Sample space

- **Sample space (population)** Ω:
  - Set of possible outcomes of some experiment.
  - Example:
    - Experiment: randomly select a student among all UST postgraduate students.
    - Sample space Ω: the set of all UST postgraduate students.

The set of possible outcomes of an "experiment" is called the sample space

- Throwing a six sided die: $\{1, 2, 3, 4, 5, 6\}$.
- Will Denmark win the world cup: {yes,no}.
- The values in a deck of cards: $\{2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A\}$.

- Elements of the sample spaces are called **samples**.
  - Subsets of sample spaces are **events**.

- Examples:
  - Sample space Ω: the set of all UST postgraduate students.
  - $E_{\text{female}} = \{\text{female students}\}$
    the randomly selected student is a female.
  - $E_{\text{male}} = \{\text{male students}\}$
    the randomly selected student is a male.
  - $E_{\text{MPhil}} = \{\text{MPhil students}\}$
    the randomly selected student is an MPhil student.
  - $E_{\text{PhD}} = \{\text{PhD students}\}$
    the randomly selected student is a PhD student.

- The event that we will get an even number when throwing a die: $\{2, 4, 6\}$.
- The event that Denmark wins: {yes}.
- The event that we will get a 6 or below when drawing a card: $\{2, 3, 4, 5, 6\}$.

Probability measure

- A **probability measure** is a mapping from the set of **events** to [0, 1]

$$P : 2^{\Omega} \rightarrow [0, 1]$$

that satisfies Kolmogorov's axioms:

1. $P(\Omega) = 1$.
2. $P(A) \geq 0 \; \forall A \subseteq \Omega$
3. **Additivity**: $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$.

- Example:
  - Sample space Ω: the set of all UST postgraduate students.
  - Define probability measure: $P(A) = |A|/|\Omega|$.
    - $P(E_{\text{female}})$ = 'fraction of female postgraduate students

Random Variables

- **Random variable X**:
  - Function defined over sample space.
  - Example:
    - Gender of (randomly selected) student,
    - Programme of (randomly selected) student

- **Domain of a random variable Ω_X**:
  - the set of possible states of X.
  - Example:

$$\Omega_{\text{Gender}} = \{\text{f}, \text{m}\}$$

- For any state x of a random variable X, let

$$\Omega_{X=x} = \{\omega \in \Omega | X(\omega) = x\}$$

This is an event

- Example:
  $\Omega_{\text{Gender=f}} = \{$ female postgraduate students in UST$\} = E_{\text{female}}$.
- Note: we use upper case letters, e.g. X, for variables and lower case letters, e.g. x, for states of variables.
- Note the difference between Ω_X and $\Omega_{X=x}$

Probability mass function (distribution)

- **Probability mass function** of a random variable X:

$$P(X) : \Omega_X \rightarrow [0, 1]$$

$$P(X = x) = P(\Omega_{X=x})$$

- Examples:
  - $P(\text{Gender=f}) = P(E_{\text{female}}) = 1/6$ (Assumption)
  - $P(\text{Gender=m}) = P(E_{\text{male}}) = 5/6$.
  - $P(\text{Programme=MPhil}) = P(E_{\text{MPhil}}) = 1/3$ (Assumption
  - $P(\text{Programme=PhD}) = P(E_{\text{PhD}}) = 2/3$.

Because of Kolmogorov's axioms, a probability mass function completely determines a probability measure.

Frequentist interpretation

- **Frequentist interpretation**:
- Probability is long term frequency

- Example:
  - X is result of coin tossing. $\Omega_X = \{\mathrm{H}, \mathrm{T}\}$
  - $P(X{=}\mathrm{H}) = 1/2$ means that
    - *the frequency of getting heads* approaches 1/2 as the number of tosses goes to infinite.
  - Justified by the Law of Large Numbers:
    - X_i: result of the i-th tossing; 1 – H, 0 — T
    - Law of Large Numbers:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{2} \qquad \text{with probability 1}$$

- The frequentist interpretation is meaningful only when experiment can be repeated.

Subjectivist interpretation

- Probabilities are logically consistent degrees of beliefs.

- Comes into play when experiment not repeatable.
- Depends on a person's background knowledge.

- Subjective: another person with different background knowledge might have different probability.
- Experiment not repeatable. If I go to library and find out the truth, my background knowledge is no longer the same.

- The subjectivist interpretation was not widely accepted until 1970s

- This is a major reason why probability theory did not play a big role in AI before 1980.
  - Because probability was defined as statistical frequency and hence was seen as a technique that was appropriate only when statistical data were available.
  - Not many interesting applications with statistical data at that time. Now, more common.
- Now both interpretations are accepted. In practice, subjective beliefs and statistical data complement each other.
  - We rely on subjective beliefs (prior probabilities) when data are scarce.
  - As more and more data become available, we rely less and less on subjective beliefs.

  - As we will learn later, probability has a numerical aspect as well as a structural aspect.
    - We will rely more on the subjectivity interpretation when it comes to building structures than estimating numbers. Our belief on "causality" often plays an important role when building structures.
- The subjectivist interpretation makes concepts such as conditional independence easy to understand.

Joint probability mass function

- **Probability mass function** of a random variable X:

$$P(X) : \Omega_X \rightarrow [0, 1]$$

- Suppose there are n random variables X_1, X_2, ..., X_n.
- A **joint probability mass function**, $P(X_1, X_2, \ldots, X_n)$, over those random variables is:
  - a probability mass function defined on the Cartesian product of their state spaces:

$$\prod_{i=1}^{n} \Omega_{X_i} \rightarrow [0, 1]$$

Joint probability distribution

- The joint distribution $P(X_1, X_2, \ldots, X_n)$ contains information about all aspects of the relations among the n random variables.
- In theory, one can answer any query about relations among the variables based on the joint probability.

- Example:
  - Population: Apartments in Hong Kong rental market.
  - Random variables: (of a random selected apartment)
    - Monthly Rent: {low ($\leq 1k$), medium ((1k, 2k]), upper medium((2k, 4k]), high ($\geq$4k)},
    - Type: {public, private, others}
  - Joint probability distribution $P(\text{Rent}, \text{Type})$:

	public	private	others
low	.17	.01	.02
medium	.44	.03	.01
upper medium	.09	.07	.01
high	0	0.14	0.1

- What is the probability of a randomly selected apartment being a public one?

P(Type=pulic) = P(Type=public, Rent=low)+P(Type=public, Rent=medium)+ P(Type=public, Rent=upper medium)+ P(Type=public, Rent=high) = .7

P(Type=private) = P(Type=private, Rent=low)+ P(Type=private, Rent=medium)+ P(Type=private, Rent=upper medium)+ P(Type=private, Rent=high)= .25

	public	private	others	P(Rent)
low	.17	.01	.02	.2
medium	.44	.03	.01	.48
upper medium	.09	.07	.01	.17
high	0	0.14	0.1	.15
P(Type)	.7	.25	.05	

- Called marginal probability because written on the margins.

Marginal probability

$$\mathrm{P}(\mathrm{Type}) = \sum_{\mathrm{Rent}} \mathrm{P}(\mathrm{Type}, \mathrm{Rent})$$

- The operation is called **marginalization**: Variable "Rent" is marginalized from the joint probability P(Type, Rent).

- Notations for more general cases:
  - $$P(X, Y) = \sum_{U,V} P(X, Y, U, V).$$
  - $\mathbf{Y} \subset \{X_1, X_2, \ldots, X_n\}$, $\mathbf{Z} = \{X_1, X_2, \ldots, X_n\} - \mathbf{Y}$,

    $$P(\mathbf{Y}) = \sum_{\mathbf{Z}} P(X_1, X_2, \ldots, X_n)$$

- A joint probability gives us a full picture about how random variables are related.
- Marginalization lets us to focus one aspect of the picture.

The probabilistic approach to reasoning under uncertainty

- A problem domain is modeled by a list of variables $X_1, X_2, \dots, X_n$,
- Knowledge about the problem domain is represented by a joint probability $P(X_1, X_2, \dots, X_n)$.

Example: Alarm (Pearl 1988)

- Story: In LA, burglary and earthquake are not uncommon. They both can cause alarm. In case of alarm, two neighbors John and Mary may call.
- Problem: Estimate the probability of a burglary based who has or has not called.
- Variables: Burglary (B), Earthquake (E), Alarm (A), JohnCalls (J), MaryCalls (M).
- Knowledge required by the probabilistic approach in order to solve this problem:

$$P(B, E, A, J, M)$$

$$P(B, E, A, J, M)$$

B	E	A	J	M	Prob	B	E	A	J	M	Prob
y	y	y	y	y	.00001	n	y	y	y	y	.0002
y	y	y	y	n	.000025	n	y	y	y	n	.0004
y	y	y	n	y	.000025	n	y	y	n	y	.0004
y	y	y	n	n	.00000	n	y	y	n	n	.0002
y	y	n	y	y	.00001	n	y	n	y	y	.0002
y	y	n	y	n	.000015	n	y	n	y	n	.0002
y	y	n	n	y	.000015	n	y	n	n	y	.0002
y	y	n	n	n	.0000	n	y	n	n	n	.0002
y	n	y	y	y	.00001	n	n	y	y	y	.0001
y	n	y	y	n	.000025	n	n	y	y	n	.0002
y	n	y	n	y	.000025	n	n	y	n	y	.0002
y	n	y	n	n	.0000	n	n	y	n	n	.0001
y	n	n	y	y	.00001	n	n	n	y	y	.0001
y	n	n	y	n	.00001	n	n	n	y	n	.0001
y	n	n	n	y	.00001	n	n	n	n	y	.0001
y	n	n	n	n	.00000	n	n	n	n	n	.996

Inference with joint probability distribution

- What is the probability of burglary given that Mary called, $P(B{=}y|M{=}y)$?
- Compute *marginal probability*:

$$P(B,M) = \sum_{E,A,J} P(B,E,A,J,M)$$

B	M	Prob
y	y	.000115
y	n	.000075
n	y	.00015
n	n	.99971

- Compute answer (reasoning by conditioning):

$$\begin{aligned} P(B{=}y|M{=}y) &= \frac{P(B{=}y, M{=}y)}{P(M{=}y)} \\ &= \frac{.000115}{.000115 + 000075} = 0.61 \end{aligned}$$

Conditional probability

- For events A and B:

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

- Meaning:
  - $P(A)$: my probability on A (without any knowledge about B)
  - $P(A|B)$: My probability on event A assuming that I know event B is true.
- What is the probability of a randomly selected private apartment having "low" rent?

$$P(Rent{=}low|Type{=}private) = \frac{P(Rent{=}Low,\ Tpe{=}private)}{P(Type{=}private)} = .01/.25{=}.04$$

In contrast:

$$P(Rent{=}low) = 0.2.$$

Properties of Conditional Probability

- The conditional probability of an event A, given an event B with $\mathbf{P}(B) > 0$, is defined by

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)},$$

  and specifies a new (conditional) probability law on the same sample space Ω. In particular, all properties of probability laws remain valid for conditional probability laws.

- Conditional probabilities can also be viewed as a probability law on a new universe B, because all of the conditional probability is concentrated on B.

- If the possible outcomes are finitely many and equally likely, then

$$\mathbf{P}(A \mid B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.$$

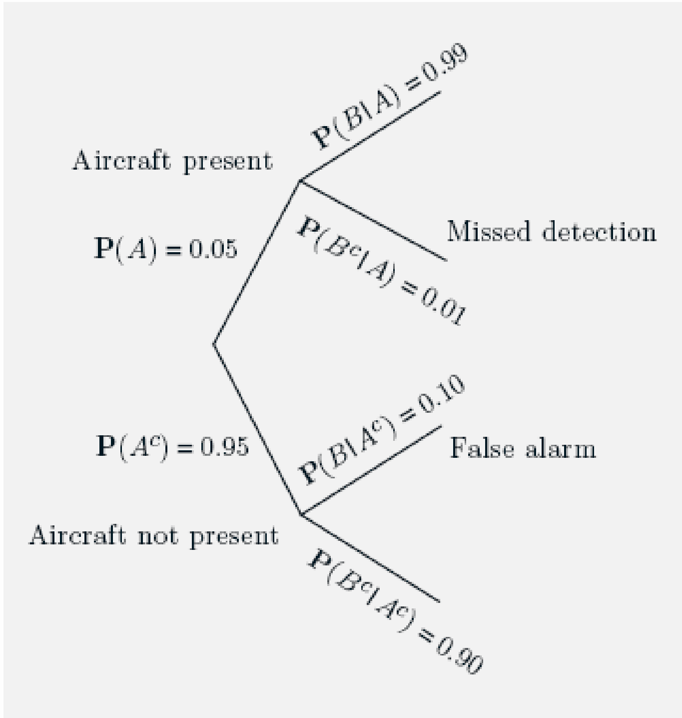
Example 1.9. Radar Detection. If an aircraft is present in a certain area, a radar detects it and generates an alarm signal with probability 0.99. If an aircraft is not present, the radar generates a (false) alarm, with probability 0.10. We assume that an aircraft is present with probability 0.05. What is the probability of no aircraft presence and a false alarm? What is the probability of aircraft presence and no detection?

$$A = \{\text{an aircraft is present}\},$$

$$B = \{\text{the radar generates an alarm}\},$$

$$A^c = \{\text{an aircraft is not present}\},$$

$$B^c = \{\text{the radar does not generate an alarm}\}.$$



$$\mathbf{P}(\text{not present, false alarm}) = \mathbf{P}(A^c \cap B) = \mathbf{P}(A^c)\mathbf{P}(B \mid A^c) = 0.95 \cdot 0.10 = 0.095,$$
$$\mathbf{P}(\text{present, no detection}) = \mathbf{P}(A \cap B^c) = \mathbf{P}(A)\mathbf{P}(B^c \mid A) = 0.05 \cdot 0.01 = 0.0005.$$

- P(Rent|Type)

	public	private	others
low	.17/.7	.01/.25	.02/.05
medium	.44/.7	.03/.25	.01/.05
upper medium	.09/.7	.07/.25	.01/.05
high	0/.7	0.14/.25	0.1/.05

- Note that

$$\sum_{Rent} P(Rent|Type) = 1.$$

Marginal independence

- Two random variables X and Y are **marginally independent**, written $X \perp Y$, if
  - for any state x of X and any state y of Y,

$$P(X{=}x|Y{=}y) = P(X{=}x), \text{ whenever } P(Y = y) \neq 0.$$

- Meaning: Learning the value of Y does not give me any information about X and vice versa. Y contains no information about X and vice versa.
- Equivalent definition:

$$P(X{=}x, Y{=}y) = P(X{=}x)P(Y{=}y)$$

- Shorthand for the equations:

$$P(X|Y) = P(X), P(X, Y) = P(X)P(Y).$$

- Examples:
  - X:result of tossing a fair coin for the first time,
    Y: result of second tossing of the same coin.
  - X: result of US election, Y: your grades in this course.
- Counter example:X – oral presentation grade , Y – project report grade.

Conditional independence

- Two random variables X and Y are **conditionally independent** given a third variable Z,written $X \perp Y | Z$, if

$$P(X{=}x|Y{=}y, Z{=}z) = P(X{=}x|Z{=}z) \text{ whenever } P(Y{=}y, Z{=}z) \neq 0$$

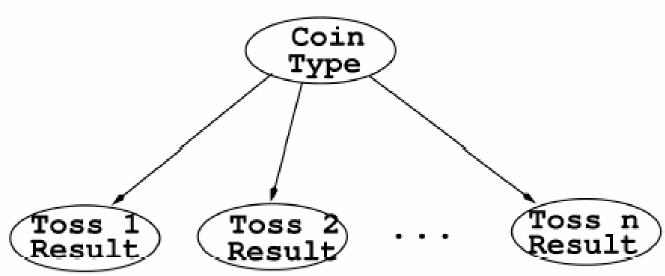
- Meaning:
  - *If I know the state of Z already, then learning the state of Y does not give me additional information about X.*
  - Y might contain some information about X.
  - However all the information about X contained in Y are also contained in Z.
- Shorthand for the equation:

$$P(X|Y, Z) = P(X|Z)$$

- Equivalent definition:

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head. Tossing such a coin results in head 80% of the time. The other coins are fair.
- Randomly draw a coin from the bag and toss it a few time.
- X_i: result of the i-th tossing, Y: whether the coin is produced by the malfunctioning machine.
- The X_i's are not marginally independent of each other:
  - If I get 9 heads in first 10 tosses, then the coin is probably a biased coin. Hence the next tossing will be more likely to result in a head than a tail.
  - Learning the value of X_i gives me some information about whether the coin is biased, which in term gives me some information about X_j.



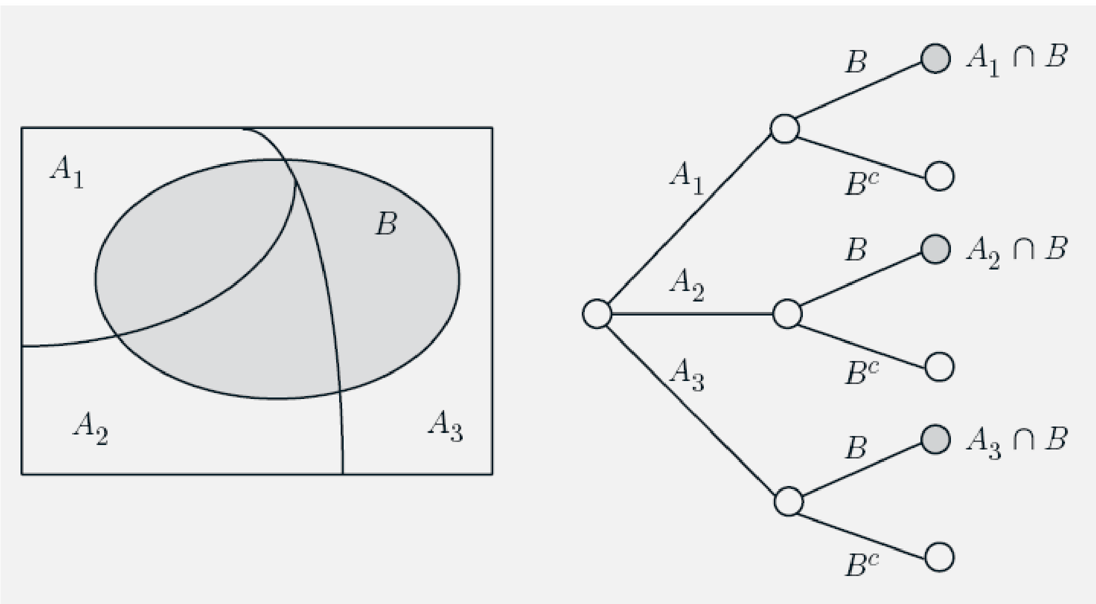
- However, they are conditionally independent given Y:
  - If the coin is not biased, the probability of getting a head in one toss is $1/2$ regardless of the results of other tosses.
  - If the coin is biased, the probability of getting a head in one toss is 80% regardless of the results of other tosses.
  - If I already knows whether the coin is biased or not, learning the value of X_i does not give me additional information about X_j.

Total Probability Theorem

Total Probability Theorem

Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in exactly one of the events $A_1, \ldots, A_n$) and assume that $\mathbf{P}(A_i) > 0$, for all i. Then, for any event B, we have

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1 \cap B) + \cdots + \mathbf{P}(A_n \cap B) \\ &= \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n).\end{aligned}$$



Example 1.13. You enter a chess tournament where your probability of winning a game is 0.3 against half the players (call them type 1), 0.4 against a quarter of the players (call them type 2), and 0.5 against the remaining quarter of the players (call them type 3). You play a game against a randomly chosen opponent. What is the probability of winning?

Let A_i be the event of playing with an opponent of type i. We have

$$\mathbf{P}(A_1) = 0.5, \qquad \mathbf{P}(A_2) = 0.25, \qquad \mathbf{P}(A_3) = 0.25.$$

Also, let B be the event of winning. We have

$$\mathbf{P}(B \mid A_1) = 0.3, \qquad \mathbf{P}(B \mid A_2) = 0.4, \qquad \mathbf{P}(B \mid A_3) = 0.5.$$

Thus, by the total probability theorem, the probability of winning is

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \mathbf{P}(A_2)\mathbf{P}(B \mid A_2) + \mathbf{P}(A_3)\mathbf{P}(B \mid A_3) \\ &= 0.5 \cdot 0.3 + 0.25 \cdot 0.4 + 0.25 \cdot 0.5 \\ &= 0.375.\end{aligned}$$

Prior, posterior, and likelihood

- **Prior probability**: belief about a hypothesis h before obtaining observations, $P(h)$.
  - Example: Suppose 10% of people suffer from Hepatitis B. A doctor's prior probability about a new patient suffering from Hepatitis B is 0.1.
- **Posterior probability**:belief about a hypothesis after obtaining observations.
- **Likelihood** of hypothesis given observation:
  - Conditional probability of observation given hypothesis $L(h|o) = P(o|h)$
  - Example: o: eye-color=yellow; h_1: Hepatitis B; h_2: no Hepatitis B

$$P(o|h_1) > P(o|h_2)$$

If we observe o, h_1 is more likely than h_2.
As a function of h, $P(o|h)$ measures the likelihood of h.

Bayes' Theorem

- **Bayes' Theorem**: relates prior probability, likelihood, and posterior probability:

$$P(h|o) = \frac{P(h)P(o|h)}{P(o)} \propto P(h)P(o|h) = P(h)L(h|o)$$

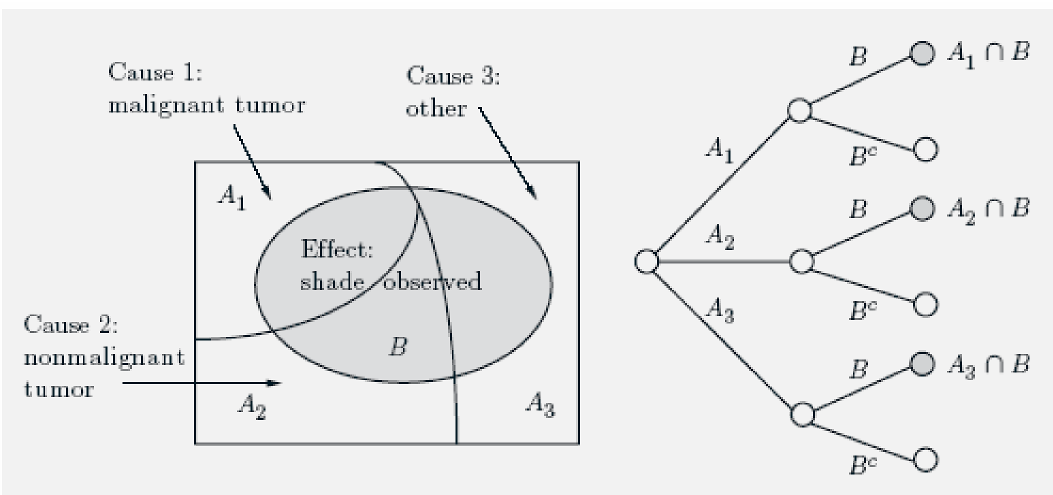
  where $P(o)$ is normalization constant to ensure $\sum_h P(h|o) = 1$.

  In words: posterior $\propto$ prior $\times$ likelihood

- Example:

$$P(\text{disease}|\text{symptoms}) = \frac{P(\text{disease})P(\text{symptoms}|\text{disease})}{P(\text{symptoms})}$$

  - $P(\text{symptom}|\text{disease})$ from understanding of disease,
  - $P(\text{disease}|\text{symptoms})$ needed in clinical diagnosis.



Let us return to the radar detection problem

$$A = \{\text{an aircraft is present}\},$$
$$B = \{\text{the radar generates an alarm}\}.$$

We are given that

$$\mathbf{P}(A) = 0.05, \qquad \mathbf{P}(B \mid A) = 0.99, \qquad \mathbf{P}(B \mid A^c) = 0.1.$$

Applying Bayes' rule, with $A_1 = A$ and $A_2 = A^c$, we obtain

$$\begin{aligned}
\mathbf{P}(\text{aircraft present} \mid \text{alarm}) &= \mathbf{P}(A \mid B) \\
&= \frac{\mathbf{P}(A)\mathbf{P}(B \mid A)}{\mathbf{P}(B)} \\
&= \frac{\mathbf{P}(A)\mathbf{P}(B \mid A)}{\mathbf{P}(A)\mathbf{P}(B \mid A) + \mathbf{P}(A^c)\mathbf{P}(B \mid A^c)} \\
&= \frac{0.05 \cdot 0.99}{0.05 \cdot 0.99 + 0.95 \cdot 0.1} \\
&\approx 0.3426.
\end{aligned}$$

Let us return to the chess problem

$$\mathbf{P}(A_1) = 0.5, \qquad \mathbf{P}(A_2) = 0.25, \qquad \mathbf{P}(A_3) = 0.25.$$

Also, B is the event of winning, and

$$\mathbf{P}(B \mid A_1) = 0.3, \qquad \mathbf{P}(B \mid A_2) = 0.4, \qquad \mathbf{P}(B \mid A_3) = 0.5.$$

Suppose that you win. What is the probability $\mathbf{P}(A_1 \mid B)$ that you had an opponent of type 1?

Using Bayes' rule, we have

$$\begin{aligned}
\mathbf{P}(A_1 \mid B) &= \frac{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \mathbf{P}(A_2)\mathbf{P}(B \mid A_2) + \mathbf{P}(A_3)\mathbf{P}(B \mid A_3)} \\
&= \frac{0.5 \cdot 0.3}{0.5 \cdot 0.3 + 0.25 \cdot 0.4 + 0.25 \cdot 0.5} \\
&= 0.4.
\end{aligned}$$

In 1964 an interracial couple was convicted of robbery in Los Angeles, largely on the grounds that they matched a highly improbable profile, a profile which fit witness reports [272]. In particular, the two robbers were reported to be

- A man with a mustache
- Who was black and had a beard
- And a woman with a ponytail
- Who was blonde

- The couple was interracial
- And were driving a yellow car

The prosecution suggested that these characteristics had the following probabilities of being observed at random in the LA area:

1. A man with a mustache 1/4
2. Who was black and had a beard 1/10
3. And a woman with a ponytail 1/10
4. Who was blonde 1/3
5. The couple was interracial 1/1000
6. And were driving a yellow car 1/10

$$P(e|\neg h) = \prod_i P(e_i|\neg h) = 1/12000000$$

e_i $(i = 1, \ldots, 6)$, the joint evidence e

A Much better estimate

$$P(e_2|\neg h)P(e_3|\neg h)P(e_4|\neg h)P(e_6|\neg h) = 1/3000.$$

The Bayesian approach

$$P(h|e) = \frac{P(e|h)P(h)}{P(e|h)P(h) + P(e|\neg h)P(\neg h)}$$

$$P(h|e) = \frac{P(h)}{P(h) + P(\neg h)/3000}$$

6.5 million people

this gives us 1,625,000 eligible males and as many females

$$P(h|e) = \frac{1/1625000}{1/1625000 + (1 - 1/1625000)/3000} \approx 0.002$$