5005 CEM

Data Science

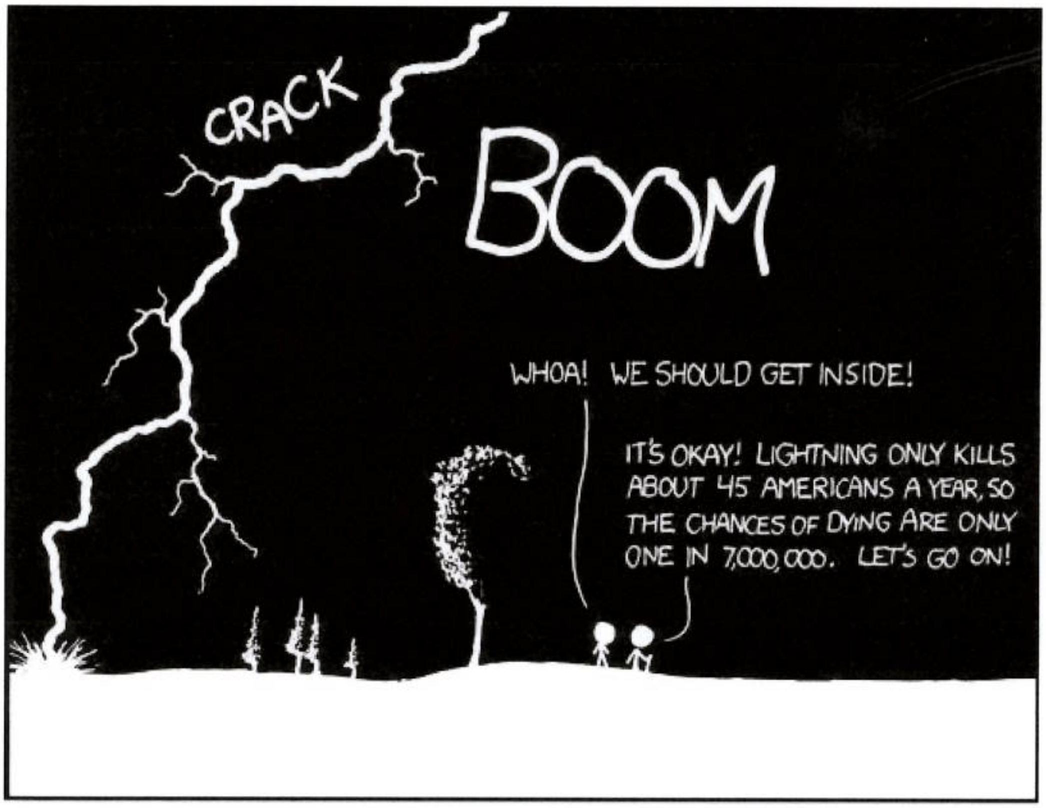
JLG20

22/1/2020

and 24/1/2020

Topic 1

Probability for Data Scientists

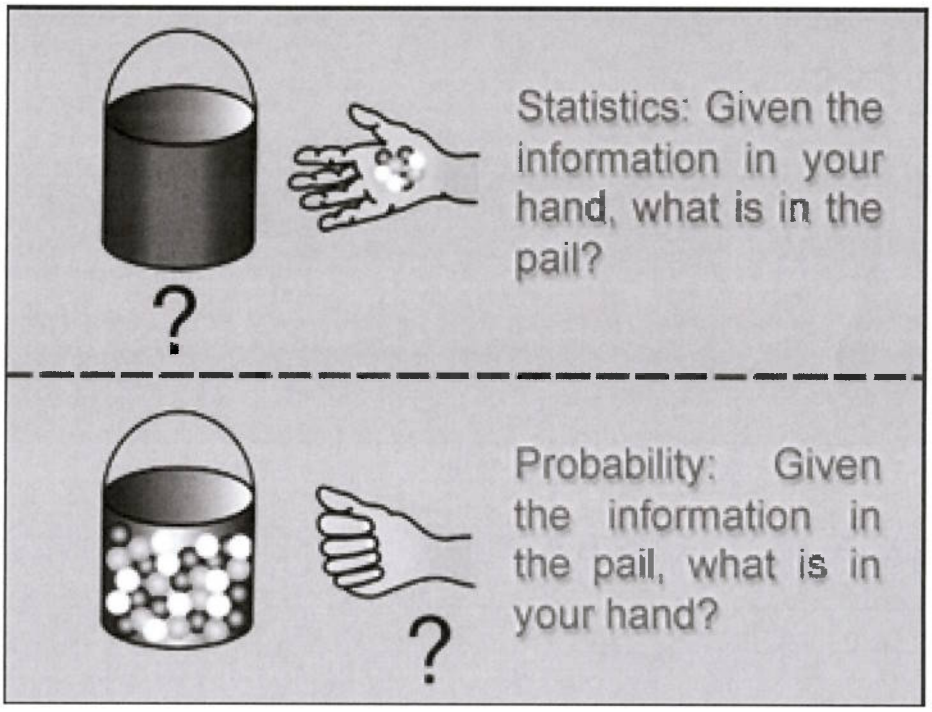

CRACK
BOOM
WHOA! WE SHOULD GET INSIDE!
IT'S OKAY! LIGHTNING ONLY KILLS ABOUT 45 AMERICANS A YEAR, SO THE CHANCES OF DYING ARE ONLY ONE IN 7,000,000. LET'S GO ON!

THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Image from http://xkcd.com/795/

1.1 What is Probability?

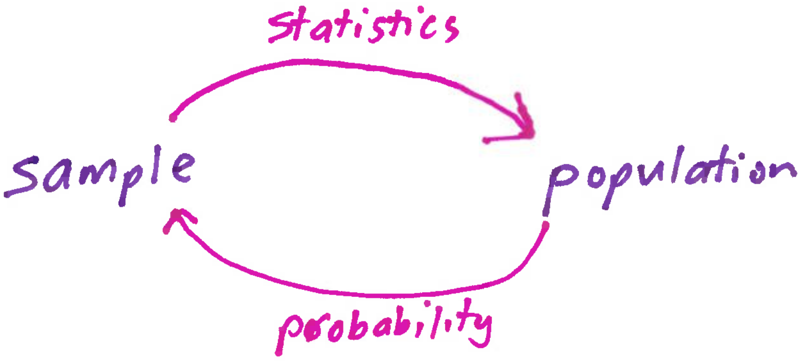
> "A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician." — Josh Blumenstock



Statistics: Given the information in your hand, what is in the pail?

Probability: Given the information in the pail, what is in your hand?

Image from http://ocw.mit.edu/courses/sloan-school-of-management/15-082j-network-optimization-spring-2003/chp_probability_stat.jpg



1.1.1 Sample Space and Events

definitions

- An **experiment** (or **trial**) is any operation or procedure whose *outcome* cannot be predicted with certainty, e.g., flip a coin
- The **sample space**, S, consists of *all possible outcomes* associated with the experiment.
- An **event**, E, is some *subset* of the sample space, i.e., $E \subseteq S$.

↑ subset

Example 1.1 Consider the experiment of flipping a coin. The sample space is a head or a tail, i.e., $S = \{H, T\}$. □

Example 1.2 Consider the experiment of tossing a coin twice in a row. The sample space is a head or a tail for each toss, i.e.,

$$S = \{HH, HT, TH, TT\}$$

Obtaining at least one head is the event

$$E = \{HH, HT, TH\}$$

□

Practice Problem. Consider the experiment of tossing a single 6-sided die. The sample space is

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

Obtaining an even number is the event

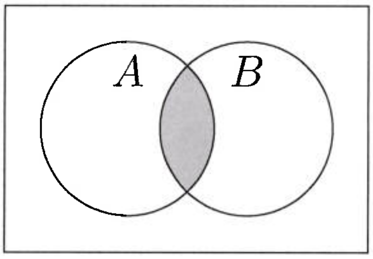
$$E = \{2, 4, 6\}$$

□

Suppose A and B are events, i.e., $A \subseteq S$ and $B \subseteq S$. Because A and B are subsets of S, we will regularly use the following (hopefully familiar) definitions from set theory.

- The **intersection** of A and B is the set of outcomes in A **and** B, i.e., $A \cap B$.

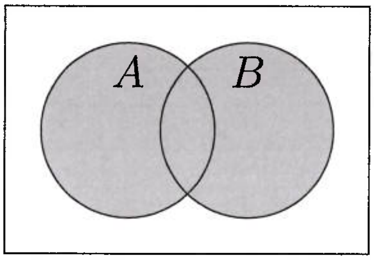
$A \cap B$



Venn diagram

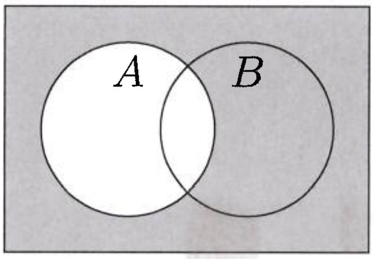
- The **union** of A and B is the set of outcomes in A **or** B or both, i.e., $A \cup B$.

$A \cup B$



- The **complement** of A is the set of outcomes in S that are **not** in A, i.e., $\overline{A}$.

$\overline{A}$



symbol

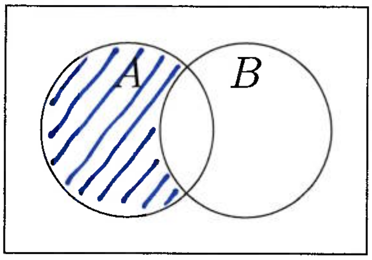
↓

$\emptyset$

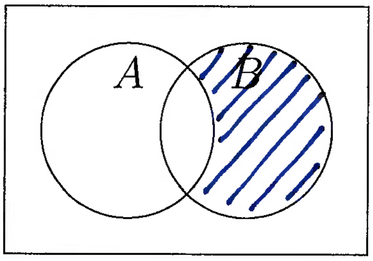
- The set with no elements is called the **empty set**, written $\emptyset$. The empty set $\emptyset$ is considered to be a subset of every set.

Practice Problem. Shade the following sets on the Venn diagram.

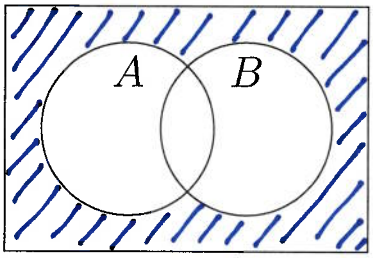
$A \cap \overline{B}$



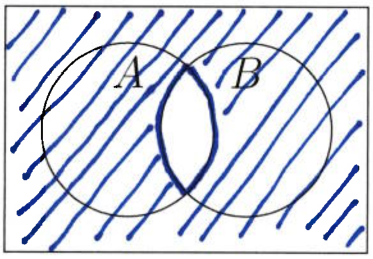
$\overline{A} \cap B$



$\overline{A \cup B}$



$\overline{A \cap B}$



1.1.2 Definition of Probability

probability = 0

- When an event is **impossible** we say the probability of it happening is 0.

  Example — it is impossible to live without oxygen and so the probability of doing this is 0.

probability = 1

- When an event is **certain** we say the probability of it happening is 1.

  Example — it is certain that a metal bar will sink when placed in water and so the probability of this happening is 1.

- Most events are **neither impossible nor certain**.

  The probability of such events lies between 0 and 1.

  Events which are *likely* to happen have probabilities close to 1.

  Events which are *unlikely* to happen have probabilities close to 0.

  An event which is as likely to happen as not has a probability of $\frac{1}{2}$, e.g., throwing a head with a fair coin.

- Events A and B are **mutually exclusive** if

$$A \cap B = \emptyset$$

i.e., they have *no common outcomes*.

Important definition

empty set

Fundamental rules or laws

■ *Axioms of Probability*

Let $P(A)$ denote the probability of the event A in a finite sample space S.

Axiom 1. $0 \leq P(A) \leq 1$ for each event A in S

Axiom 2. $P(S) = 1$

Axiom 3. If A and B are mutually exclusive events in S, then

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

■ What is $P(A \cup B)$ when A, B not mutually exclusive? (see p 1014)

Question —

How do we know the probability of an event?

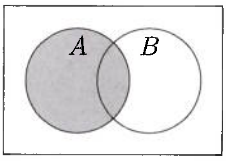
Equally Likely Outcomes

- Suppose an experiment has n equally likely outcomes, and an event E occurs if any one of k of these outcomes occurs as an outcome of the experiment.
- We define the probability of event E as

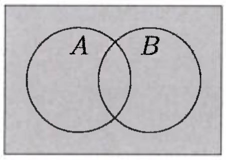
$$P(E) = \frac{k}{n}$$

If all outcomes are equally likely, the **probability** of event A is

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



divided by



Example 1.3 Suppose a box of 100 items contains 5 defective items. If one item is **randomly selected** then

$$P(\text{item is defective}) = \frac{5}{100} = 0.05$$

□

Practice Problem. A fair 6-sided die is ~~tossed~~ rolled. The event A is defined as "the number obtained is a multiple of 3".

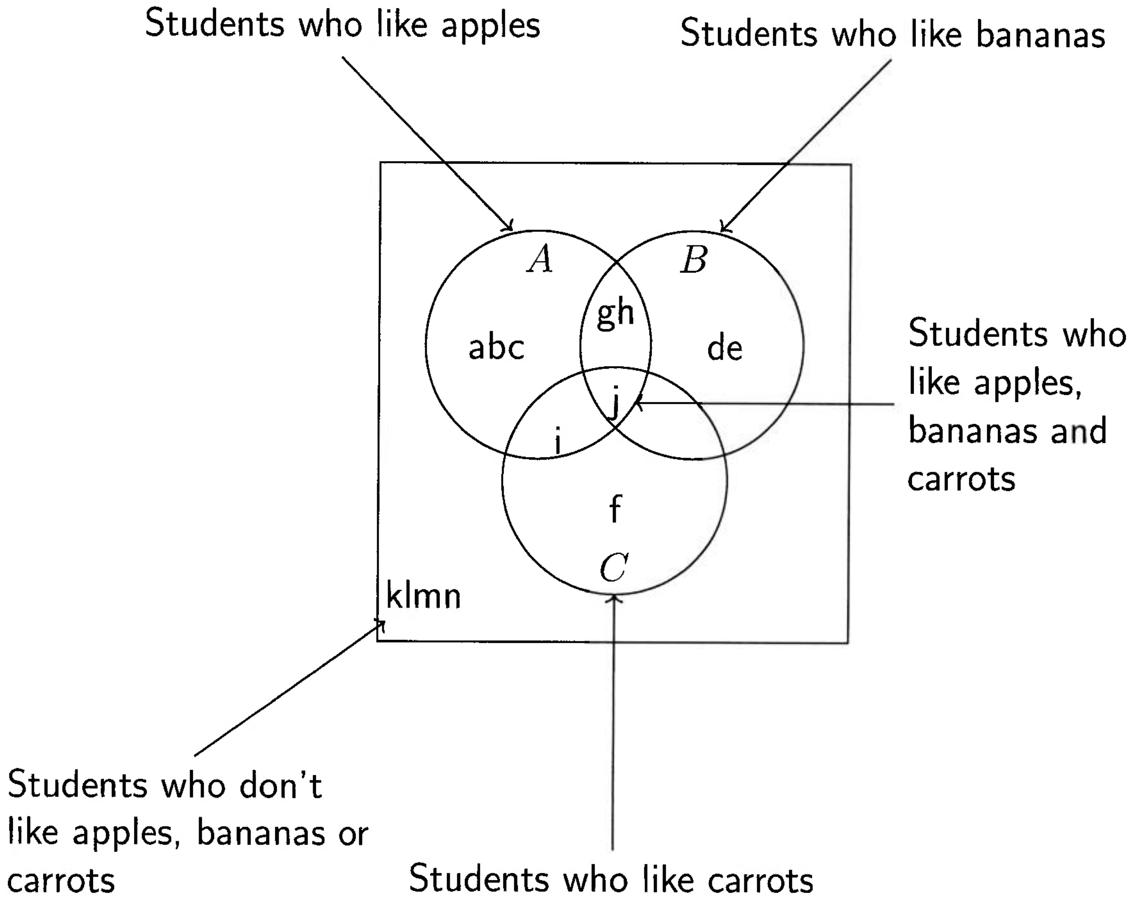
$P(A) = \frac{2}{6}$ ← no need to simplify fractions

$P(\text{prime on 20-sided dice})$
$= \frac{8}{20}$

primes: 2, 3, 5, 7
11, 13, 17, 19, ...

□

Example 1.4 Suppose there are a group of 14 students. For simplicity, assume their names are a, b, c, . . . , n. Some like to eat apples, some like to eat bananas and some like to eat carrots, as shown below.


Students who like apples
Students who like bananas
A
B
abc
gh
de
j
i
f
C
klmn
Students who like apples, bananas and carrots
Students who don't like apples, bananas or carrots
Students who like carrots

If one student is **randomly selected**, what is the probability that the student

- likes apples

$$P(A) = \frac{|A|}{|S|} = \frac{7}{14} = 0.5$$

- likes bananas

$$P(B) = \frac{|B|}{|S|} = \frac{5}{14}$$

- likes carrots

$$P(C) = \frac{|C|}{|S|} = \frac{3}{14}$$

- likes apples and bananas

$$P(A \cap B) = \frac{|A \cap B|}{|S|} = \frac{3}{14}$$

- likes bananas and carrots

$$P(B \cap C) = \frac{|B \cap C|}{|S|} = \frac{1}{14}$$

- likes apples, bananas and carrots

$$P(A \cap B \cap C) = \frac{|A \cap B \cap C|}{|S|} = \frac{1}{14}$$

- does not like apples

$$P(\overline{A}) = \frac{|\overline{A}|}{|S|} = \frac{7}{14} = 0.5$$

- likes bananas or likes carrots

$$P(B \cup C) = \frac{|B \cup C|}{|S|} = \frac{7}{14}$$

- likes bananas but does not like apples

$$P(B \cap \overline{C}) = \frac{|B \cap \overline{C}|}{|S|} = \frac{2}{14}$$

- likes bananas and carrots but does not like apples

$$P(B \cap C \cap \overline{A}) = \frac{|B \cap C \cap \overline{A}|}{|S|} = \frac{0}{14} = 0$$

□

Practice Problem. A fair coin is tossed twice. The sample space is $S = \{HH, HT, TH, TT\}$. Find the probability that the outcome has

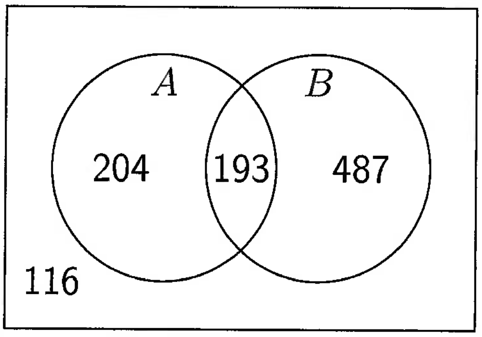
- no heads : $E = \{TT\}$ $P = \frac{1}{4}$
- exactly one head : $E = \{HT, TH\}$ $P = \frac{2}{4}$
- two heads : $E = \{HH\}$ $P = \frac{1}{4}$

□

Relative Frequency → Counting

- If an experiment is repeated many times, the probability of an event is the proportion of times the event occurs in n repetitions of the experiment.

Example 1.5 Suppose an experiment is repeated 1000 times and we count how many times events A only occurs, B only occurs, A and B both occur, and neither A nor B occur.



$$\begin{aligned}
P(A) &= \tfrac{204+193}{1000} = 0.397 \\
P(B) &= \tfrac{487+193}{1000} = 0.68 \\
P(A \cap B) &= \tfrac{193}{1000} = 0.193 \\
P(A \cup B) &= \tfrac{204+193+487}{1000} = 0.884 \\
P(\overline{A \cup B}) &= 1 - P(A \cup B) = 1 - 0.884 = 0.116
\end{aligned}$$

□

1.1.3 Some Rules of Probability

n events

Axiom 3 (see page 1007) can be extended from two mutually exclusive events to n mutually exclusive events.

- Events $A_1, A_2, \ldots, A_n$ are **mutually exclusive** if the occurrence of any one of them implies that none of the others can occur, i.e.,

$$A_i \cap A_j \;=\; \emptyset \quad \text{for every } i \text{ and } j \text{ where } i \neq j$$

- Events $A_1, A_2, \ldots, A_n$ are **exhaustive** if it is certain that at least one of them occurs, i.e.,

$$A_1 \cup A_2 \cup \cdots \cup A_n = S$$

*

■ *Addition Law of Probability*

If $A_1, A_2, \ldots, A_n$ are n mutually exclusive events within a sample space, then

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

Add probabilities of mutually exclusive events.

Example 1.6

An Irish rugby club has 40 players, of whom 7 are called O'Brien, 6 are called O'Connell, 4 are called O'Hara, 8 are called O'Neill and there are 15 others.

If the captain of the team is chosen at random, determine the probability that the captain is

(i) called either O'Brien or O'Connell

(ii) is not called either O'Hara or O'Neill

Solution.

The sample space consists of the 40 players, each of whom is equally likely to be selected as captain.

- Let B be the event "the captain is an O'Brien".
- Let C be the event "the captain is an O'Connell".
- Let H be the event "the captain is an O'Hara".
- Let N be the event "the captain is an O'Neill".

Events B, C, H and N are mutually exclusive, since a player cannot have two surnames.

(i) $P(B \cup C) = \frac{7}{40} + \frac{6}{40} = \frac{13}{40}$

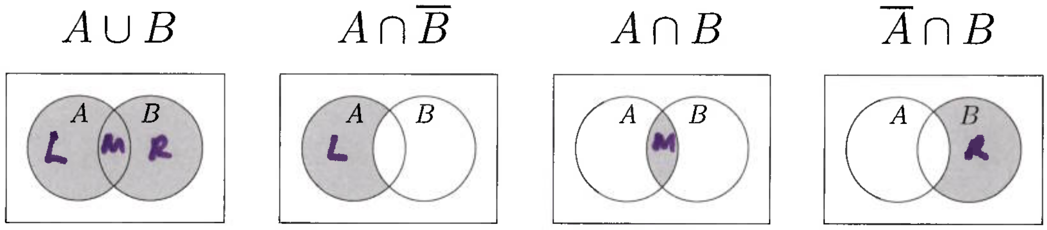
(ii) $P(\text{neither } H \text{ nor } N) = 1 - P(H \cup N) = 1 - (P(H) + P(N)) = 1 - (\frac{4}{40} + \frac{8}{40}) = \frac{28}{40}$

□

Question —

What is $P(A \cup B)$ when A and B are **<u>not</u>** mutually exclusive?

Recall — (from page 1005)



$A \cup B$ $\quad$ $A \cap \overline{B}$ $\quad$ $A \cap B$ $\quad$ $\overline{A} \cap B$

Left
Middle
Right

We can see that

$$(A \cup B) = \overbrace{(A \cap \overline{B}) \;\cup\; (A \cap B) \;\cup\; (\overline{A} \cap B)}^{\text{no overlap}}$$

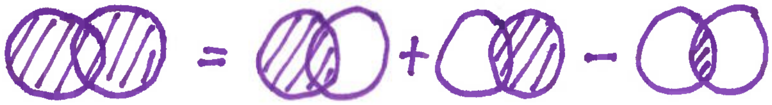
i.e., $A \cup B$ can be expressed as the union of *three mutually exclusive events*.

$$\begin{aligned}
P(A \cup B) &= P(A \cap \overline{B}) + P(A \cap B) + P(\overline{A} \cap B) \\
&= P(A \cap \overline{B}) + P(A \cap B) + P(\overline{A} \cap B) \\
&\qquad\qquad + P(A \cap B) - P(A \cap B) \\
&= \Big(P(A \cap \overline{B}) + P(A \cap B)\Big) \\
&\qquad + \Big(P(\overline{A} \cap B) + P(A \cap B)\Big) - P(A \cap B) \\
&= P(A) + P(B) - P(A \cap B)
\end{aligned}$$

In general:

- For any events A and B

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



Example 1.7 For some sample space S it is known that

$$P(A) = 0.5 \qquad \text{and} \qquad P(B) = 0.6$$

Determine the minimum and maximum possible values of $P(A \cap B)$.

Solution.

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)} \rightarrow \text{always applies}$$

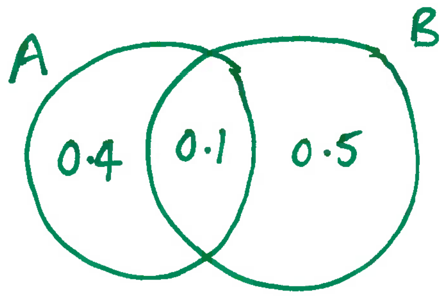
$$P(A \cup B) = 0.5 + 0.6 - P(A \cap B)$$

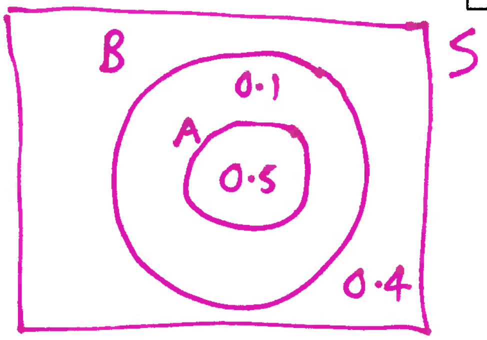
$$P(A \cup B) = 1.1 - P(A \cap B)$$

- Since $P(A \cup B) \leq 1$ the minimum value of $P(A \cap B)$ is 0.1. Note that when $P(A \cap B) = 0.1$, we have $P(A \cup B) = 1$.

- Since $A \cap B \subseteq A$ and $A \cap B \subseteq B$ the maximum value of $P(A \cap B)$ is $\min\{P(A), P(B)\} = 0.5$.

□





Question —

What is $P(\overline{A})$?

Since A and $\overline{A}$ are complementary we have

(no overlap)

- $A \cap \overline{A} = \emptyset$, i.e., A and $\overline{A}$ are mutually exclusive, so
$$P(A \cup \overline{A}) = P(A) + P(\overline{A})$$

sample space

- $A \cup \overline{A} = S$, so
$$P(A \cup \overline{A}) = 1$$

- Therefore
$$P(A) + P(\overline{A}) = 1$$

*

■ For any event A

$$P(\overline{A}) \;=\; 1 - P(A)$$

Complement law of probability.

Practice Problem. The events A and B are such that

$$P(A) = 0.4 \qquad P(\overline{B}) = 0.3 \qquad P(A \cap B) = 0.2$$

Determine $P(A \cup B)$ and $P(\overline{A} \cap \overline{B})$.

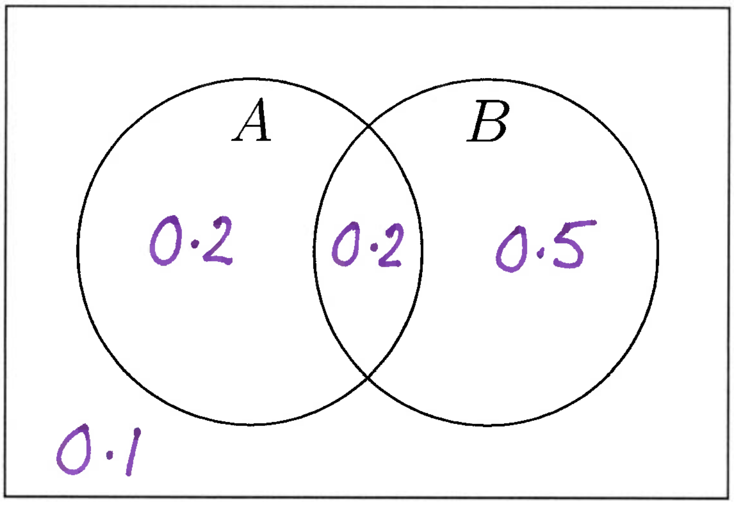
$$P(B) = 1 - P(\overline{B}) = 1 - 0.3 = 0.7$$

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.4 + 0.7 - 0.2 \\ &= 0.9 \end{aligned}$$

De Morgan's Laws: $\overline{A} \cap \overline{B} = \overline{A \cup B}$

$\overline{A} \cup \overline{B} = \overline{A \cap B}$

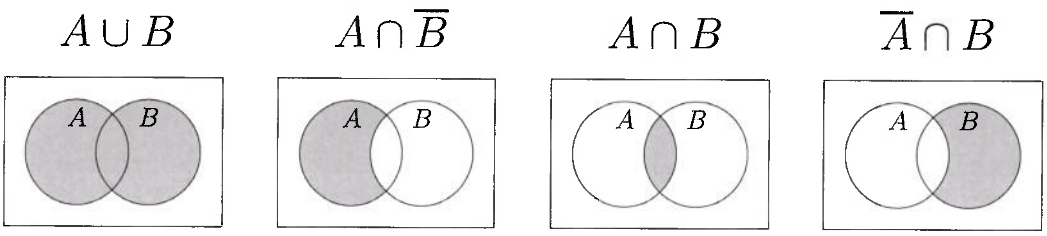
$$\begin{aligned} P(\overline{A} \cap \overline{B}) &= P(\overline{A \cup B}) = 1 - P(A \cup B) \\ &= 1 - 0.9 = 0.1 \end{aligned}$$



□

Summary

- *Experiment* (or *trial*) — any operation or procedure whose outcome cannot be predicted with certainty.
- *Sample space*, S — set of all possible outcomes associated with the experiment.
- *Event* — some subset of the sample space.
- Suppose, an experiment has n equally likely outcomes, and an event E occurs if any one of k of these outcomes is an outcome of the experiment. Then $P(E) = \frac{k}{n}$.
- If an experiment is repeated many times, the probability of an event is the proportion of times the event occurs in n repetitions.



$A \cup B$ $\quad A \cap \overline{B}$ $\quad A \cap B$ $\quad \overline{A} \cap B$

- Suppose the sample space S is finite.

  Axiom 1. $0 \leq P(A) \leq 1$ for each event A in S

  Axiom 2. $P(S) = 1$

  Axiom 3. If A and B are mutually exclusive events in S, then

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

Mutually exclusive events have empty intersections.

- For any events A and B

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

- For any event A we have $P(\overline{A}) = 1 - P(A)$