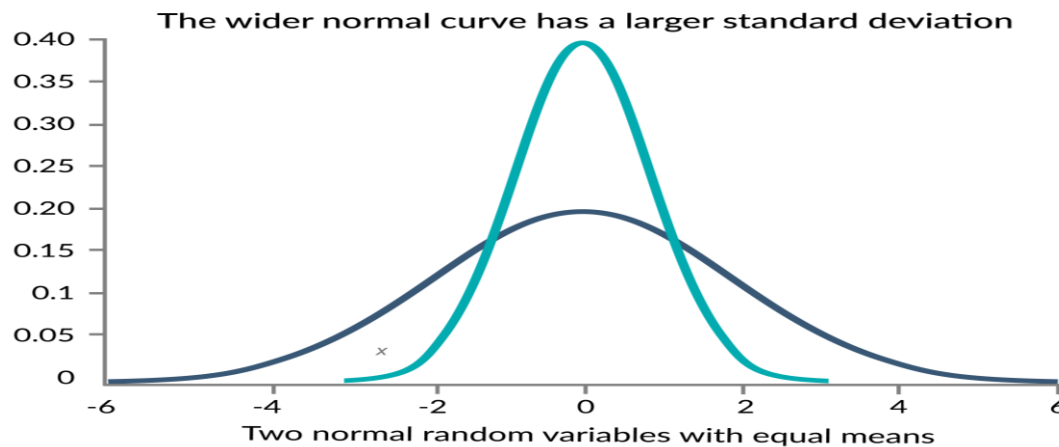
- A normal probability distribution is mound-shaped and symmetric and has a single peak at the middle of the distribution, at which the mean and the median and the mode coincide.
- Half the area under the curve is above this center point and the other half is below it and the two halves of the curve are mirror images of each other.
- Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.



# Normal Probability Distribution

---

- In the figure below, two normal distributions with equal means but different standard deviations are sketched. The wider distribution possesses the larger standard deviation. Since the total area under each normal curve must be 1 unit, the narrower distribution is taller.
- An infinite number of different normal distributions that possess different means and standard deviations



# Standard Normal Distribution

---

- The standard normal distribution is a special normal distribution where the mean is 0 and the standard deviation is 1.
- It is also called the z-distribution

$$z = \frac{x - \mu}{\sigma}$$

$$x = \sigma z + \mu$$

- The curve is symmetric with respect to the vertical line passing through  $z = 0$
- The total area under the curve equals one



# Probability Distributions

---

- **Discrete Probability Distributions:** It describes the probability of occurrence of each value of a discrete random variable.
- Also known as Probability mass function (PMF)
- Example Distributions:
  - Binomial Distribution ,
  - Poisson Distribution , and
  - Bernoulli distributions



# Example 6

---

Toss a coin twice and let  $X$  be the number of tails observed. Construct a probability distribution for  $X$ .

**Solution:**

Recall that the sample space of this experiment is:  $S = \{HH, HT, TH, TT\}$ . As far as the variable  $X$  is concerned, the possible outcomes are: either no tail will be observed ( $X = 0$ ), or only one tail will be observed ( $X = 1$ ), or 2 tails will be observed ( $X = 2$ ). Note that  $X$  is a discrete r.v. taking separated values. Next, we will assign probabilities to each possible value of  $X$ :

The outcome of  $HH$  corresponds to the case when  $X = 0$ . Therefore, the probability that  $X = 0$ , denoted by  $P(X = 0)$  or simply  $P(0)$ , is given by:

$$P(X=0)=P(HH)=1/4=0.25$$

Similarly,

$$P(X=1)=P(HT,TH)=1/4+1/4=0.5$$

$$P(X=2)=P(TT)=1/4=0.25$$

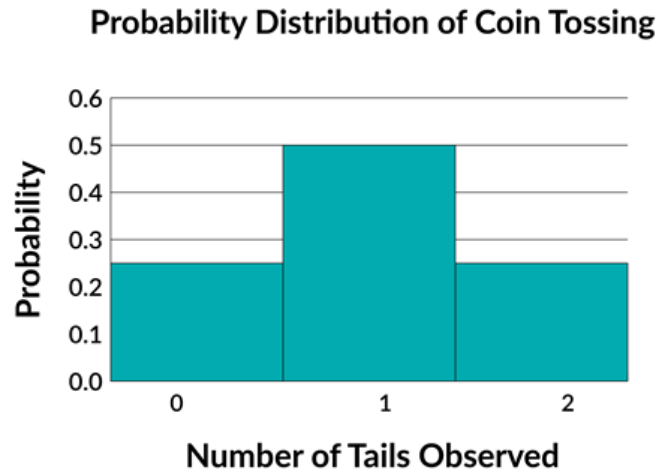
Number of Trials of $x$	Probability of Outcome $P(x)$
0	0.25
1	0.50
2	0.25





# Discrete Probability Distributions

---



- The probability of a particular value is between 0 and 1, inclusive. That is,  $0 \leq P(x) \leq 1$
- The sum of the probabilities of all values is 1. That is,  $\sum P(x) = 1$

# Discrete Probability Distributions

---

The probability distribution of a discrete r.v.  $X$  is given below.

$x$	$P(x)$
-3	0.15
0	0.05
2	0.40
4	0.15
5	0.10
6	0.05
8	0.10

Based on the above table, let's find the following probabilities:

- A.  $P(X = 4)$
- B.  $P(X > 5)$
- C.  $P(X < 2)$



# Discrete Probability Distributions

---

Solution:

- A. To find  $P(4)$ , we must use the fact that all probabilities must add up to 1 (the second characteristic of a discrete probability distribution).  $P(4) = 1 - (0.15 + 0.05 + 0.40 + 0.10 + 0.05 + 0.10) = 0.15$
- B.  $P(X > 5) = P(6) + P(7) = 0.05 + 0.10 = 0.15$
- C.  $P(X < 2) = P(-3) + P(0) = 0.15 + 0.05 = 0.20$
-

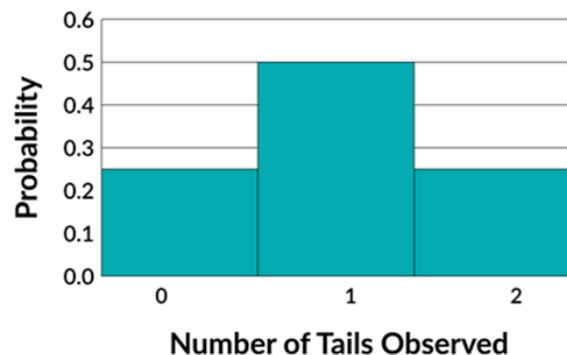
# Cumulative Probability Distribution

- For any value  $x$  of a random variable  $X$ , the cumulative probability is given by  $P(X \leq x)$ .

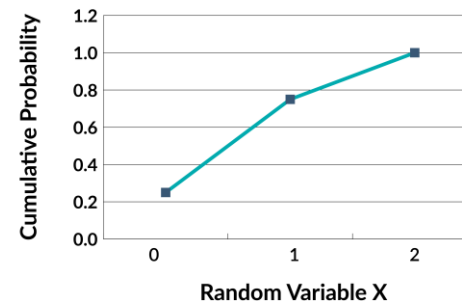
Number of Trials of $x$	Probability of Outcome $P(x)$
0	0.25
1	0.50
2	0.25

$x$	$P(X \leq x)$
0	0.25
1	0.75
2	1

Probability Distribution of Coin Tossing



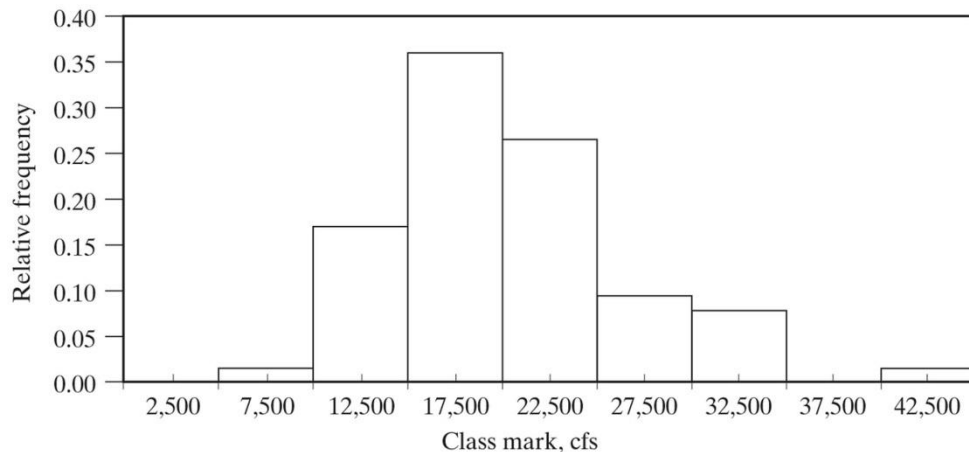
Cumulative Probability Distribution



# Cumulative Probability Distribution

---

- For any value  $x$  of a random variable  $X$ , the cumulative probability is given by  $P(X \leq x)$ .



	Probability
P(2500)	0.000
P(7500)	0.013
P(12500)	0.173
P(17500)	0.360
P(22500)	0.267
P(27500)	0.093
P(32500)	0.080
P(37500)	0.000
P(42500)	0.013

## Cumulative distribution function (CDF)

$$F(22,500) = 0.000 + 0.013 + 0.173 + 0.360 + 0.267 = 0.813 \text{ or } 81.3\%$$



# Poisson Distributions

---

The Poisson distribution is a discrete distribution which applies when we want to calculate the probability that an event will occur a given number of times in a given interval.

The Poisson distribution is defined mathematically as:

$k$ : number of times event occurs in interval, integer

$\lambda$ : average number of event occurrences in interval,  $\lambda \geq 0$

$X \sim \text{Poisson}(\lambda)$ : random variable  $X$  follows Poisson distribution

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

$$E[X] = \lambda$$



# Poisson Distributions

---

There are some conditions for which the Poisson distribution applies:

- Each event is independent of the others
- The rate of occurrence is constant (does not vary over time)
- Two events cannot occur at the same instant
- The probability of an event occurring in an interval is proportional to the length of that interval

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



# Poisson Distributions

---

- A pizza shop receives on average 12 pizza order per day (a) what is the probability of that business will receive exactly 8 order per day?
- A typical Facebook user received 7 friend request per day. What is the probability that he will receive exactly 9 friend request per day?





# Binomial Probability

---

- Binomial probability refers to the probability of exactly  $x$  successes on  $n$  repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment).
- **What is the probability of getting 6 heads, when you toss a coin 10 times?**
- The probability that the binomial r.v.  $X$  takes the value  $x$  is denoted by  $P(X = x)$ , or simply by  $P(x)$ , and is given by:

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



# Binomial Probability

---

- According to a recent survey, 60% of adult Charlotte residents are in favor of the new zoning regulations introduced by the city council. A random sample of 15 residents is chosen. Let  $X$  be the number of residents in the sample who are in favor of the new zoning regulations. Describe why  $X$  is a binomial random variable.

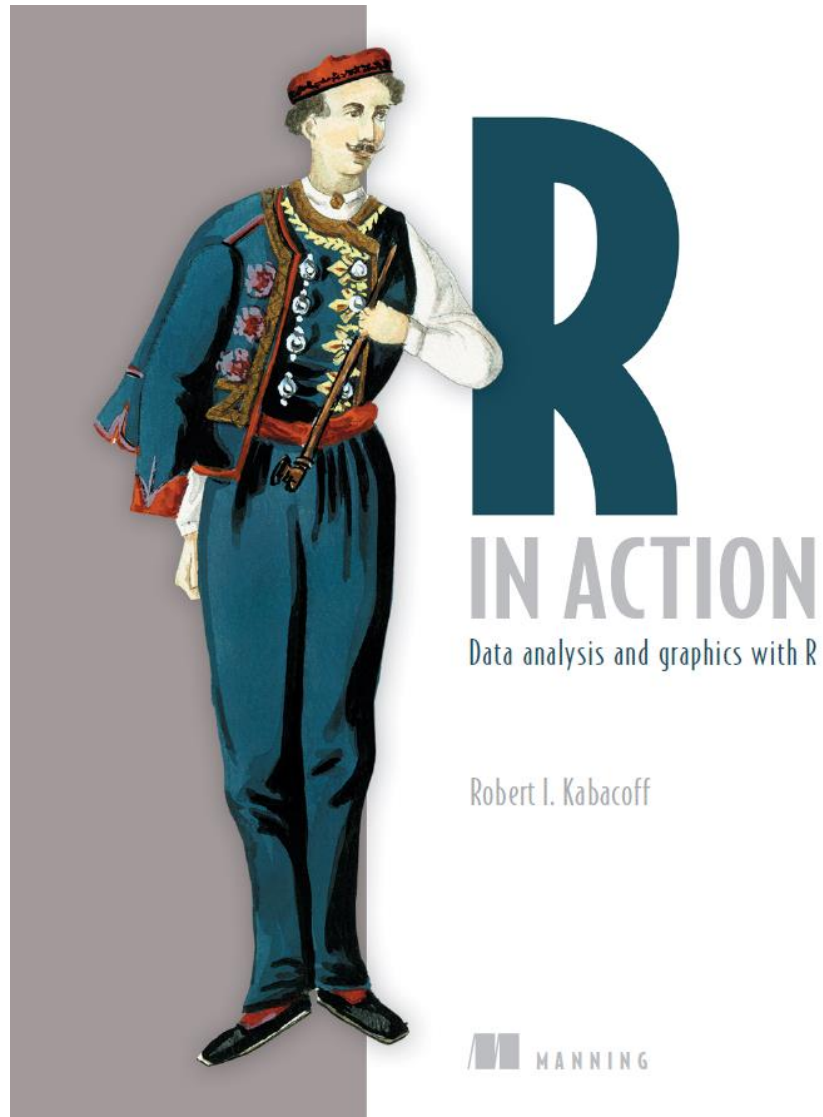
## Solution:

The experiment consists of  $n = 15$  identical inquiries from the residents about their opinion of the new zoning regulations. Since any resident's opinion does not influence another's, then the trials are independent. Each inquiry may result in one of the two possible outcomes: either a randomly chosen resident is in favor of the new regulations (success) or he/she is against them (failure). Since 60% of the population is known to be in favor, then the probability of success is  $p = 0.6$ , and this probability stays the same from trial to trial. The probability of failure is  $q = 0.4$ . The number of residents among the 15 who are in favor is the binomial random variable  $X$ .



# R in Action

---



# Summary

---

- Reviewed module 2 and related topics
- Introduced Probability
- Solve probability problems using addition and multiplication rules of probability
- Introduced various Probability distributions



Q & A

---