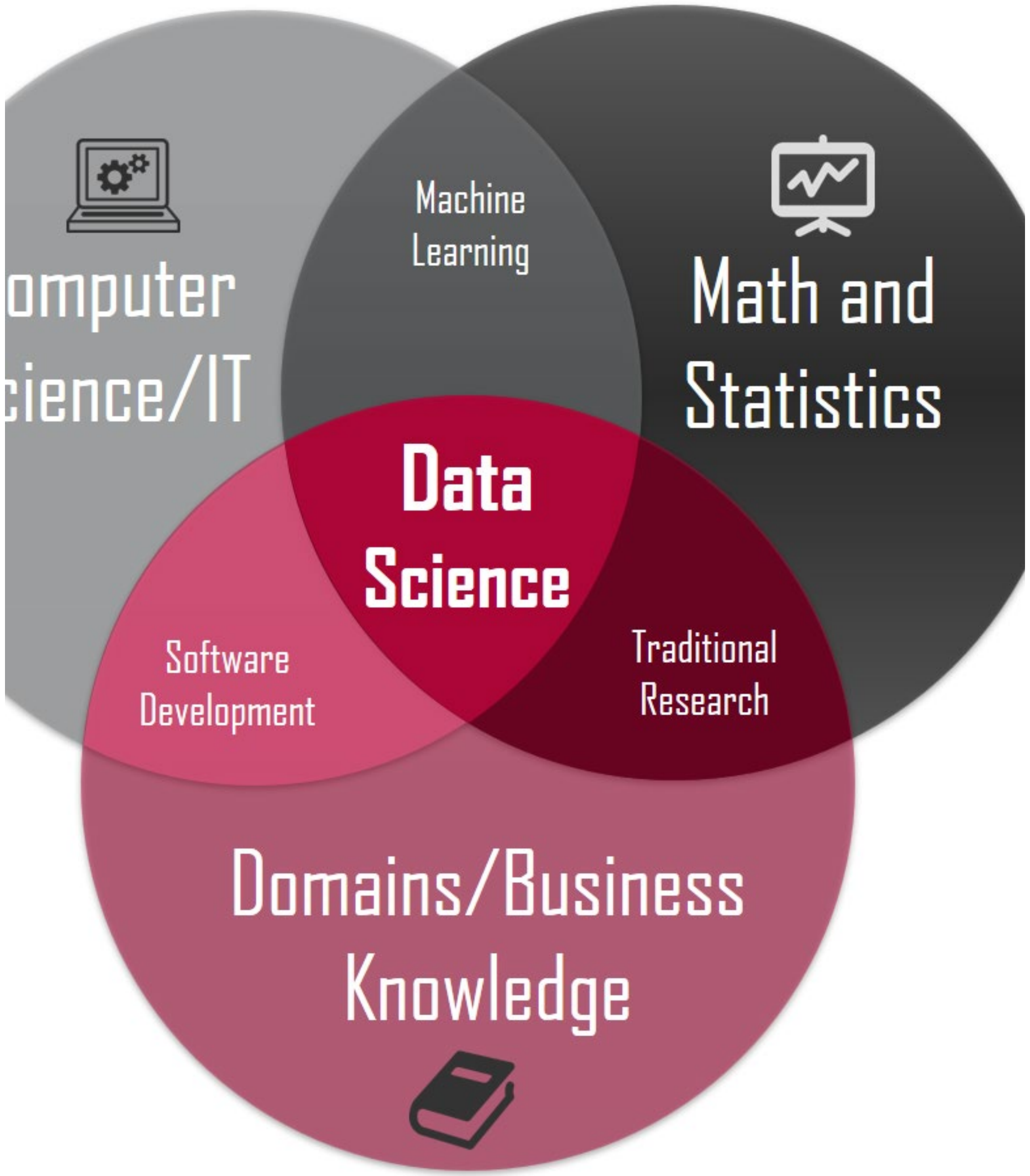


https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

THE UNIVERSITY OF WESTERN AUSTRALIA

STAT 1400

Statistics for Science

Lecture Week 6

Dr Darfiana Nur

Department Mathematics and Statistics

Aims of this week

- Aim 1 Probability – basic concepts
- Aim 2 Manipulating probabilities
- Aim 3 Solving conditional probability
- Aim 4 Probability Models
- Aim 5 Materials from Sem 1 2023 (Frequentism, Random Variables, Expectation, Discrete)

References:

- Moore et al (2021) Chapter 2
- *Rafael A. Irizarry (e-book)* **Introduction to Data Science: *Data Analysis and Prediction Algorithms with R*,** Chapters 3 and 13 http://rafalab.dfci.harvard.edu/dsbook/

Aim 1. Probability

(Chapter 4 – Moore et al 2017)

Why do we study probability?

Can be applied to many phenomena:

- In **Genetics** as a model for mutations and ensuing natural variability, at both quantitative (larger-picture) and molecular levels
- Patients undergoing **Radiation Therapy** (and their physicians alike) are concerned with the probability of cure (long-term recurrence-free survival, meaning the absence of a detectable or symptomatic tumor).
- Designing and analysing computer operating systems: the lengths of queues, optimal procedures, how robots "learn"
- Noise in electrical devices and communication systems, including traffic analysis
- Operations research: demand on inventories of goods, optimal systems for processes
- Actuarial science: used by insurance companies
- Study of complex systems: eg modern commercial or military aircraft, environmental modelling, epidemics

Biodiversity: Normal probability

- A researcher is interested in the lengths of *Salvelinus fontinalis* (brook trout), which are known to be approximately Normally distributed with mean 80 centimeters and standard deviation 5 centimeters.
- To help preserve brook trout populations, some regulatory standards need to be set limiting the size of fish that can be caught. The standard is set so that all fish shorter than 70.20 centimeters or longer than 89.80 centimeters must be thrown back.
- What proportion of the population is this?

Probabilities in court: Forensic Sciences

- Let S be event that suspect was present at the scene of a crime and S' be event that the suspect was not present
- Assume each juror has initial probability for each event
- Witness says he/she saw a tall Caucasian male running from scene. The defendant is tall Caucasian male

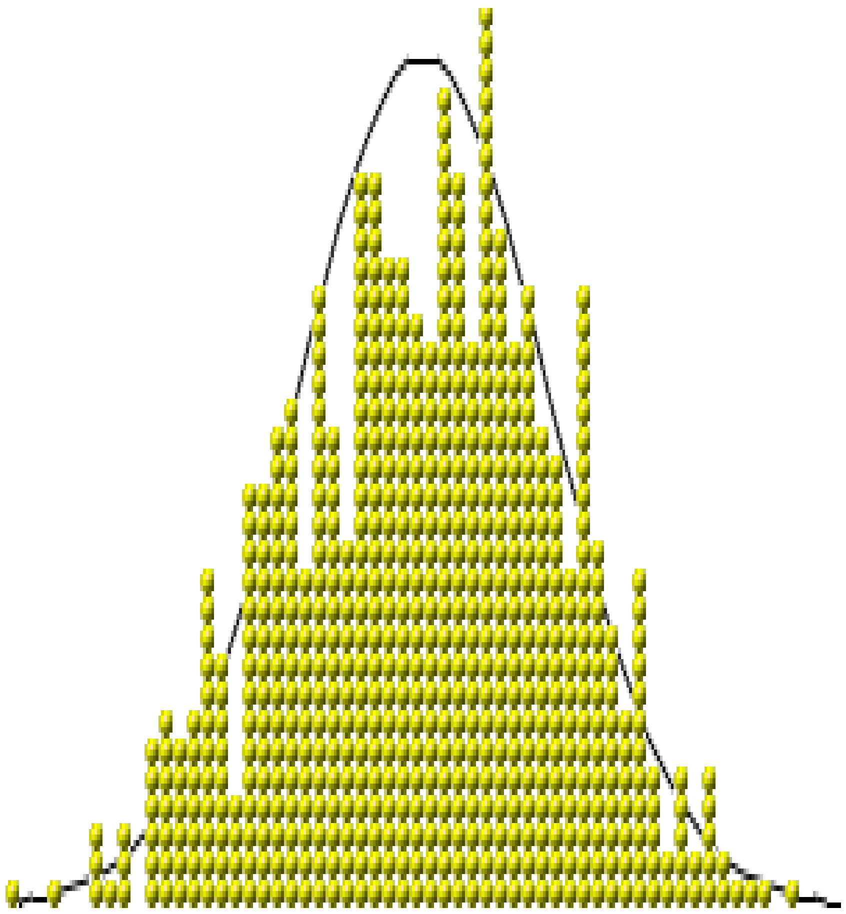
  – jurors update probabilities
- Window broken during the crime and fragments found on the defendant's clothing match the broken window

  – jurors update probabilities
- How should jurors update their probabilities? Do jurors actually think this way?

Randomness and probability

- The probability of any outcome of a random phenomenon can be defined as the proportion of times the outcome would occur in a very long series of repetitions.

  - A phenomenon is random if individual outcomes are uncertain, but there is nonetheless a regular distribution of outcomes in a large number of repetitions.



What is probability?

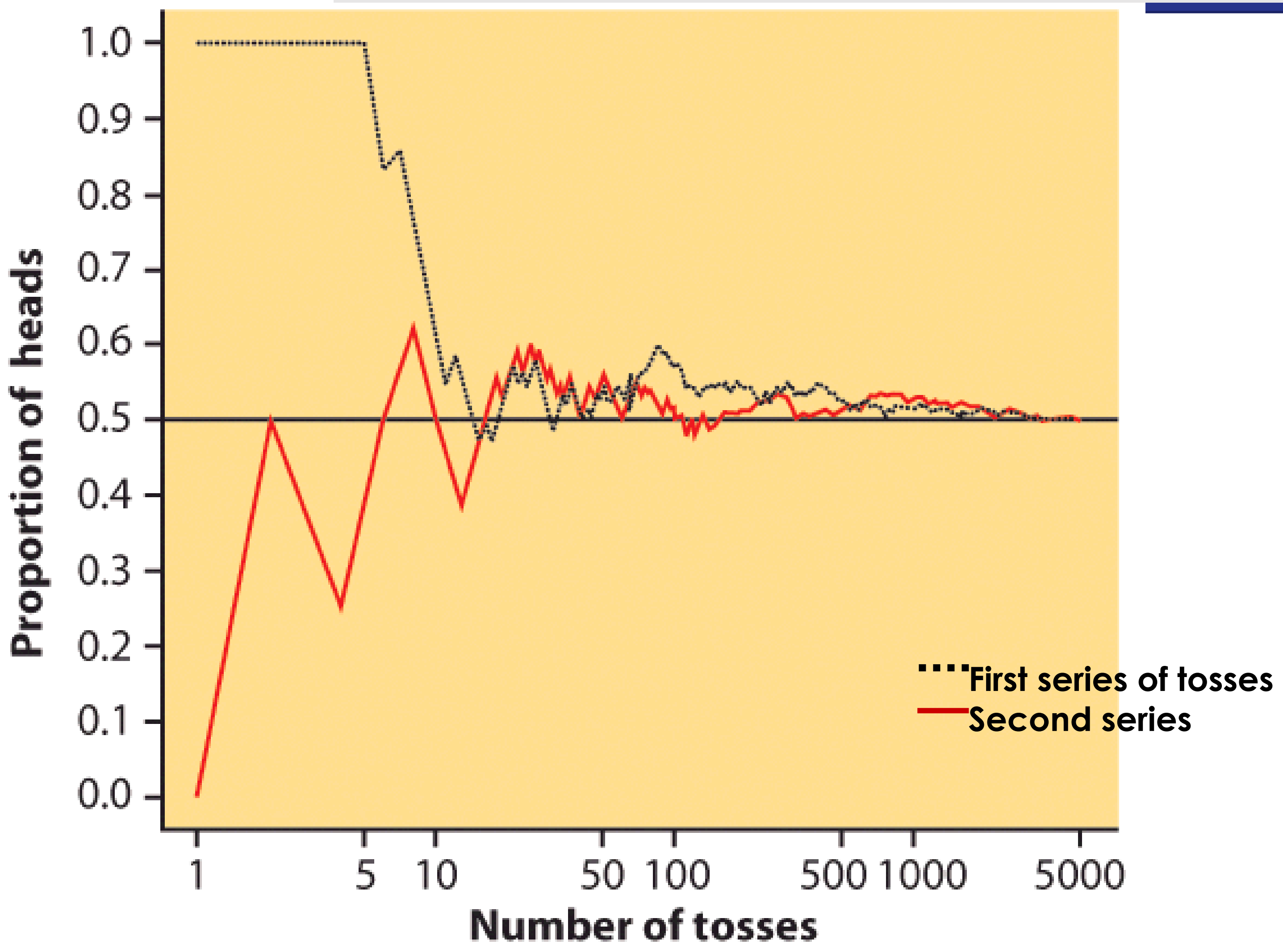
- The *probability* of any outcome of a random phenomenon is **the proportion of times the outcome would occur in a very long series of repetitions.**

- We use probability subjectively every day: e.g. How likely is it to rain? Should I take my umbrella?

- **Empirical Probability:** Based on data from an experiment. Toss a coin 100 times and see how likely the outcome head is to occur.

Coin toss

The result of any single coin toss is random. But the result over many tosses is predictable, as long as the trials are **independent** (i.e., the outcome of a new coin flip is not influenced by the result of the previous flip).



The probability of heads is 0.5 = the proportion of times you get heads in many repeated trials.

Probability models

Probability models describe, mathematically, the outcome of random processes. They consist of two parts:

1) ***S* = Sample Space:** This is **a set, or list,** of **<u>all</u>** possible outcomes of a random process. An **event** is a subset of the sample space (usually denoted as A, B, C...)
2) A **probability** for each possible event in the sample space *S*.

Example 1: *Probability Model for a Coin Toss*:

S = {Head, Tail}

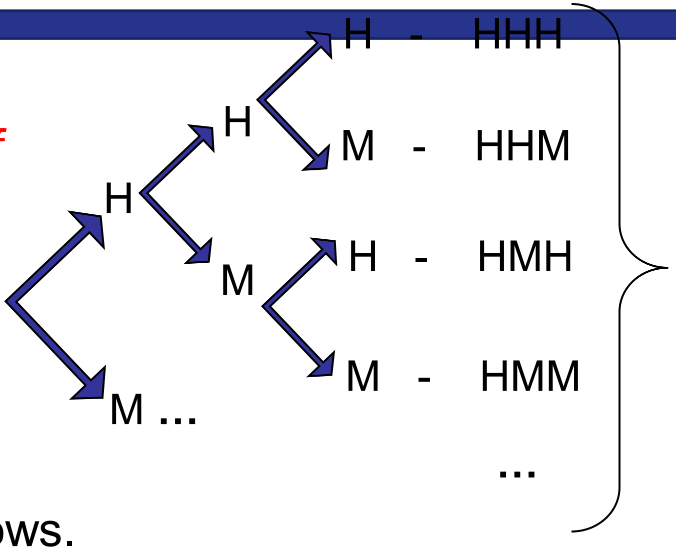
Probability of heads = 0.5

Probability of tails = 0.5

Sample spaces

It's the question that determines the sample space.

A. A basketball player shoots three free throws. What are the possible sequences of hits (H) and misses (M)?



S = { HHH, HHM, HMH, HMM, MHH, MHM, MMH, MMM }

Note: 8 elements, 2^3

B. A basketball player shoots three free throws. What is the number of baskets made?

$S = \{ 0, 1, 2, 3 \}$

C. A nutrition researcher feeds a new diet to a young male white rat. What are the possible outcomes of weight gain (in grams)?

$S = [0, \infty[= (\text{all numbers} \geq 0)$

Probability rules

Coin Toss Example:
S = {Head, Tail}
Probability of heads = 0.5
Probability of tails = 0.5

1) Probabilities range from
0 (*no chance of the event*) to
1 (*the event has to happen*).

For any event A, $0 \le P(A) \le 1$

Probability of getting a Head = 0.5
We write this as: $P(\text{Head}) = 0.5$

$P(\text{neither Head nor Tail}) = 0$
$P(\text{getting either a Head or a Tail}) = 1$

2) Because some outcome must occur on every trial, the sum of the probabilities for all possible outcomes (the sample space) must be exactly 1.

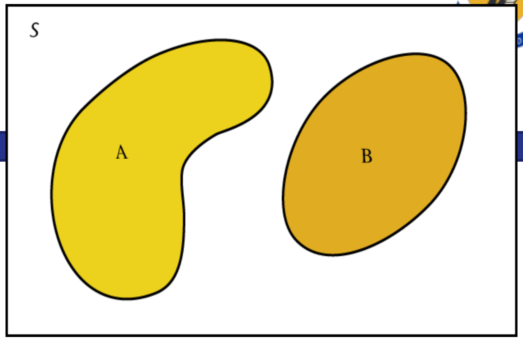
$P(\text{sample space}) = P(S) = 1$

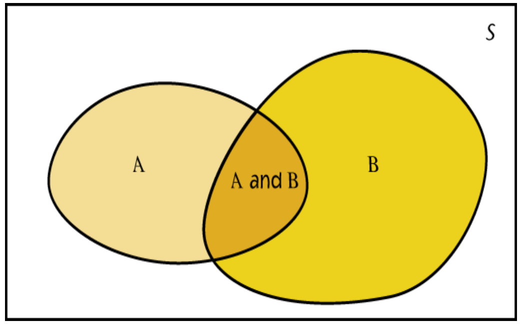
Coin toss: $S = \{\text{Head, Tail}\}$

$P(\text{head}) + P(\text{tail}) = 0.5 + 0.5 = 1$
→ $P(\text{sample space}) = 1$

Probability rules (*cont'd*)

Venn diagrams:
A and B disjoint





A and B not disjoint

3) Two events A and B are **disjoint** if they have no outcomes in common and can never happen together**.** The probability that A or B occurs is then the sum of their individual probabilities.

P(A or B) = "*P*(A U B)" = *P*(A) + *P*(B)
This is the **addition rule for disjoint events.**

Example 2: If you flip two coins, and the first flip does not affect the second flip:

S = {HH, HT, TH, TT}. The probability of each of these events is 1/4, or 0.25.

The probability that you obtain "only heads or only tails" is:

P(HH or TT) = *P*(HH) + *P*(TT) = 0.25 + 0.25 = 0.50

Probability rules (*cont'd*)

Coin Toss Example:
S = {Head, Tail}
Probability of heads = 0.5
Probability of tails = 0.5

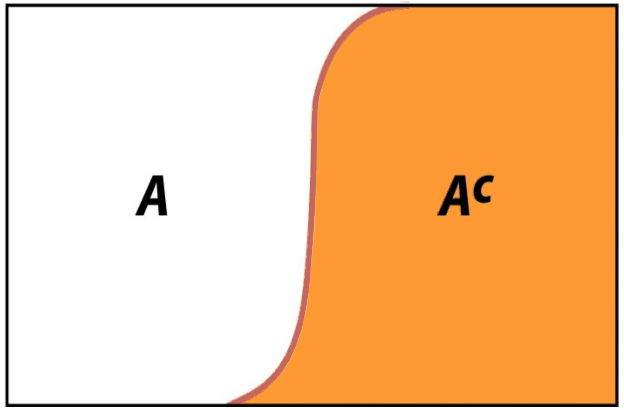
4) The complement of any event A is the event that A does not occur, written as A^c.

The complement rule states that the probability of an event not occurring is 1 minus the probability that is does occur.

$P(\text{not A}) = P(A^c) = 1 - P(A)$

Tail^c = not Tail = Head

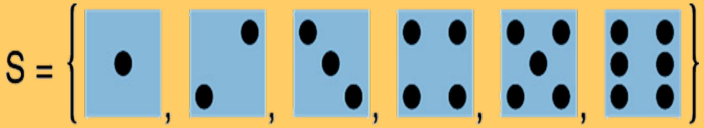
$P(\text{Tail}) = 1 - P(\text{Head}) = 0.5$



Venn diagram:
Sample space made up of an event A and its complementary A^c, i.e., everything that is not A.

Probabilities: finite number of outcomes

- **Finite sample spaces** deal with **discrete data**—data that can only take on a limited number of values. These values are often **integers** or **whole numbers.**

Throwing a die: 

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

- The individual outcomes of a random phenomenon are always disjoint.

  ➔ The probability of any event is the sum of the probabilities of the outcomes making up the event (addition rule).

Example 3: M&M candies

If you draw an M&M candy at random from a bag, the candy will have one of six colors. The probability of drawing each color depends on the proportions manufactured, as described here:

Color	Brown	Red	Yellow	Green	Orange	Blue
Probability	0.3	0.2	0.2	0.1	0.1	?

What is the probability that an M&M chosen at random is blue?

S = {brown, red, yellow, green, orange, blue}

$P(S) = P(\text{brown}) + P(\text{red}) + P(\text{yellow}) + P(\text{green}) + P(\text{orange}) + P(\text{blue}) = 1$

$P(\text{blue}) = 1 - [P(\text{brown}) + P(\text{red}) + P(\text{yellow}) + P(\text{green}) + P(\text{orange})]$

$= 1 - [0.3 + 0.2 + 0.2 + 0.1 + 0.1] = 0.1$

In-Class Exercise 1

What is the probability that a random M&M is either red, yellow, or orange?

$P(\text{red or yellow or orange}) = P(\text{red}) + P(\text{yellow}) + P(\text{orange})$

$= 0.2 + 0.2 + 0.1 = 0.5$

Probabilities: equally likely outcomes

We can assign probabilities either:

empirically → from our knowledge of numerous similar past events

Mendel discovered the probabilities of inheritance of a given trait from experiments on peas without knowing about genes or DNA.

or theoretically → from our understanding of the phenomenon and symmetries in the problem

A 6-sided fair die: each side has the same chance of turning up

Genetic laws of inheritance based on meiosis process

If a random phenomenon has *k* equally likely possible outcomes, then each individual outcome has probability 1/*k*.

And, for any event A:

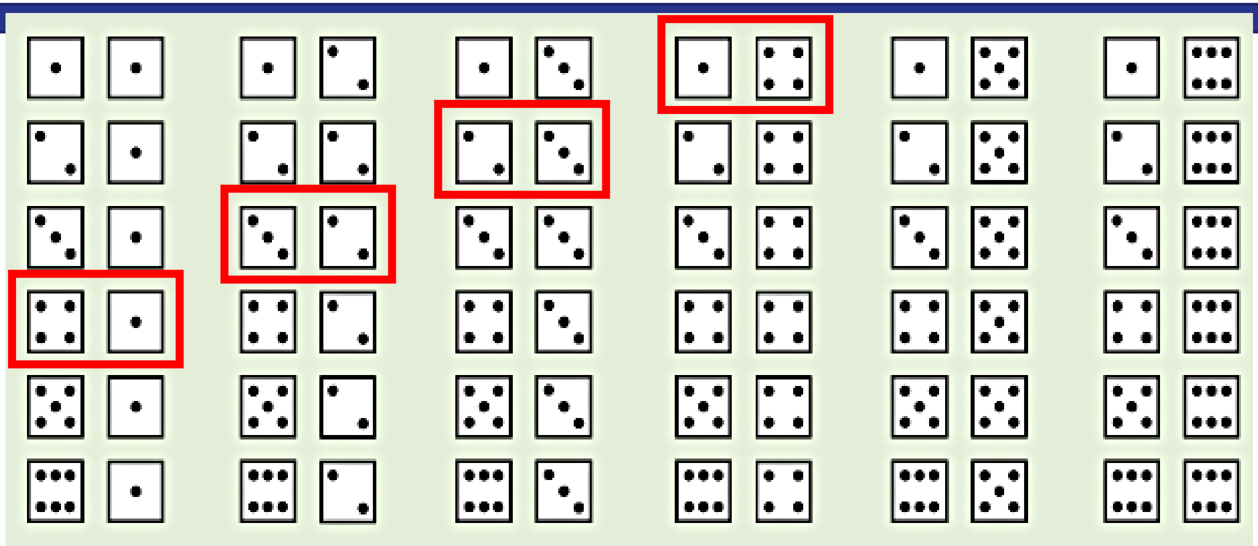
$$P(\mathrm{A}) = \frac{\text{count of outcomes in A}}{\text{count of outcomes in S}}$$

Example 4: Dice

You toss two dice. What is the probability of the outcomes summing to 5?

This is *S*:

{(1,1), (1,2), (1,3), ……etc.}



There are 36 possible outcomes in *S*, all equally likely (given fair dice).

Thus, the probability of any one of them is 1/36.

P(the roll of two dice sums to 5) =

$P(1,4) + P(2,3) + P(3,2) + P(4,1) = 4 / 36 = 0.111$

THE UNIVERSITY OF WESTERN AUSTRALIA

Example 5: The gambling industry relies on probability distributions to calculate the odds of winning. The rewards are then fixed precisely so that, **on average**, players lose and the house wins.

The industry is very tough on so called “cheaters” because their probability to win exceeds that of the house. Remember that it is a business, and therefore it has to be profitable.

Probability rules (*cont'd*)

Coin Toss Example:
S = {Head, Tail}
Probability of heads = 0.5
Probability of tails = 0.5

5) Two events A and B are independent if knowing that one occurs does not change the probability that the other occurs.

If A and B are independent, *P*(A and B) = *P*(A)*P*(B)

This is the **multiplication rule for independent events.**

Two consecutive coin tosses:
P(first Tail and second Tail) = *P*(first Tail) * *P*(second Tail) = 0.5 * 0.5 = 0.25

Venn diagram:
Event A and event B. The intersection represents the event {A and B} and outcomes common to both A and B.

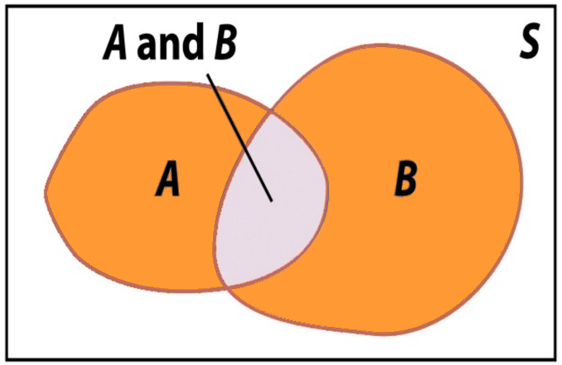


Figure 4-4
Introduction to the Practice of Statistics, Fifth Edition

Aim 2 Manipulating Probabilities

- A, B sets representing events

Intersection: means **both** A **and** B occur (A ∩ B)

Union: means **either** A **or** B occurs (A ∪ B)

- **Addition rule**: If A and B are two events, the probability of A or B is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

or

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

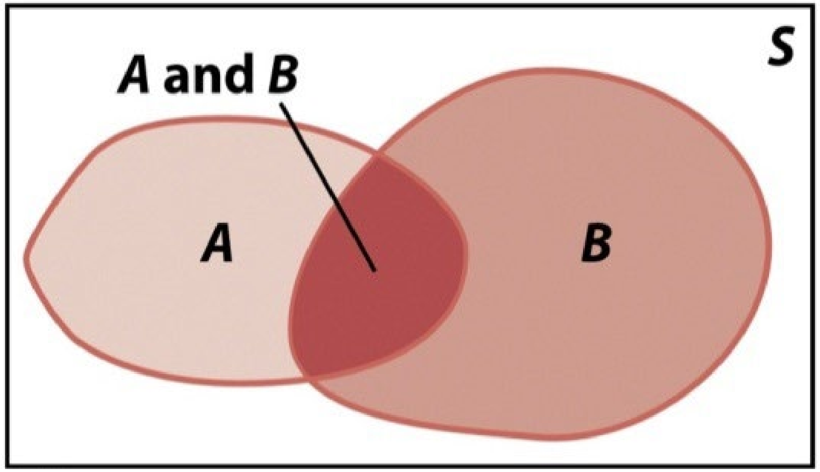
General addition rules

General addition rule for <u>any</u> two events A and B:

The probability that A occurs,

B occurs, or both events occur is:

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$



Example 6: What is the probability of randomly drawing either an ace or a heart from a deck of 52 playing cards? There are 4 aces in the pack and 13 hearts. However, 1 card is both an ace and a heart. Thus:

$P(\text{ace or heart}) = P(\text{ace}) + P(\text{heart}) - P(\text{ace and heart})$

$= 4/52 + 13/52 - 1/52 = 16/52 \approx .3$

Conditional probability

THE UNIVERSITY OF WESTERN AUSTRALIA

- **Conditional probabilities** reflect how the probability of an event can change if we know that some other event has occurred/is occurring.

  Example: The probability that a cloudy day will result in rain is different if you live in Los Angeles than if you live in Seattle.

- Our brains effortlessly calculate conditional probabilities, updating our "degree of belief" with each new piece of evidence.

- The conditional probability of event B occurs given that event A has occurred (or is occurring) is: (provided that $P(A) \neq 0$)

$$\boxed{P(B \mid A) = \frac{P(A \, and \, B)}{P(A)}}$$

General multiplication rules

The probability that any two events, A and B, both occur is:

$P(\text{A and B}) = P(\text{A})P(\text{B} \mid \text{A})$

This is the **general multiplication rule.**

Example 7: What is the probability of randomly drawing two hearts from a deck of 52 playing cards?

There are 13 hearts in the pack. Let A and B be the events that the first and second cards drawn are hearts, respectively. Assume that the first card is not replaced before the second card is drawn.

$P(\text{A}) = 13/52 = 1/4$ $\qquad\qquad$ $P(\text{B} \mid \text{A}) = 12/51$

➔ $P(\text{two hearts}) = P(\text{A})^* P(\text{B} \mid \text{A}) = (1/4)^*(12/51) = 3/51$

Notice that the probability of a heart on the second draw depends on which card was removed on the first draw.

Independence

- Two events are independent if the occurrence of one doesn't effect the probability of the other:

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

This means the multiplication rule

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

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

if independent

Independent Events

Example 8: What is the probability of randomly drawing two hearts from a deck of 52 playing cards if the first card is replaced (and the cards re-shuffled) before the second card is drawn.

P(A) = 1/4 P(B | A) = 13/52 = 1/4

➔P(A and B) = P(A)* P(B) = (1/4)*(1/4) = 1/16

Notice that the two draws are independent events if the first card is replaced before the second card is drawn.

In summary, to show that two events A and B are independent, use either one of these:

If A and B are independent, then P(A and B) = P(A)P(B)

(A and B are independent when they have no influence on each other's occurrence.)

Two events A and B that both have positive probability are **independent** if P(B | A) = P(B)

Mutually exclusive (disjoint) events

If A and B are mutually exclusive: (i.e. they cannot both occur or are disjoint)

$$P(A \text{ and } B) = 0$$

Aim 3 Solving Conditional Probability

- Conditional probabilities reflect how the probability of an event can change if we know that some other event has occurred/is occurring
- The notation: P(A|B) is a conditional probability.
- You can read the | as "given the information that."
- Conditional probability works by restricting the range of events considered
- Our brains try to calculate conditional probabilities all the time, updating our "degree of belief" with each new piece of evidence.

Conditional probability rules

THE UNIVERSITY OF WESTERN AUSTRALIA

This notation is used in these two rules:

Conditional rule:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

Multiplication rule:

$$P(A \text{ and } B) = P(B \mid A) \times P(A)$$

or

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

Contingency table or crosstab

	B	**B^C**	
A	P(A and B)	P(A and B^c)	**P(A)**
A^c	P(A^c and B)	P(A^c and B^c)	**P(A^c)**
	P(B)	**P(B^c)**	**1**

The row and column totals give the marginal probability distribution for events A and B.

The **numbers within the table** give the joint probability distribution of the 2 events.

Example 9

A type of tree has variegated leaves 35% of the time while the rest have normal green leaves. 70% of the trees have white flowers while the rest have pink flowers. 20% of the trees have variegated leaves and white flowers.

What is the probability of a randomly selected tree having variegated leaves OR white flowers?

Solution:

Define event A – tree having variegated leaves.

Define event B – tree having white flowers.

$$P(A)=0.35;\ P(B)=0.70;\ P(A \text{ and } B)=0.20$$

What is the probability of the event A or B?

Example 9 (ctnd): Contingency table

P(A)=0.35; P(B)=0.70; P(A and B)=0.20

	White (B)	Pink (not B)	**Total**
Variegated (A)	**.2**	.15	**.35**
Normal (not A)	.5	.15	.65
Total	**.7**	.3	1

Solution:

1. Draw a Contingency table as in the previous slide

2. Use the rule: $$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Therefore the probability of a randomly selected tree having variegated leaves or white flowers is 85%.

$$P(A \text{ or } B) = 0.35 + 0.7 - 0.2 \quad = 0.85$$

Example 9 (ctnd)

Solving conditional probability problems using contingency table

From previous example:

What is the probability of white flowers in a tree given that the tree has variegated leaves?

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

event A: tree having variegated leaves

event B: tree having white flowers

$$P(\text{White fl.}|\text{Variegated leaves}) = \frac{P(\text{White fl. and Varigated leaves})}{P(\text{Variegated leaves})}$$

$$= (0.2) / (0.35) = 0.571$$

Hence the probability of white flowers in a tree given that the tree has variegated leaves is 57.1%

In-Class Exercise 2

What is the probability of pink flowers in a tree given that the tree has variegated leaves?

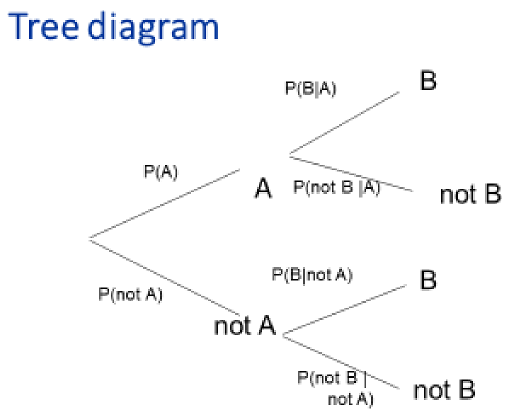
What is the probability of pink flowers in a tree given that the tree has normal leaves?

Tree Diagrams

Probability problems often require us to combine several of the basic rules into a more elaborate calculation. One way to model chance behavior that involves a sequence of outcomes is to construct a **tree diagram.**

Consider flipping a coin twice.

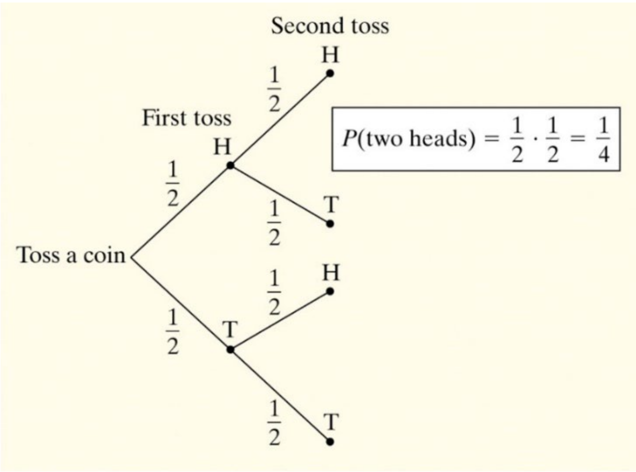
What is the probability of getting two heads?



Sample Space:

HH   HT   TH   TT

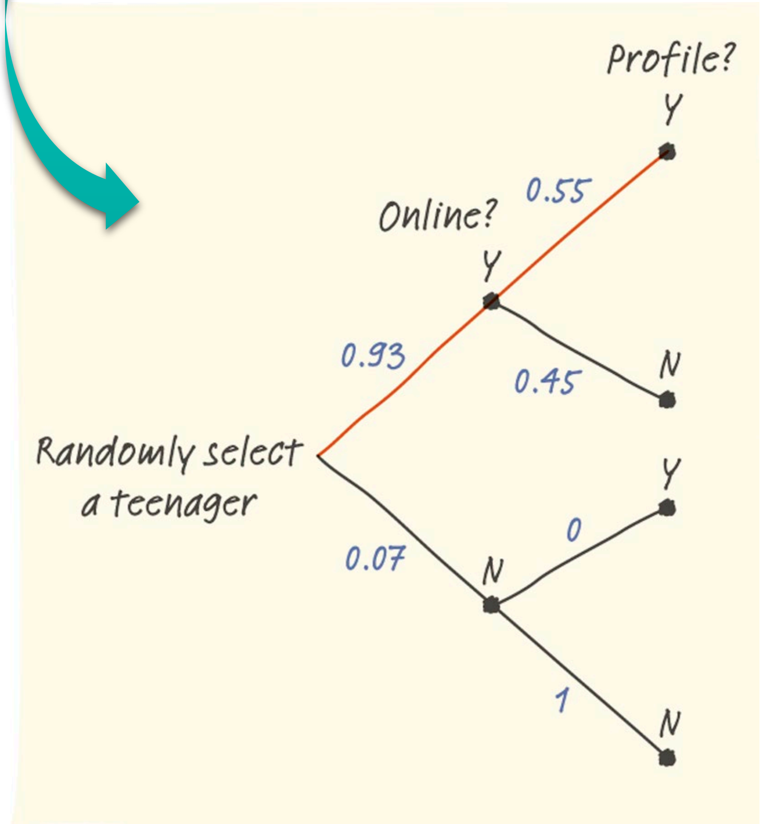
So, *P*(two heads) = *P*(HH) = 1/4



Example 10

The Pew Internet and American Life Project finds that 93% of teenagers (ages 12 to 17) use the Internet, and that 55% of online teens have posted a profile on a social-networking site.

What percent of teens are online *and* have posted a profile?



$$P(\text{online}) = 0.93$$
$$P(\text{profile} \mid \text{online}) = 0.55$$

$$P(\text{online} \quad \text{and} \quad \text{have profile})$$
$$= P(\text{online}) \cdot P(\text{profile}|\text{online})$$
$$= (0.93)(0.55)$$
$$= 0.5115$$

51.15% of teens are online *and* have posted a profile.

Example 11: Students smoking habit and parents smoking habit (in count)

Student smoking habit

Parent smoking habit

	Smoke	Doesn't smoke	Total
Both smoke or one smokes	816	3203	4019
Neither smoke	188	1168	1356
Total	1004	4371	5375

Students smoking habit and parents smoking habit: Contingency table (in %)

Student smoking habit

Parent smoking habit

	Smoke	Doesn't smoke	Total
Both smoke or one smokes	816/5375= 15.18%	3203/5375= 59.59%	4019/5375 =74.77%
Neither smoke	188/5375= 3.50%	1168/5375= 21.73%	1356/5375 =25.23%
Total	1004/5375= 18.68%	4371/5375= 81.32%	5375/5375 =100.00%

In-Class Exercise 3

1. What is the probability of a student smokes given that neither of the parent smoke?
2. What is the probability of a student does not smoke given that both parents smoke or one parent smokes?

Aim 4. Slides from Sem1 2023

- **Frequentism**: Probability as a frequency
- **Random variable:**
  - ✓ **discrete** random variable
- **Expectation**