Statistical and Mathematical Methods for Data Analysis

Dr. Syed Faisal Bukhari

Associate Professor

Department of Data Science

Faculty of Computing and Information Technology

University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- **Elementary Statistics**, 13th Edition, Mario F. Triola

Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- **Probability Demystified**, Allan G. Bluman
- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- **Python for Probability, Statistics, and Machine Learning,** José Unpingco
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- **Think Stats: Probability and Statistics for Programmers,** Allen Downey

References

Readings for these lecture notes:

- **Probability & Statistics for Engineers & Scientists**, Ninth edition, Ronald E. Walpole, Raymond H. Myer
- **Probability Demystified**, Allan G. Bluman
- https://en.wikipedia.org/wiki/Law_of_large_numbers

These notes contain material from the above three resources.

Distribution of points

Midterm = 30 points

Final term = 40 points

Sessional points = 30 points

I. Assignments = 2×4 = 10 points

II. Hands-on Python in class = 0.5×5 = 2.5 points

III. Quizzes = 0.5×5 = 2.5 points

IV. Journal/conference paper presentation = 5

V. Mini project (its report should be in an IEEE journal paper format) = 10 points

“There is only one thing that makes a dream impossible to achieve: the fear of failure.”

— Paulo Coelho, The Alchemist

Basic concepts

The **probability** of an event A is the sum of the weights of all **sample points** in A.

Therefore,

I. $0 \leq P(A) \leq 1$

II. $P(\varphi) = 0$

III. $P(S) = 1.$

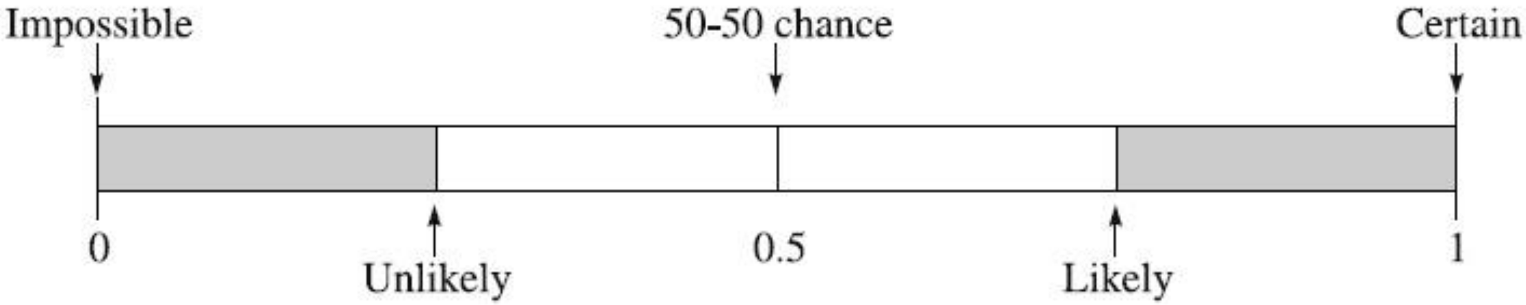
Basic concepts

- When the probability of an event is close to **zero**, the occurrence of the event is relatively **unlikely**. For example, if the chances that you will win a certain lottery are **0.00l** or one in one thousand, you probably won’t win, unless of course, you are very **“lucky.”**

- When the probability of an event is **0.5** or $\mathbf{^1/_2}$, there is a **50–50 chance** that the event will happen—the same.

Basic concepts

When the probability of an event is close to one, the event is almost sure to occur. For example, if the chance of it snowing tomorrow is **90%,** more than **likely**, you'll see some snow.



Empirical Probability [1]

Probabilities can be computed for situations that do not use sample spaces. In such cases, frequency distributions are used and the probability is called **empirical probability**.

Rank	Frequency
Freshmen	4
Sophomores	6
Juniors	8
Seniors	7
TOTAL	25

Empirical Probability [2]

$$P(E) = \frac{\text{Frequency of E}}{\text{Sum of the frequencies}}$$

$P(E) = {}^{1}/_{4}$

Empirical probability is sometimes called **relative frequency probability**.

Law of large numbers

- In probability theory, the **law of large numbers (LLN)** is a theorem that describes the **result** of performing the **same experiment a large number of times**.

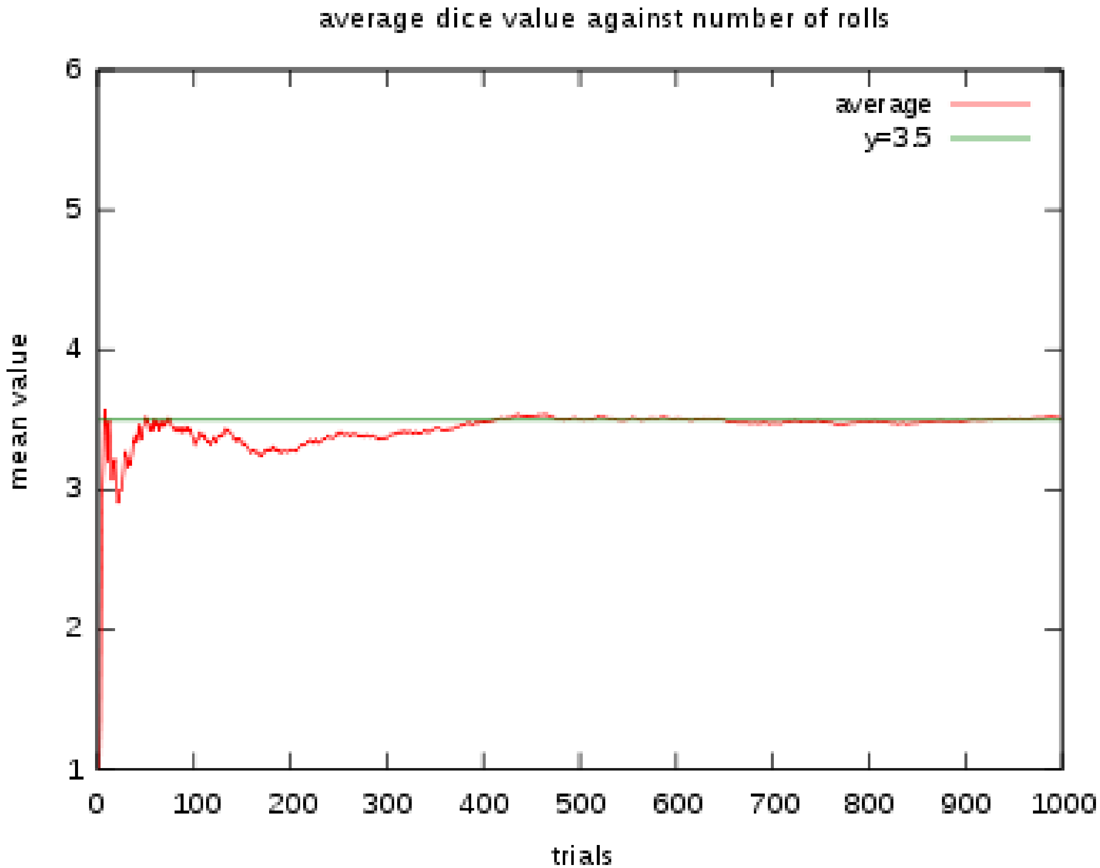
- According to the law, the **average** of the results obtained from a **large number of trials** should be close to the **expected value**, and will tend to become **closer** as more trials are performed.

Law of large numbers

- The LLN is important because it **"guarantees" stable long-term** results for the averages of some random events.

- **For example**, while a casino may lose money in a single spin of the **roulette** wheel, its earnings will tend towards a predictable percentage over a large number of spins.

Out come of a die
1
2
3
4
5
6
$\sum x = 21$
$\bar{x} = \frac{\sum x}{n} = \frac{21}{6} = 3.5$



An illustration of the law of large numbers using a particular run of rolls of a single die. As the number of rolls in this run increases, the **average** of the values of all the results approaches **3.5**.

Law of Large Numbers

Questions:

What happens if we toss the coin **100 times** ? Will we get **50** heads?

What will happen if we toss a coin **1000 times**? Will we get exactly **500** heads?

Law of Large Numbers

- **Solution:** Probably not.

- However, as the number of **tosses increases**, the ratio of the number of heads to the total number of tosses will get closer to $^{1}/_{2}$.

- This phenomenon is known as the **law of large numbers**.

Subjective Probability

A third type of probability is called **subjective probability**. Subjective probability is based upon an **educated guess**, **estimate**, **opinion**, or **inexact information**.

Sample Spaces

There are two **specific devices** that will be used to find sample spaces for probability experiments. They are **tree diagrams** and **tables**.

A **tree diagram** consists of **branches** corresponding to the outcomes of two or more probability experiments that are done in sequence.

Sample Spaces

- In order to construct a tree diagram, use branches corresponding to the outcomes of the **first experiment**. These branches will emanate from a single point.

- Then from each branch of the first experiment draw branches that represent the outcomes of the **second experiment**.

- You can continue the process for further experiments of the sequence if necessary.

Tree Diagram [1]

Example: A **coin** is tossed and a **die** is rolled. Draw a tree diagram and find the sample space.

Solution:

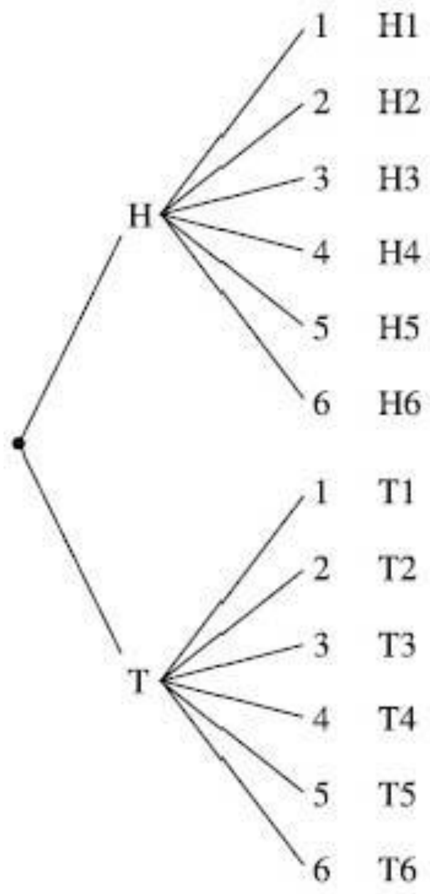
Since there are two outcomes (heads and tails for the coin), draw two branches from a single point and label one **H** for head and the other one **T** for tail.

From each one of these outcomes, draw and label six branches representing the outcomes **1, 2, 3, 4, 5,** and **6** for the die.

Trace through each branch to find the outcomes of the experiment.

Tree Diagram [2]

Example: A coin is tossed and a die is rolled. Draw a tree diagram and find the sample space.



Tree Diagram [3]

Example: A coin is tossed and a die is rolled. Find the probability of getting

a. A head on the coin and a 3 on the die.

b. A head on the coin.

c. A 4 on the die.

Solution:

a. P(H3) = ${}^{1}/_{12}$ = **0.0833 (or 8.33%)**

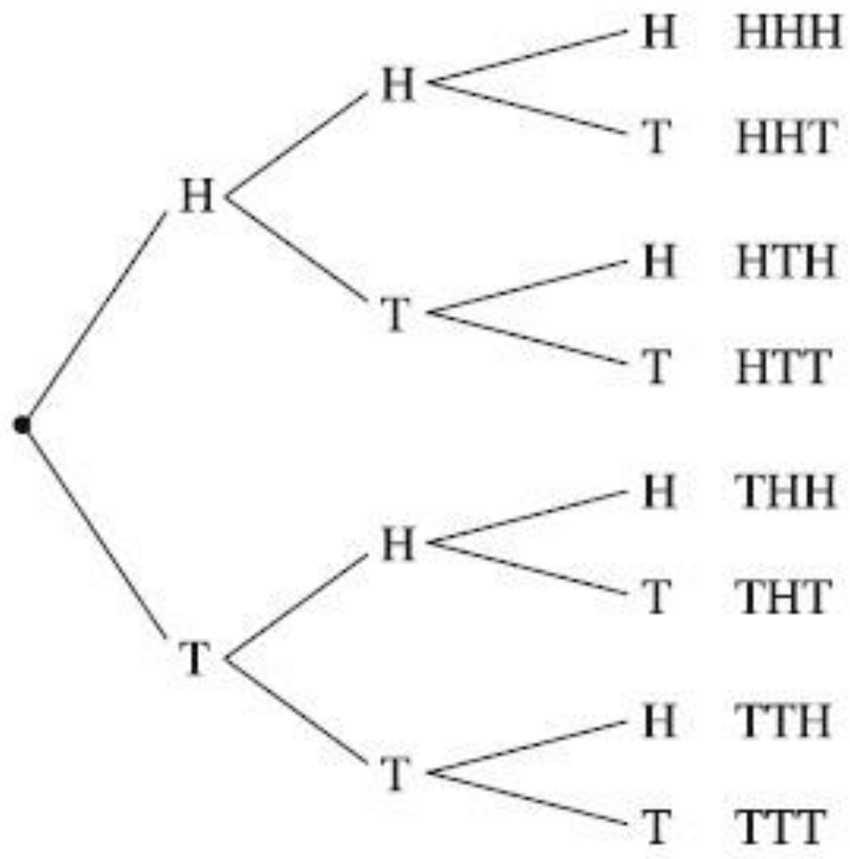
b. P(head on the coin) = ${}^{6}/_{12}$ = ${}^{1}/_{2}$ = **0.5 (or 50%)**

c. P(4 on the die) = ${}^{2}/_{12}$ = ${}^{1}/_{6}$ = **0.1667 (16.67%)**

Tree Diagram [4]

Example: Three coins are tossed. Draw a tree diagram and find the sample space.

Solution



Tree Diagram [5]

Example: Three coins are tossed. Find the probability of getting

a. Two heads and a tail in any order.

b. Three heads.

c. No heads.

d. **At least** two tails.

e. **At most** two tails.

Solution:

a. P(2 heads and a tail) = 3/8 = **0.375 (or 37.5 %)**

b. P(HHH) = $^1/_8$ = **0.125 (or 12.5 %)**

c. P(TTT) = $^1/_8$ = **0.125 (or 12.5 %)**

d. P(at least two tails) = $^4/_8$ = $^1/_2$ = **0.5 (or 50 %)**

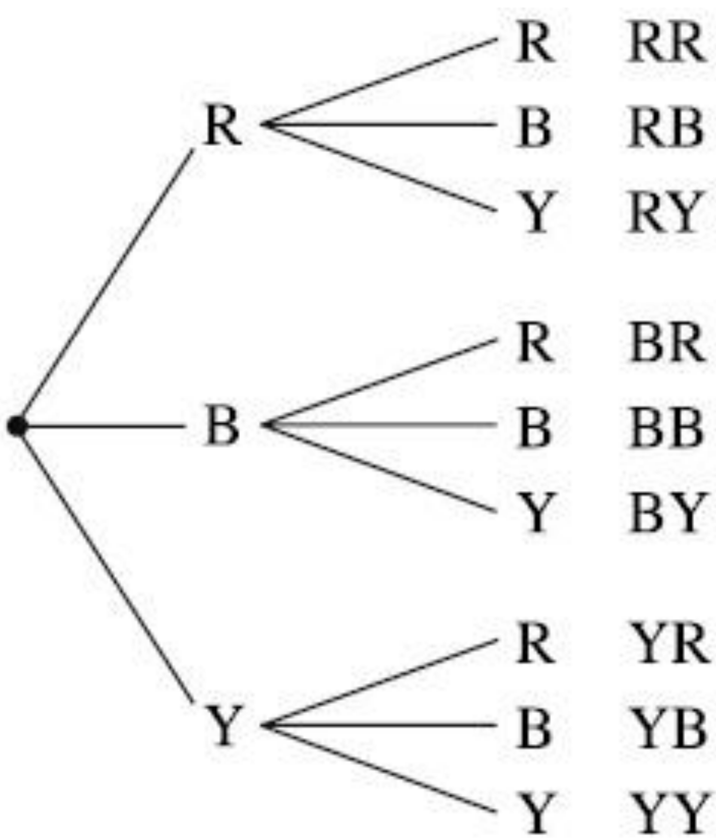
e. P(at most two tails) = $^7/_8$ = **0.875 (or 87.5 %)**

$\Rightarrow \boldsymbol{At\ most\ two\ tails\ mean\ no\ three\ tails}$

Tree Diagram [6]

Example: A box contains a **red ball (R),** a **blue ball (B),** and a **yellow ball (Y**). **Two balls** are selected at random in succession. Draw a **tree diagram** and find the sample space if the first ball is **replaced** before the **second ball** is selected.

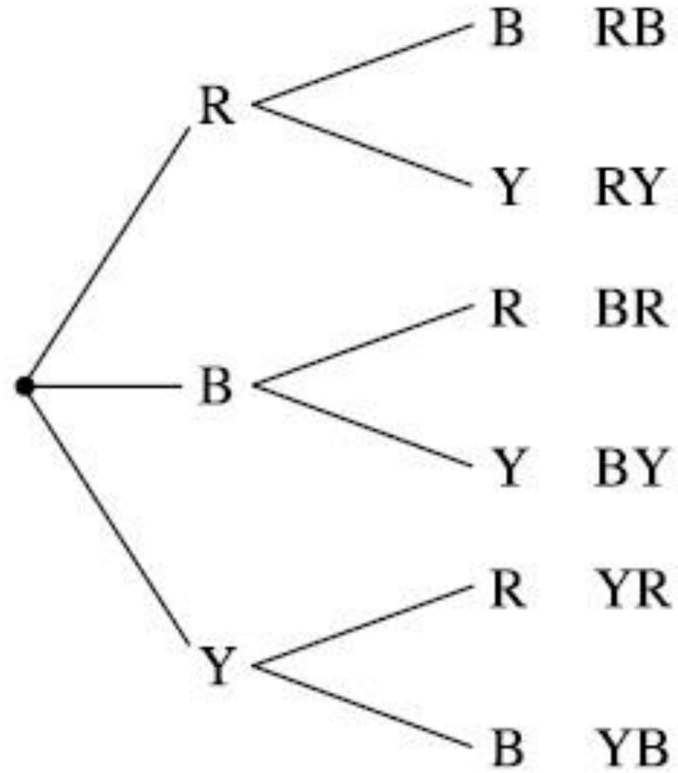
Solution



Tree Diagram [7]

Example: A box contains a **red ball (R),** a **blue ball (B)**, and a **yellow ball (Y)**. Two balls are selected at random in **succession**. Draw a **tree diagram** and find the sample space if the first ball is **not replaced** before the second ball is selected.

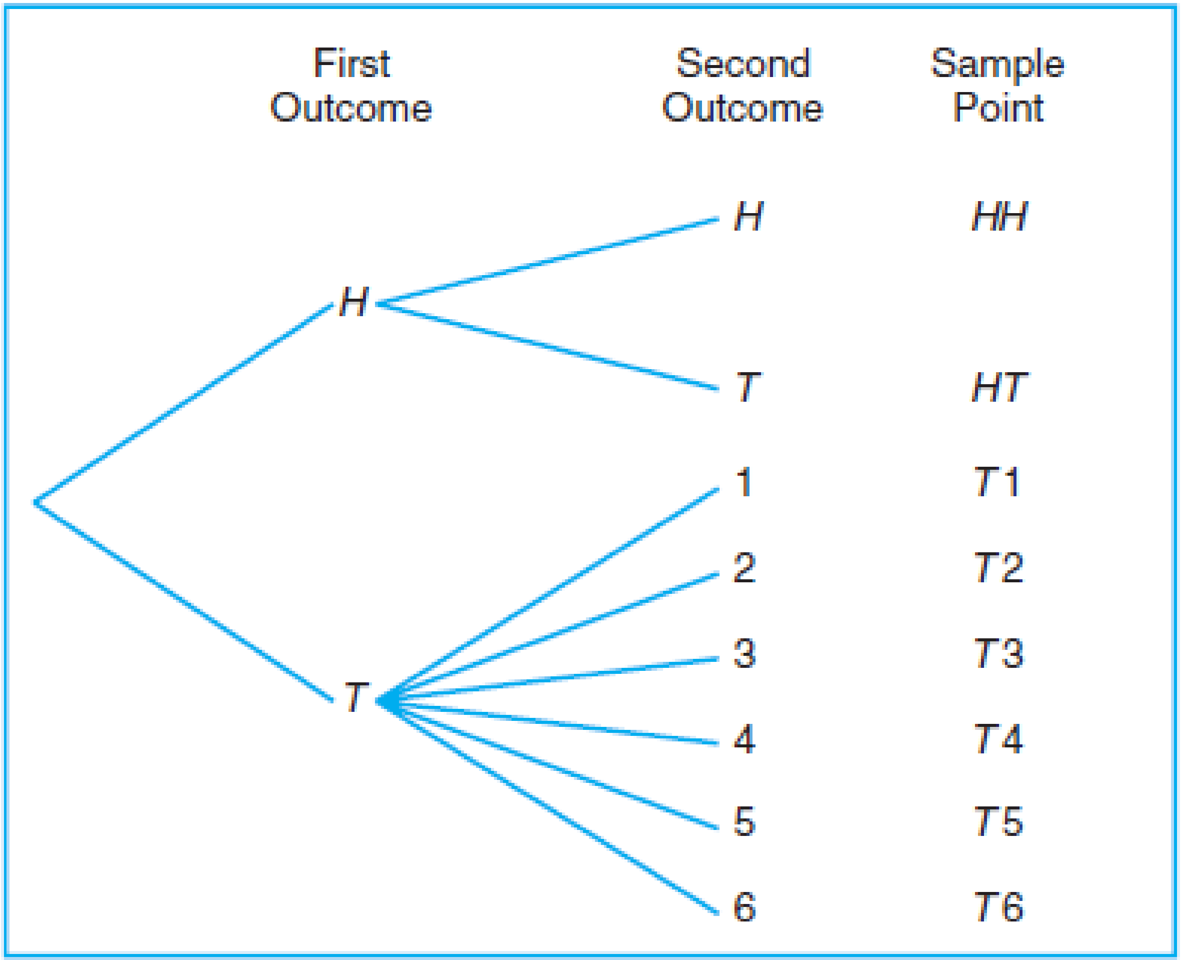
Solutions



Tree Diagram [9]

Example An experiment consists of **flipping a coin** and then flipping it a **second time** if a **head occurs**. If a **tail** occurs on the **first flip**, then a **die is tossed** once. To list the elements of the sample space providing the most information, construct the **tree diagram**

Solution



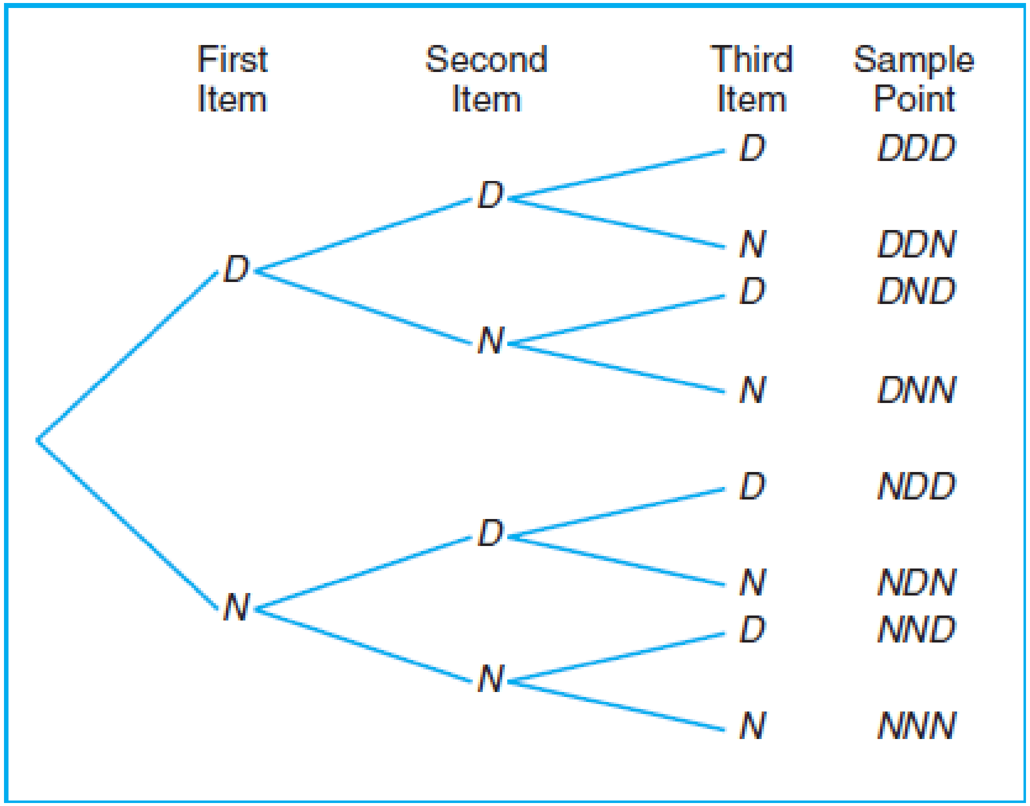
First Outcome
Second Outcome
Sample Point
H
T
H
T
1
2
3
4
5
6
HH
HT
T1
T2
T3
T4
T5
T6

Tree Diagram [10]

Example Suppose that **three items** are selected at random from a manufacturing process. Each item is inspected and classified **defective**, **D**, or **nondefective**, **N**. To list the elements of the sample space providing the most information, construct the **tree diagram.**

Solution

- S = {HH, HT, T1, T2, T3, T4, T5, T6}.



Tables [1]

Another way to find a sample space is to use a **table.**

Example: Find the sample space for selecting a card from a standard deck of 52 cards.

There are four suits—**hearts** and **diamonds**, which are **red**, and **spades** and **clubs**, which are **black**. Each suit consists of 13 cards—ace through king. Face cards are kings, queens, and jacks.

A standard deck of 52 cards

Heart	A ♥	2 ♥	3 ♥	4 ♥	5 ♥	6 ♥	7 ♥	8 ♥	9 ♥	10 ♥	J ♥	Q ♥	K ♥
Diamond	A ♦	2 ♦	3 ♦	4 ♦	5 ♦	6 ♦	7 ♦	8 ♦	9 ♦	10 ♦	J ♦	Q ♦	K ♦
Spade	A ♠	2 ♠	3 ♠	4 ♠	5 ♠	6 ♠	7 ♠	8 ♠	9 ♠	10 ♠	J ♠	Q ♠	K ♠
Club	A ♣	2 ♣	3 ♣	4 ♣	5 ♣	6 ♣	7 ♣	8 ♣	9 ♣	10 ♣	J ♣	Q ♣	K ♣

Tables [2]

Example: A single card is drawn at random from a standard deck of cards. Find the probability that it is

a. The 4 of diamonds.

b. A queen.

Solution:

a. P(The 4 of diamonds) = $^{1}/_{52}$ = **0.0192 (or 1.9231%)**

b. P(A queen) = $^{4}/_{52}$ = $^{1}/_{13}$ = **0.0769 (or 7.6923 %)**

Tables [3]

A table can be used for the sample space when two dice are rolled.

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

Tables [4]

Example: When two dice are rolled, find the probability of getting a sum of nine.

Solution:

Let A be the event of getting a "sum of 9"

P(A) = ${}^{4}/_{36}$ = ${}^{1}/_{9}$ = **0.1111 (or 11.11 %)**