Probability and Statistics

Kristel Van Steen, PhD^2

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 1: PROBABILITY THEORY

1 What’s in a name

1.1 Relevant questions in a probabilistic context

1.2 Relevant questions in a statistics context

2 Probability and statistics: two related disciplines

2.1 Probability

3 Different flavors of probability

3.1 Classical or a priori probability

3.2 Set theory

3.3 Sample space and probability measures

3.4 A posteriori or frequency probability

4 Statistical independence and conditional probability

4.1 Independence

4.2 Conditional probability

Law of total probability

Bayes’ theorem

Bayesian odds

Principle of proportionality

5 In conclusion

5.1 Take-home messages

5.2 The birthday paradox

1 What’s in a name …

If someone asks you what probability is, can you point out a key question to him/her?

(madamebutterflytoo.com)

1.1 Relevant questions in a probabilistic context

The bear cubs problem

There are two bears - white and dark. We may reasonably ask several questions:

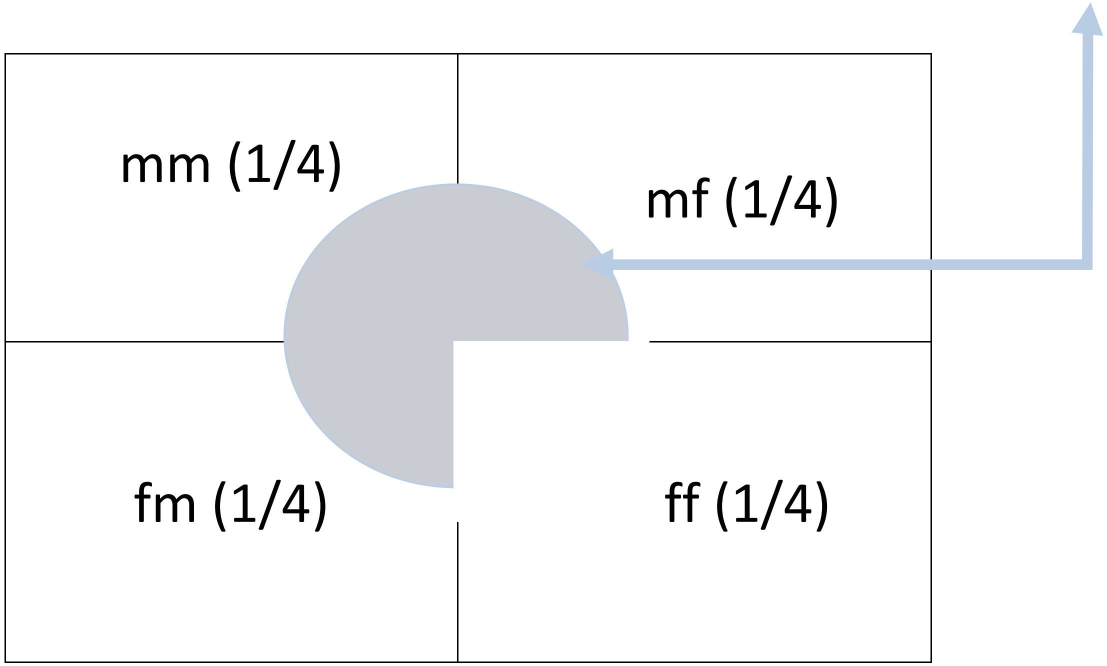
- What is the probability that both bears are male?
  Writing 'm' for male and 'f' for female and counting the lighter bear first we get four possible outcomes (ff, mf, fm, mm) of which only one should be considered favorable. The answer, therefore, is 1/4.

- Now assume I told you that one of the bears is male. What is the probability that both are males?
  Of the three possible outcomes (mf, fm, mm) only the last where both bears are male is favorable. The answer is 1/3.
  - The sample space of the problem is actually (Mf, fM, Mm, mM) … Isn't the answer 1/2?
    - Only the first bear is male. In this event, 0 prob that both are male. Only the second bear is male. In this event, 0 prob that both are male. Both bears are male. In this event, prob 1 that both are male.
  - Note that in general (mf,fm,mm,ff) are 4 equally likely events. Assuming one of these events, the probability of (at least) one bear being male is respectively 1, 1, 1, 0. If one bear was found male, the probabilities of the four possibilities change but the proportionality remains: 1/3,1/3,1/3,0.

Note the different probability assessments: sample space????

	Prob		Prob
ff	¼	mf	1/3
mf or fm	½	fm	1/3
mm	¼	mm	1/3

Event: at least one bear is male



I am telling you that the lighter bear is known to be male. What is now the probability that both of them are males?

- First solution: Since it's now given that the lighter bear is male there are only two possible outcomes (mf, mm). Thus the probability that both are male goes up to 1/2. Note how each additional piece of information changed the number of possibilities and, hence, the probability of the outcome.
- Second solution: The sequence of three questions is supposed to lead one on to wondering what difference it makes to specify that the white bear is male. Since it's now known that the white bear is male, its sex is removed from the realm of random. All that matters is the sex of the dark bear who is believed to be male with the probability of 1/2.
  - A short way to express the same idea is as follows:
    P("both are male" | "white is male") = P("dark is male")
    where P(A|B) means the (conditional) probability of A provided B is known to take place.

If someone asks you what statistics is, can you point out a key question to him/her?

1.2 Relevant questions in a statistics context

- Conceptual questions
  - What is the difference between a “statistic” and a “parameter”?

Statistic: characteristic of a sample

What is the average salary of 2000 people randomly sampled in Spain?

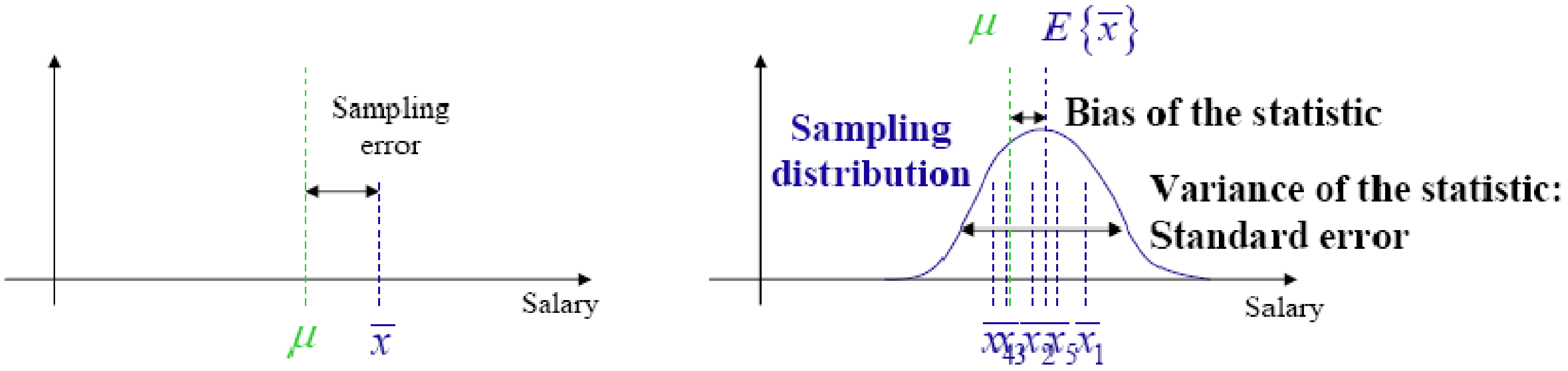
$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Parameter: characteristic of a population

What is the average salary of all Spaniards?

$$\mu$$

- o



- What is the distribution of the statistic?
  - Known
  - Unknown but well-behaved mean (central limit theory)
- Versatile use of the normal distribution?

Mode:

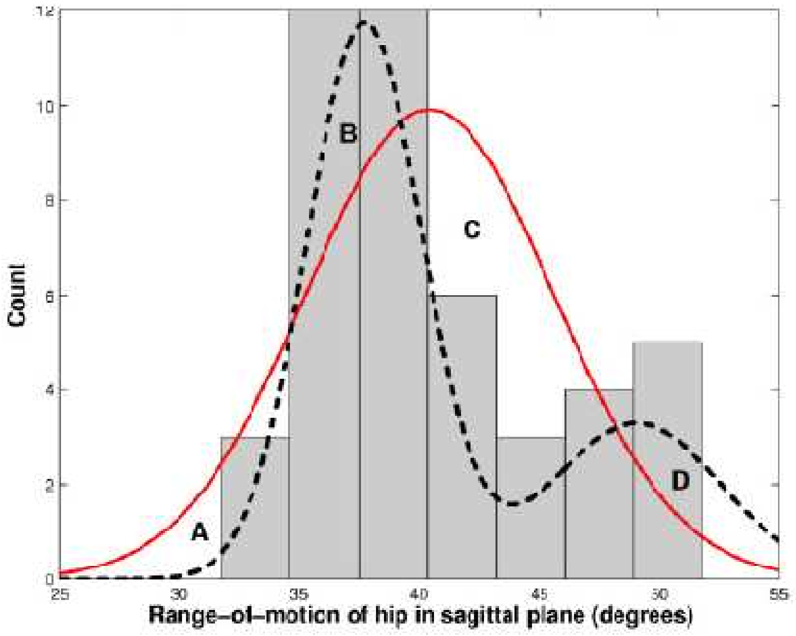
$$x^* = \arg\max f_X(x)$$

Most frequently occurring

(-) Not unique (multimodal)

(−) representative of the most "typical" result

If a variable is multimodal, most central measures fail!



o Are my data really independent?

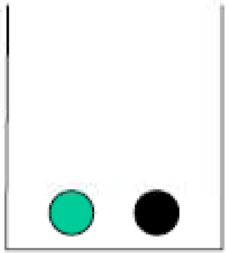
Independence is different from mutual exclusion

In general,

$$\begin{tikzcd}
p(A \cap B) = p(A)\,p(B \mid A) \arrow[d, "p(B \mid A) = 0"] \arrow[dd, bend left=60, "\text{Knowing A does not give any information about the next event}"] \\
p(A \cap B) = 0 \\
p(A \cap B) = p(A)\,p(B)
\end{tikzcd}$$

Mutual exclusion is when two results are impossible to happen at the same time.

Independence is when the probability of an event does not depend on the results that we have had previously.



Example: Sampling with and without replacement

What is the probability of taking a black ball as second draw, if the first draw is green?

- Questions related to collecting data
    - Basics of experimental design
    - What are controlling variables?
    - How many samples do I need for my test?
    - What if I cannot get more samples? [Resampling: Bootstrapping, jackknife]

- Questions related to extracting information
    - Can I see any interesting association between two variables, or between two populations?
    - Which models could have generated these data?
    - How to estimate a parameter of a distribution?
    - What is my confidence in the results?
    - What if my data are “contaminated”? [Robust statistics]

- Questions related to hypothesis testing
    - How can I know if what I see is “true”?
    - What is a hypothesis test? What is the statistical power? What is a p-value? How to use it? What is the relationship between sample size, sampling error, effect size and power? What are the assumptions of hypothesis testing?
    - How to select the appropriate statistical test?
        - Tests about a population central tendency
        - Tests about a population variability
        - Tests about a population distributions
    - What are the dangers of testing multiple times?

2 Probability and statistics: two related disciplines

2.1 Probability

- One of the fundamental tools of statistics is *probability*.
- Probability is derived from the verb to probe meaning to "find out" what is not too easily accessible or understandable. The word "proof" has the same origin that provides necessary details to understand what is claimed to be true.
- Probability originated from the study of games of chance and gambling during the 16th century.

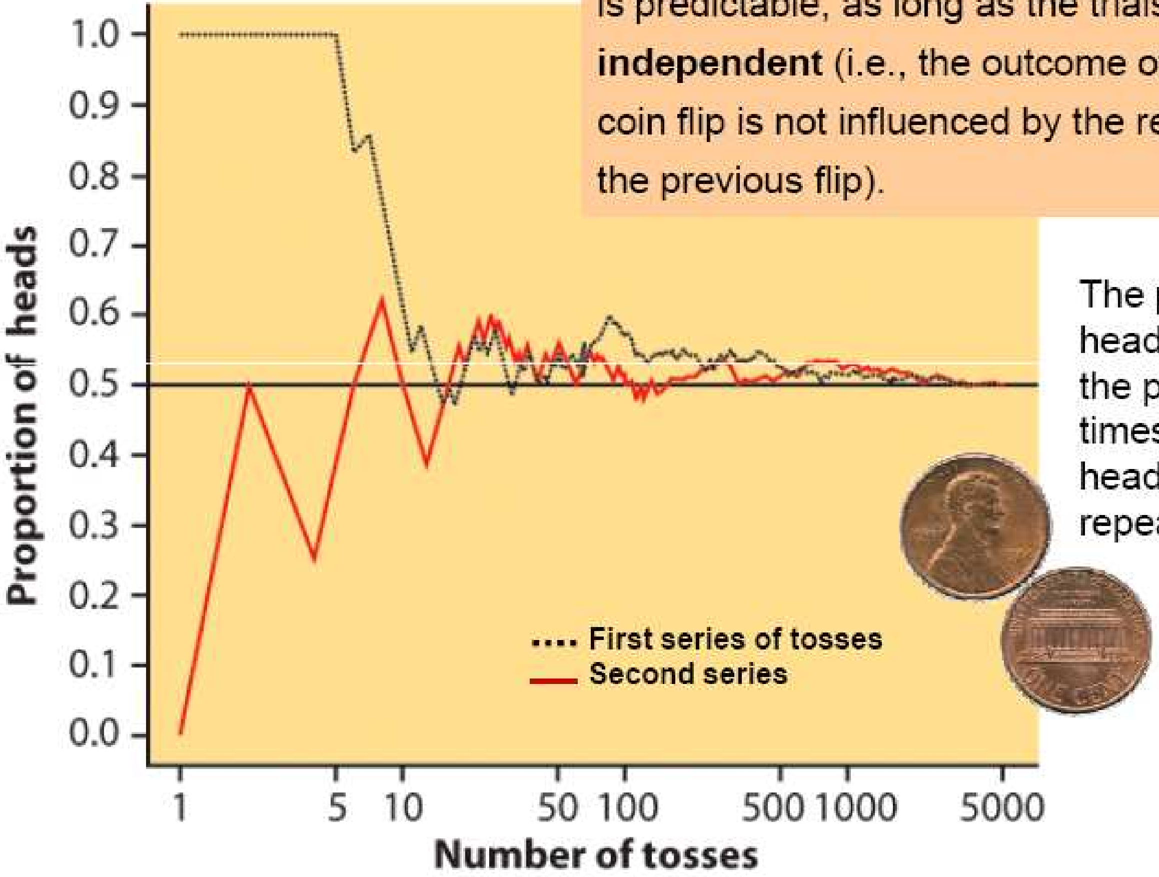
- Probability theory was a branch of mathematics studied by Blaise Pascal and Pierre de Fermat in the seventeenth century. Currently in 21st century, probabilistic modeling is used to control the flow of traffic through a highway system, a telephone interchange, or a computer processor; find the genetic makeup of individuals or populations; quality control; insurance; investment; and other sectors of business and industry.

Blaise Pascal

Pierre de Fermat

Example: coin tossing

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.

- "Fair" in "flipping a fair coin" means, technically, that the probability of heads on a given flip is 50%, and the probability of tails on a given flip is 50%.
- This doesn't mean that every other flip will give a head — after all, three heads in a row is no surprise.
    - Five heads in a row would be more surprising
    - When you've seen twenty heads in a row you're sure that something fishy is going on.
- What the 50% probability of heads does mean is that, as the number of flips increases, we expect the number of heads to approach half the number of flips.
    - So even though the outcome of a particular trial (tossing a coin or spinning a roulette wheel) may be uncertain, there is a predictable long-term outcome
    - Seven heads on ten flips is no surprise; 700,000 heads on 1,000,000 tosses is highly unlikely (note the equal ratio!).

- In probability, we start with a model describing what events we think are going to occur, with what likelihoods.
- The events may be random, in the sense that we don't know for sure what will happen next, but we do quantify our degree of surprise when various things happen.
- In other words, the probabilist starts with a probability model (something which assigns various percentage likelihoods of different things happening), then tells us which things are more and less likely to occur.

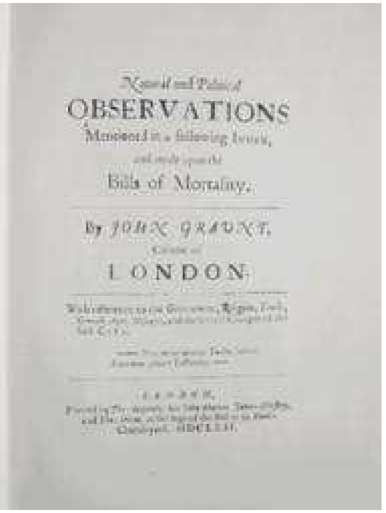
Key points about probability

1. Rules → data: Given the rules, describe the likelihoods of various events occurring.
2. Probability is about prediction — looking forward.
3. Probability is mathematics.

2.2 Statistics

- The original idea of *statistics* was the collection of information about and for the "state". The word statistics derives directly, not from any classical Greek or Latin roots, but from the Italian word for state.
- The birth of statistics occurred in mid-17th century. John Graunt, a native of London, began reviewing a weekly church publication issued by the local parish clerk that listed the number of births, christenings, and deaths in each parish. These so called Bills of Mortality also listed the causes of death. Graunt, who was a shopkeeper, organized these data in the form we call *descriptive statistics*, which was published as Natural and Political Observations Made upon the Bills of Mortality.

(http://www.statisticalforecasting.com/)



Natural and Political
OBSERVATIONS
Mentioned in a following Index,
and made upon the
Bills of Mortality.

By JOHN GRAUNT,
Citizen of
LONDON.

With reference to the Government, Religion, Trade, [illegible]

[illegible]

LONDON,
[illegible]
MDCLXII.

- With this in mind, statistics has to borrow some concepts from sociology, such as the concept of *population*. It has been argued that since statistics usually involves the study of human behavior, it cannot claim the precision of the physical sciences.
- Although new and ever growing diverse fields of human activities are using statistics, the field itself remains obscure to the larger public.

During the 20th Century statistical thinking and methodology have become the scientific framework for literally dozens of fields including education, agriculture, economics, biology, and medicine, and with increasing influence recently on the hard sciences such as astronomy, geology, and physics. In other words, we have grown from a small obscure field into a big obscure field.

(Professor Bradley Efron)

Example: coin tossing revisited

- Suppose you are given a list of heads and tails (= data). You, as the statistician, are in the following situation:
    - You do not know ahead of time that the coin is fair. Maybe you've been hired to decide whether the coin is fair (or, more generally, whether a gambling house is committing fraud).
    - You may not even know ahead of time whether the data come from a coin-flipping experiment at all.
- Suppose the data are three heads out of 7.
    - Your first guess might be that a fair coin is being flipped, and these data don't contradict that hypothesis. Based on these data, you might hypothesize that the rules governing the experiment are that of a fair coin: your probability model for predicting the future is that heads and tails each occur with 50% likelihood.

- Suppose there are ten heads in a row, though, or twenty.
    - You might start to reject the hypothesis of a fair coin and replace it with the hypothesis that the coin has heads on both sides. Then you would predict that the next toss will certainly be heads: your new probability model for predicting the future is that heads occur with 100% likelihood, and tails occur with 0% likelihood.
- Suppose the data are “heads, tails, heads, tails, heads, tails”.
    - Again, your first fair-coin hypothesis seems plausible.
    - If on the other hand you have heads alternating with tails not three pairs but 50 pairs in a row, then you reject that model. It begins to sound like the coin is not being flipped in the air, but rather is being flipped with a spatula. Your new probability model is that if the previous result was tails or heads, then the next result is heads or tails, respectively, with 100% likelihood.

In a sense, probability doesn't need statistics, but statistics uses probability.

Key points about statistics

1. Rules ← data: Given only the data, try to guess what the rules were. That is, some probability model controlled what data came out, and the best we can do is guess — or approximate — what that model was. We might guess wrong; we might refine our guess as we get more data.
2. Statistics is about looking backward.
3. Statistics is an art. It uses mathematical methods, but it is more than maths.
4. Once we make our best *statistical guess* about what the probability model is (what the rules are), based on looking *backward*, we can then use that *probability* model to predict the *future* →
   The purpose of statistics is to make inference about unknown quantities from samples of data

3 Different flavors of probability

3.1 Classical or a priori probability

- The classical definition of probability is prompted by the close association between the theory of probability of the early ages and games of chance.

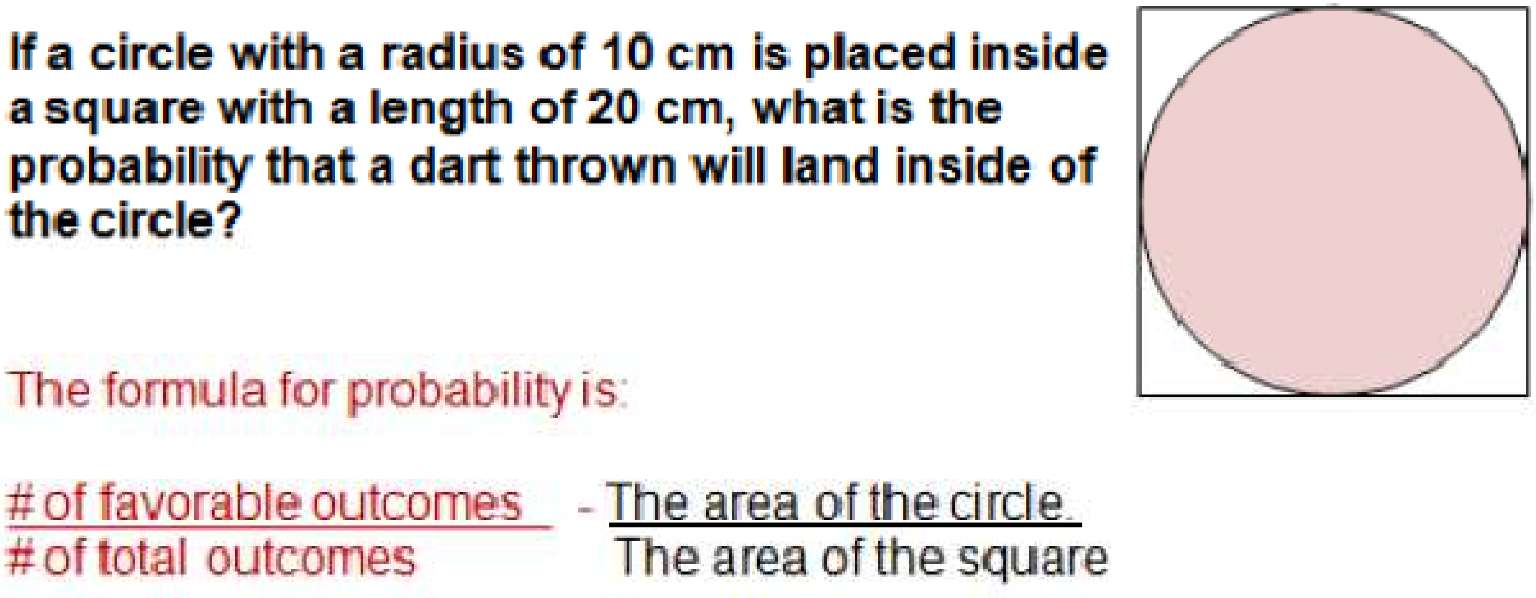
  Classical probability: If a random experiment can result in n mutually exclusive and equally likely outcomes and if n_A of these outcomes have an attribute A, then the probability of A is the fraction n_A/ n.

- In this context

  An event: a possible outcome or set of possible outcomes of an experiment or observation. Typically denoted by a capital letter (e.g., A = result of coin toss) [Note: ALWAYS check the particular notations in text books]

Geometric probability

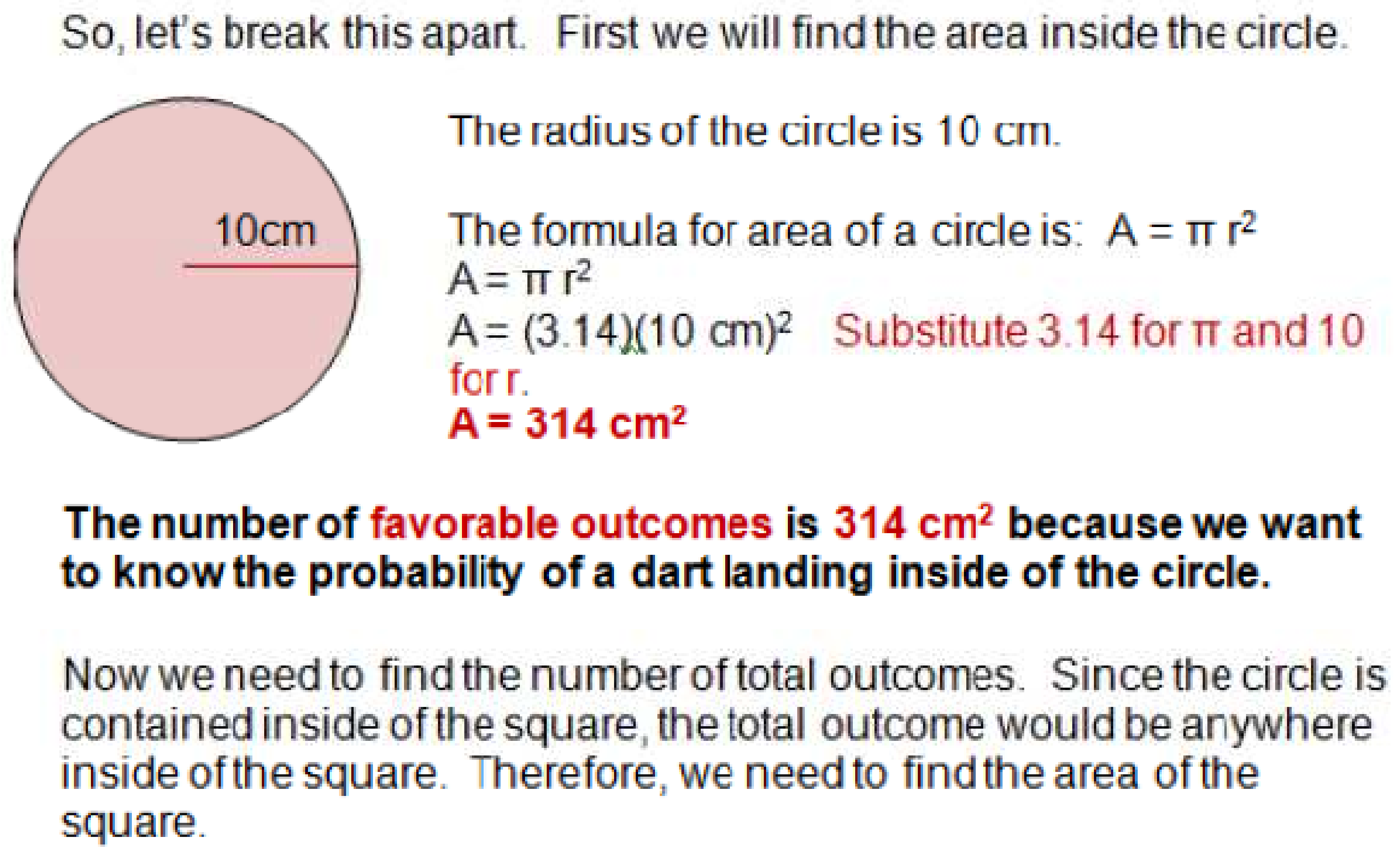
- This is the study of the probabilities involved in geometric problems, e.g., the distributions of length, area, volume, etc. for geometric objects under stated conditions.



If a circle with a radius of 10 cm is placed inside a square with a length of 20 cm, what is the probability that a dart thrown will land inside of the circle?

The formula for probability is:

$$\frac{\text{\# of favorable outcomes}}{\text{\# of total outcomes}} - \frac{\text{The area of the circle.}}{\text{The area of the square}}$$



So, let's break this apart. First we will find the area inside the circle.

The radius of the circle is 10 cm.

The formula for area of a circle is: $A = \pi r^2$

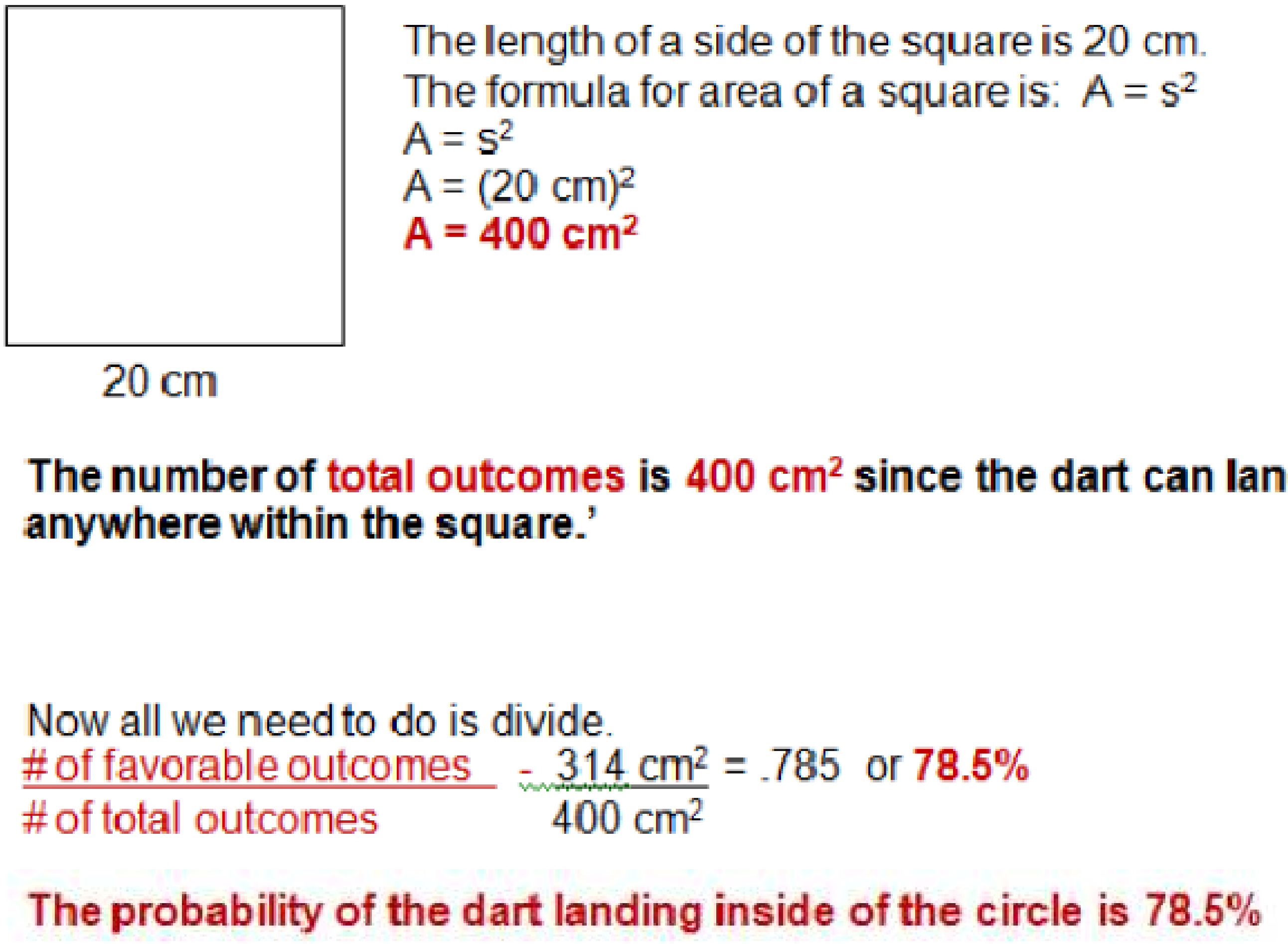
$A = \pi r^2$

$A = (3.14)(10\text{ cm})^2$ Substitute 3.14 for π and 10 for r.

$A = 314\text{ cm}^2$

The number of favorable outcomes is 314 cm^2 because we want to know the probability of a dart landing inside of the circle.

Now we need to find the number of total outcomes. Since the circle is contained inside of the square, the total outcome would be anywhere inside of the square. Therefore, we need to find the area of the square.



20 cm

The length of a side of the square is 20 cm.

The formula for area of a square is: $A = s^2$

$A = s^2$

$A = (20\text{ cm})^2$

$\mathbf{A = 400\text{ cm}^2}$

The number of total outcomes is 400 cm^2 since the dart can land anywhere within the square.'

Now all we need to do is divide.

$$\frac{\#\text{ of favorable outcomes}}{\#\text{ of total outcomes}} - \frac{314\text{ cm}^2}{400\text{ cm}^2} = .785 \text{ or } \mathbf{78.5\%}$$

The probability of the dart landing inside of the circle is 78.5%

(www.algebra-class.com/)

3.2 Set theory

Elements of set theory

- Understanding set theory helps people to … see things in terms of systems, organize things into groups, begin to understand logic
- A set is a collection of objects possessing some common properties. These objects are called *elements* of the set. Sets are denoted by capital letters and elements usually by small letters:

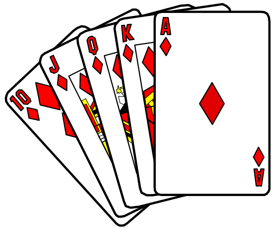
$$A = \{1, 2, 3, 4, 5, 6\},$$

$$B = \{s, f\}.$$

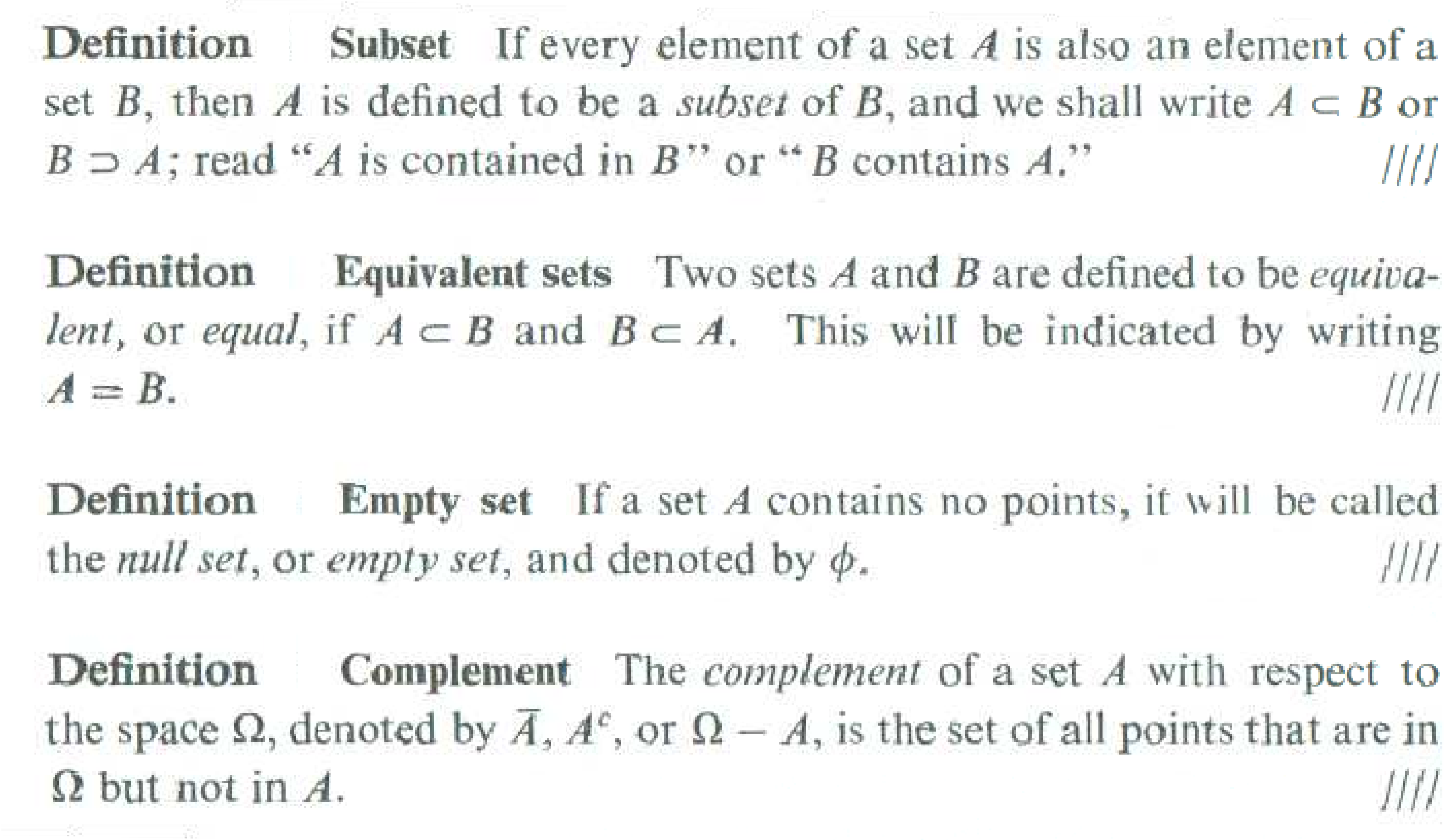
- We use the convention $a \in A$ to mean “element a belongs to set A”

Important set definitions

- Sets containing a finite number of elements are called “finite sets”. Sets containing an infinite number of elements are called “infinite sets”.
- An infinite set is called “enumerable” or “countable” if all of its elements can be arranged in such a way that there is a one-to-one correspondence between them and all positive integers.
    - What is $C = \{x : x \geq 0\}$?
- One particular set is called the “space” and often denoted by S, U or Ω. This “largest” set contains all elements of all the sets under consideration
    - In a deck of ordinary playing cards, each card is an element in the universal set and some subsets are face cards, numbered cards, suits



Important set definitions



Definition **Subset** If every element of a set A is also an element of a set B, then A is defined to be a *subset* of B, and we shall write $A \subset B$ or $B \supset A$; read "A is contained in B" or "B contains A." ////

Definition **Equivalent sets** Two sets A and B are defined to be *equivalent*, or *equal*, if $A \subset B$ and $B \subset A$. This will be indicated by writing $A = B$. ////

Definition **Empty set** If a set A contains no points, it will be called the *null set*, or *empty set*, and denoted by ϕ. ////

Definition **Complement** The *complement* of a set A with respect to the space Ω, denoted by $\bar{A}$, A^c, or $\Omega - A$, is the set of all points that are in Ω but not in A. ////

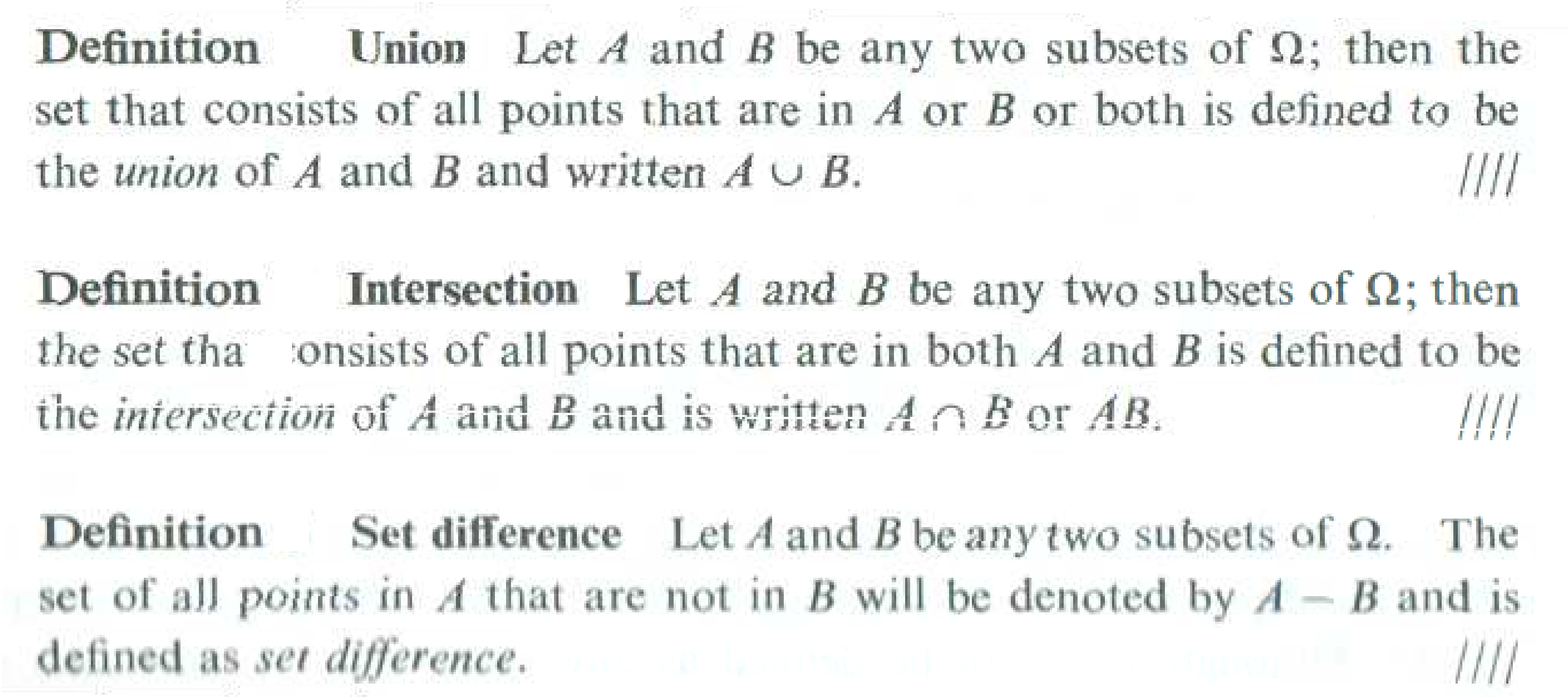
- The **power set** is the set of all subsets that can be created from a given set
  - The *cardinality* (size) of the power set is 2 to the power of the given set's cardinality
  - A power set is usually denoted by $\mathscr{P}$
  - Example*:*

    A = {a, b, c} where |A| = 3 (i.e., the cardinality is 3)

    $\mathscr{P}$(A) = {{a, b}, {a, c}, {b, c}, {a}, {b}, {c}, A, ϕ}
    and | $\mathscr{P}$(A)| = 8

    In general, if |A| = n, then | $\mathscr{P}$(A) | = 2^n

Set operations



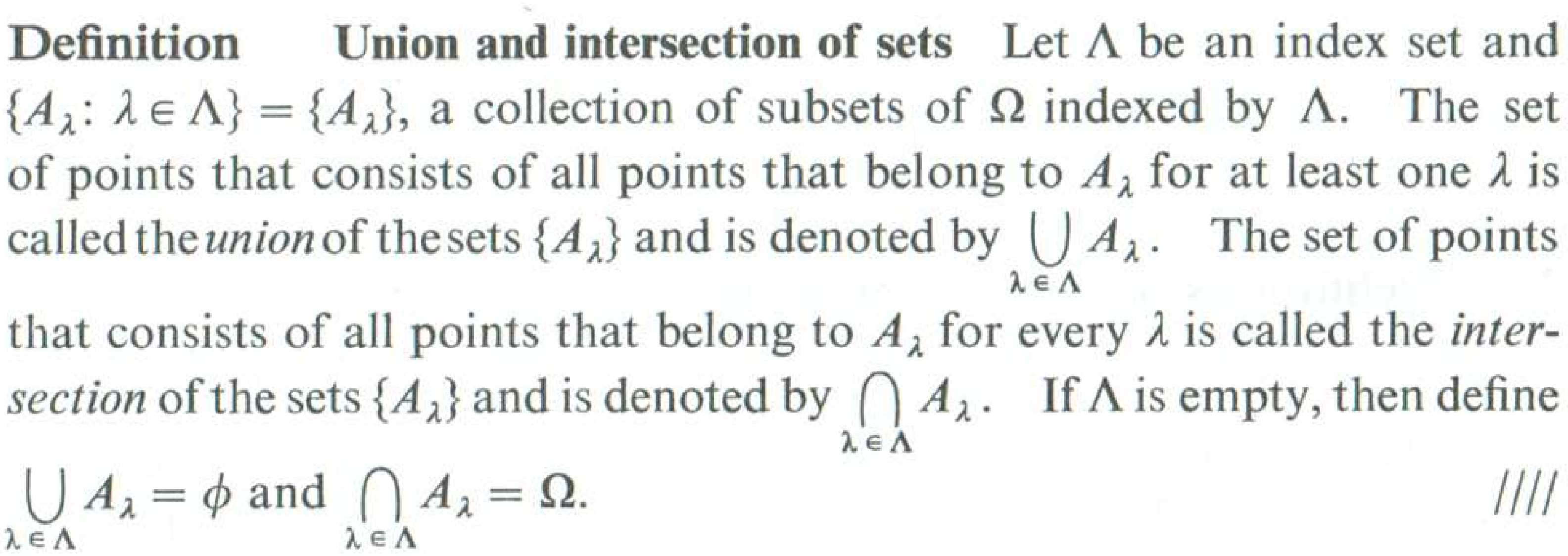
Definition **Union** *Let* A and B be any two subsets of Ω; then the set that consists of all points that are in A or B or both is defined to be the *union* of A and B and written $A \cup B$. ////

Definition **Intersection** Let A *and* B be any two subsets of Ω; then the set that consists of all points that are in both A and B is defined to be the *intersection* of A and B and is written $A \cap B$ or AB. ////

Definition **Set difference** Let A and B be *any* two subsets of Ω. The set of all *points* in A that are not in B will be denoted by $A - B$ and is defined as *set difference*. ////

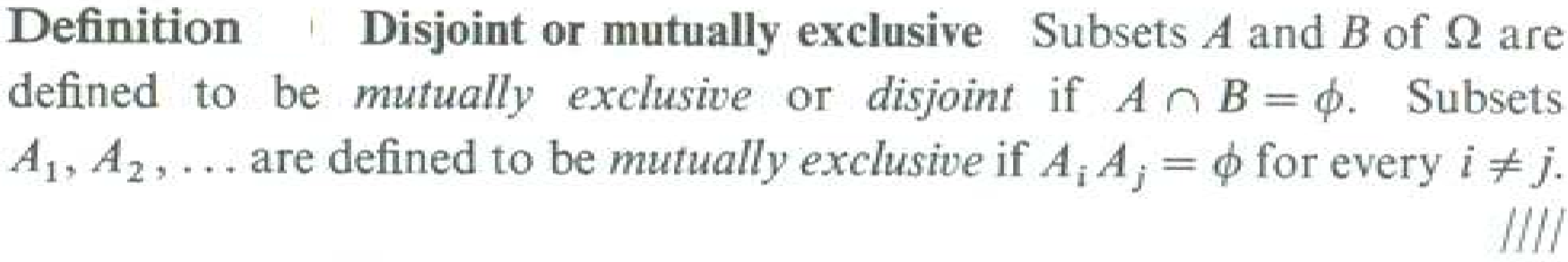
Set operations

- The aforementioned definitions of union and intersection can be directly generalized to those involving any arbitrary number (finite or countable infinite) of sets. In particular:



Definition **Union and intersection of sets** Let Λ be an index set and $\{A_\lambda : \lambda \in \Lambda\} = \{A_\lambda\}$, a collection of subsets of Ω indexed by Λ. The set of points that consists of all points that belong to A_λ for at least one λ is called the *union* of the sets $\{A_\lambda\}$ and is denoted by $\bigcup_{\lambda \in \Lambda} A_\lambda$. The set of points that consists of all points that belong to A_λ for every λ is called the *intersection* of the sets $\{A_\lambda\}$ and is denoted by $\bigcap_{\lambda \in \Lambda} A_\lambda$. If Λ is empty, then define $\bigcup_{\lambda \in \Lambda} A_\lambda = \phi$ and $\bigcap_{\lambda \in \Lambda} A_\lambda = \Omega$. ////

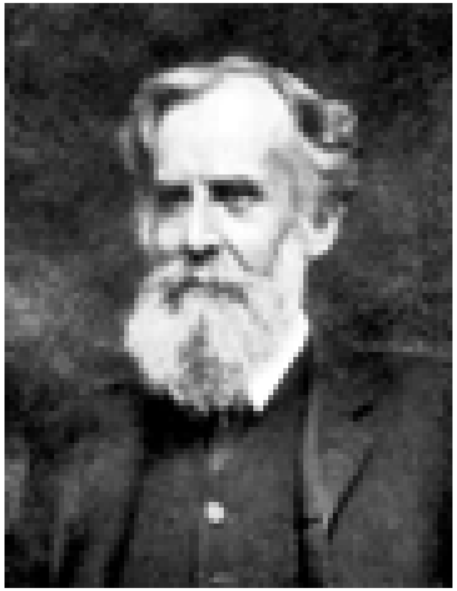
- Recall:

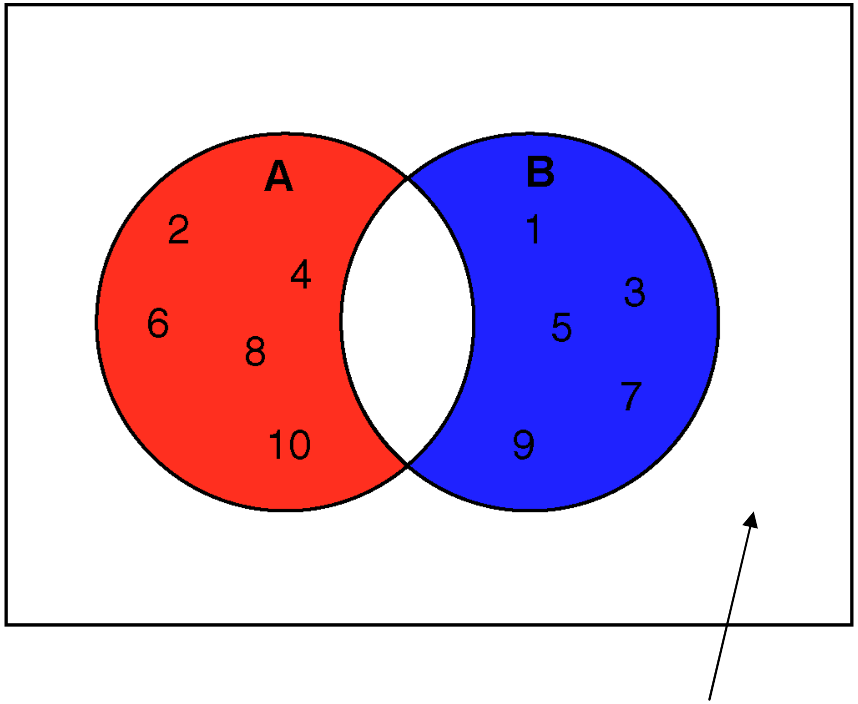


Definition **Disjoint or mutually exclusive** Subsets A and B of Ω are defined to be *mutually exclusive* or *disjoint* if $A \cap B = \phi$. Subsets $A_1, A_2, \ldots$ are defined to be *mutually exclusive* if $A_i A_j = \phi$ for every $i \neq j$.
////

- The symbol “+” is often reserved to denote the union of two sets which are disjoint.
  - For example: $A \cup B = A \cup (\overline{A}B) = A + (\overline{A}B)$

- John Venn devised a simple way to diagram set operations (Venn Diagrams)

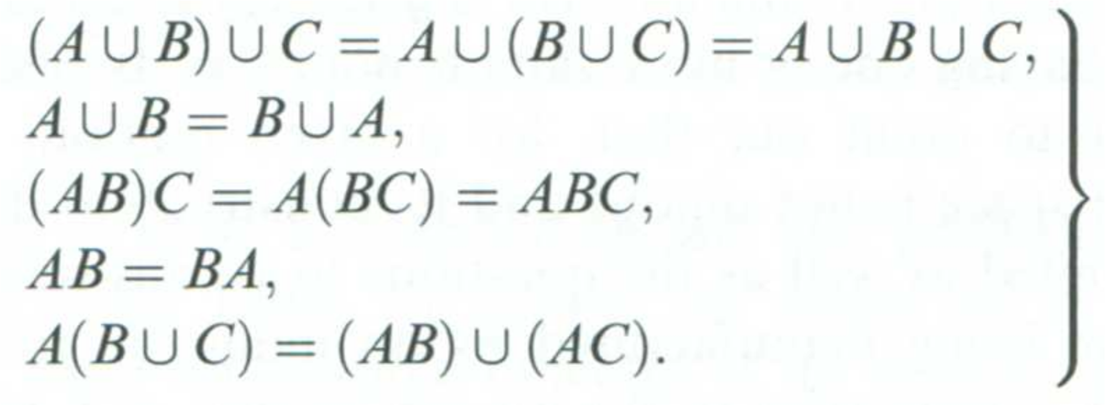




Ω: the universal space

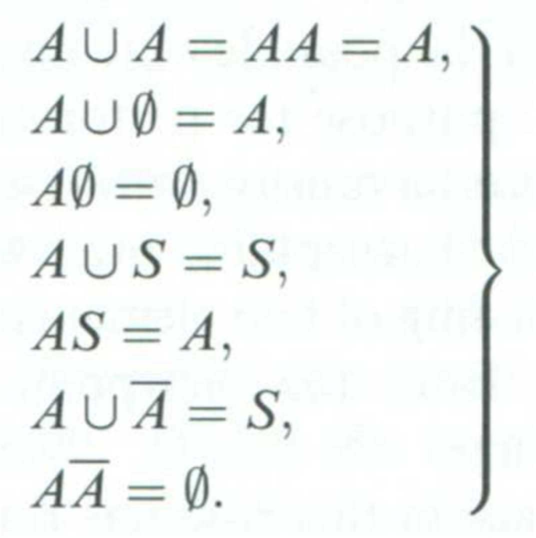
Now reconsider: $A \cup B = A \cup (\overline{A}B) = A + (\overline{A}B)$

- Venn diagrams make it easy to verify that union and intersection operations are associative, commutative and distributive:



$$\left.\begin{aligned}
&(A \cup B) \cup C = A \cup (B \cup C) = A \cup B \cup C,\\
&A \cup B = B \cup A,\\
&(AB)C = A(BC) = ABC,\\
&AB = BA,\\
&A(B \cup C) = (AB) \cup (AC).
\end{aligned}\right\}$$

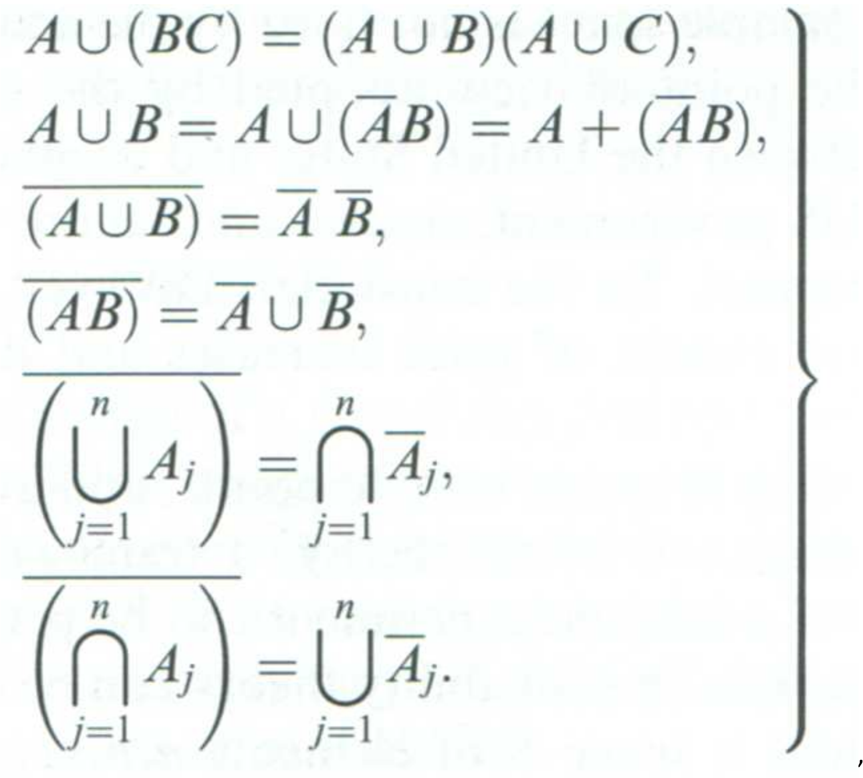
- Also easily verified:



$$\left.\begin{aligned}
&A \cup A = AA = A,\\
&A \cup \emptyset = A,\\
&A\emptyset = \emptyset,\\
&A \cup S = S,\\
&AS = A,\\
&A \cup \overline{A} = S,\\
&A\overline{A} = \emptyset.
\end{aligned}\right\}$$

- The second relation below gives the union of two sets in terms of the union of two disjoint sets. This representation will turn out to be very useful in probability calculations.

- The last two relations below are referred to as “***DeMorgan’s Laws***



$$\left.\begin{aligned}
&A \cup (BC) = (A \cup B)(A \cup C),\\
&A \cup B = A \cup (\overline{A}B) = A + (\overline{A}B),\\
&\overline{(A \cup B)} = \overline{A}\ \overline{B},\\
&\overline{(AB)} = \overline{A} \cup \overline{B},\\
&\overline{\left(\bigcup_{j=1}^{n} A_j\right)} = \bigcap_{j=1}^{n} \overline{A}_j,\\
&\overline{\left(\bigcap_{j=1}^{n} A_j\right)} = \bigcup_{j=1}^{n} \overline{A}_j.
\end{aligned}\right\}$$

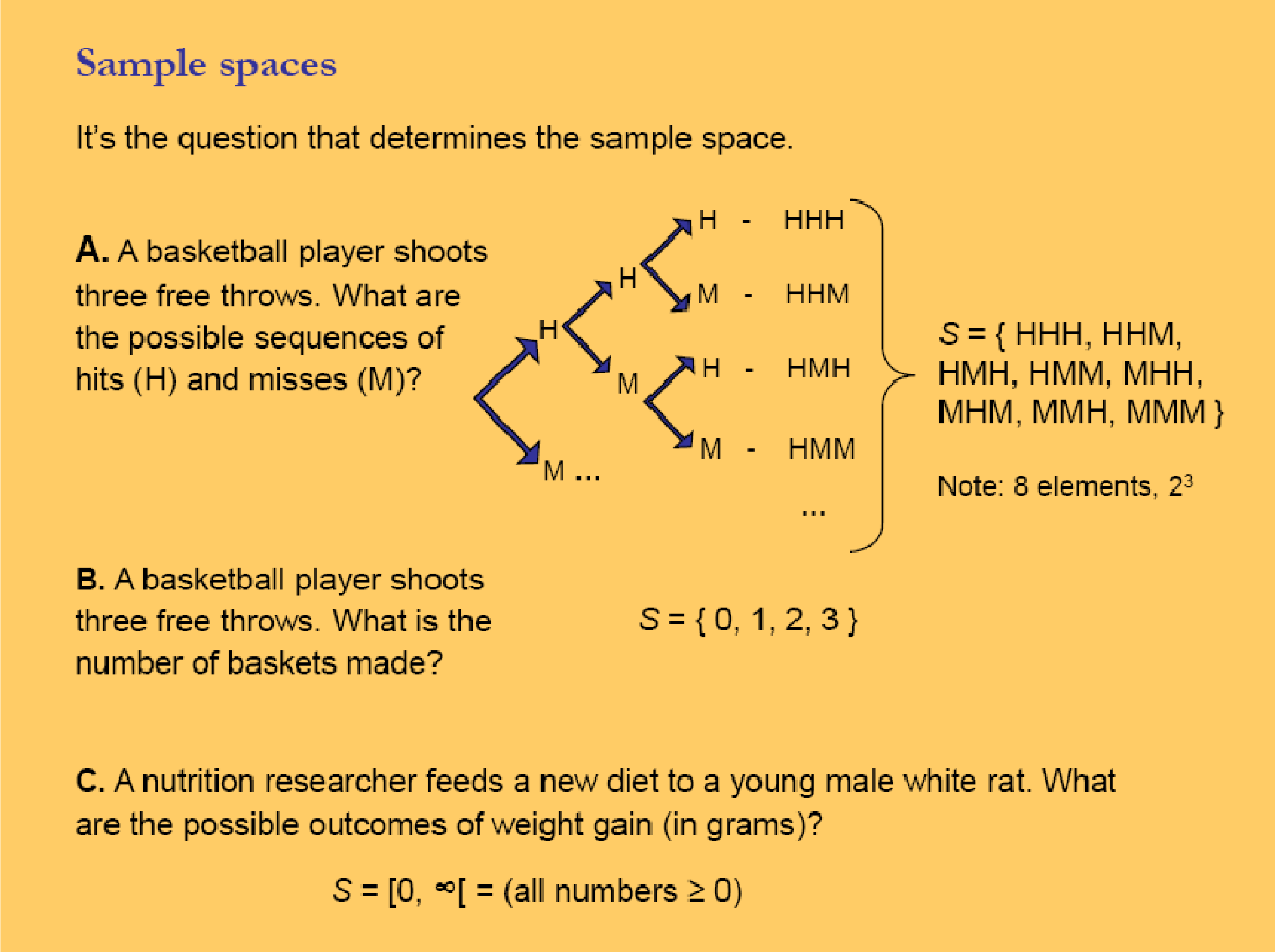
”

3.3 Sample space and probability measures

- In probability theory we are concerned with an experiment with an outcome depending on chance: a *random experiment*
- All possible distinct outcomes of a random experiment are assumed to be known and are elements of a fundamental set known as the *sample space* Each possible outcome is called *a sample point.*

  As before*, an event* ***is*** a possible outcome or set of possible outcomes of an experiment or observation.
- These descriptions nicely fit into the framework of set theory. Therefore all relations between outcomes or events in probability theory can be described by sets and operations.

- Note: For a given random experiment, the associated samples space is NOT unique!



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)?

H
H
H - HHH
M - HHM
M
H - HMH
M - HMM
...
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)?

$S = [0, \infty[$ = (all numbers ≥ 0)

- Note: Working with a wrong sample space can lead to strange results…

“Hi, I'm an amateur so I'm sorry if this is something well known and uninteresting. Is 1 + 1 = 1 in probability theory?:

Consider tossing a coin and throwing a dice. Let the set of all possible outcomes for the coin be C. which implies p(C) = 1. Let the set of all possible outcomes for the dice be D, which implies p(D) = 1. Now p(C∪D) which is the probability that either the events D or C occur is also 1.

Here's the interesting bit: C and D are disjoint sets and therefore p(C∪D)= p(C) + P(D) which implies 1 = 1 + 1.

But then I started having doubts because I made some unproved assumptions such as p(C) and p(D) and so on, are actually defined in such a situation as this and whether C and D are truly disjoint.

Help! “

The answer lies in a proper delineation of the sample space for this problem. If you throw either dice or a coin but you do not know (or do not specify) which, then the sample space is

{H, T, 1, 2, 3, 4, 5, 6}

so that P(C) = 1 and P(D) = 1 are both false.

If you throw both a dice and a coin then the sample space is

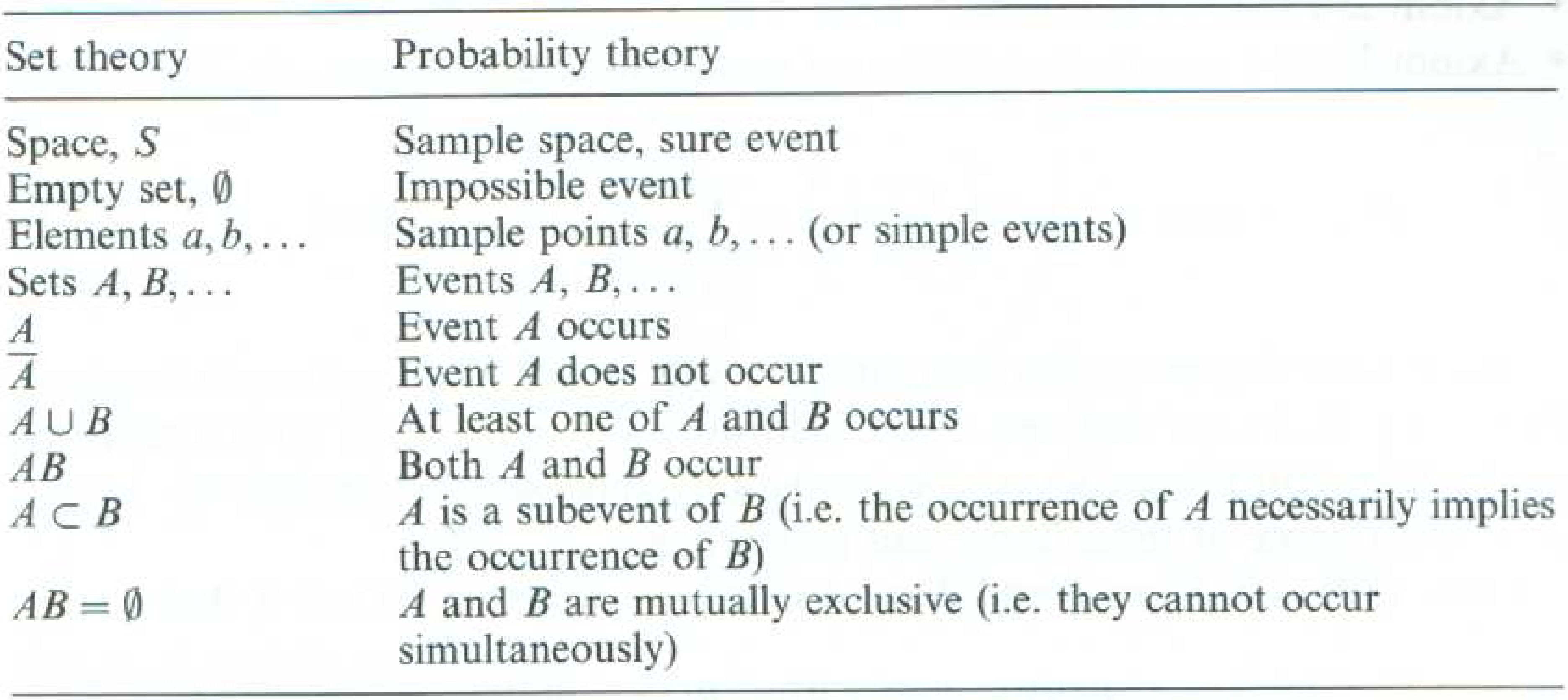
{H, T} × {1, 2, 3, 4, 5, 6}

in which case the events C and D are simply not defined.

If you just throw a coin then certainly P(C) = 1. If, in another experiment, you throw a dice then, too, P(D) = 1. But in this case the event CUD is undefined because the events C and D do not belong to the same space.

(http://www.cut-the-knot.org/)

Corresponding statements in set theory and probability



Set theory	Probability theory
Space, S	Sample space, sure event
Empty set, $\emptyset$	Impossible event
Elements $a, b, \ldots$	Sample points a, $b, \ldots$ (or simple events)
Sets $A, B, \ldots$	Events A, $B, \ldots$
A	Event A occurs
$\overline{A}$	Event A does not occur
$A \cup B$	At least one of A and B occurs
AB	Both A and B occur
$A \subset B$	A is a subevent of B (i.e. the occurrence of A necessarily implies the occurrence of B)
$AB = \emptyset$	A and B are mutually exclusive (i.e. they cannot occur simultaneously)

The notion of probability revisited

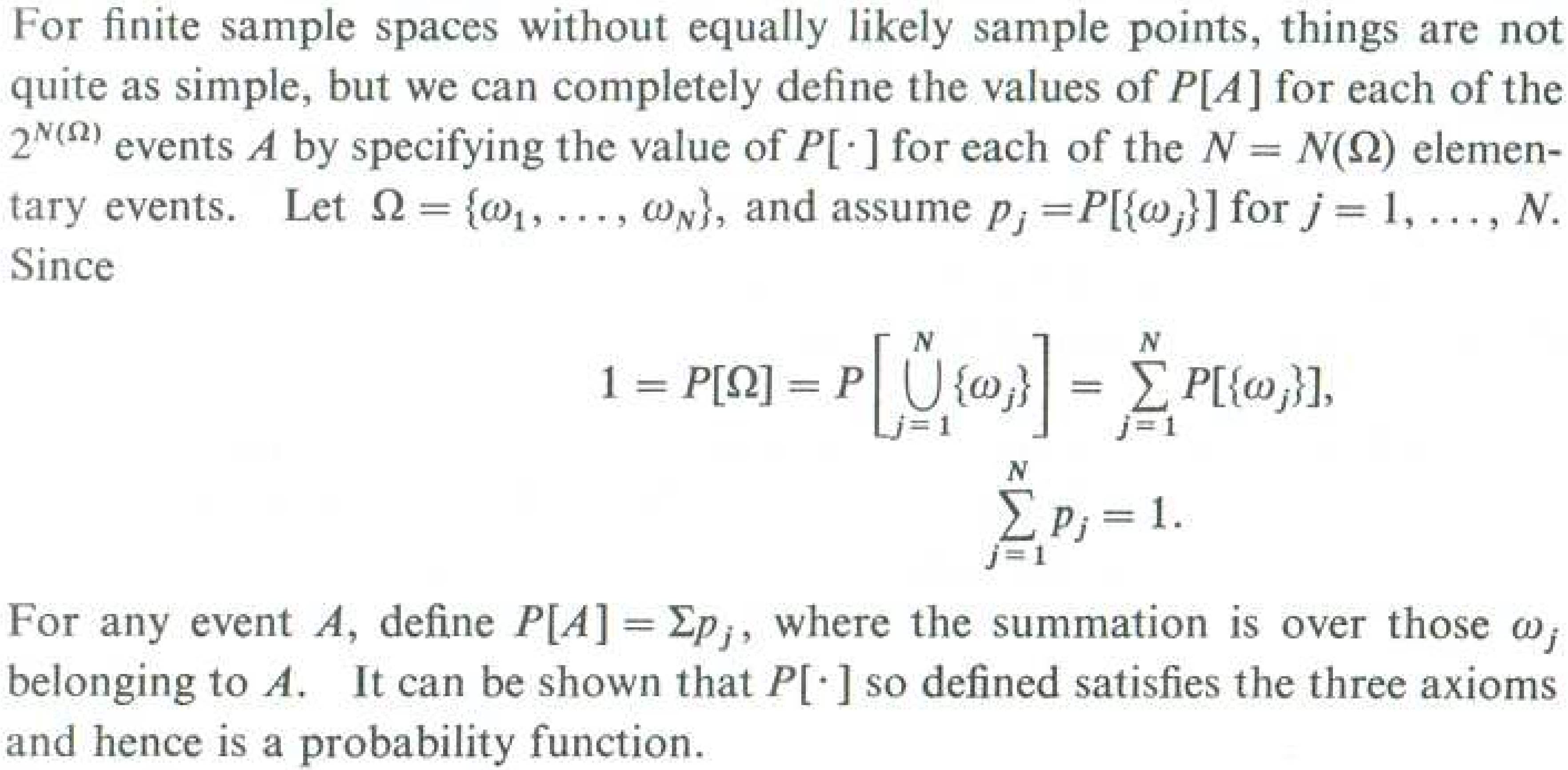
- Given a random experiment, a finite number P(A) is assigned to every event A in the sample space S of all possible events.
- The number P(A) is a function of set A and is assumed to be defined for all sets in S. It is thus a set function
- P(A) is called the *probability measure of A* or simply the *probability of A***.**
- It adheres to the ***following axioms***:
  - o Axiom 1: $P(A) \geq 0$ (nonnegative)
  - o Axiom 2: $P(S) = 1$ (normed)
  - o Axiom 3: For a countable number of mutually exclusive events $A_1, A_2, ...$ in S,

$$P(A_1 \cup A_2 \cup ...) = P(\sum_j A_j) = \sum_j P(A_j) \quad \text{(additive)}$$

A probability measure for finite samples spaces with equally likely points

Classical probability: If a random experiment can result in n mutually exclusive and equally likely outcomes and if n_A of these outcomes have an attribute A, then the probability of A is the fraction $n_A/ n.$

A probability measure for finite samples spaces without equally likely points

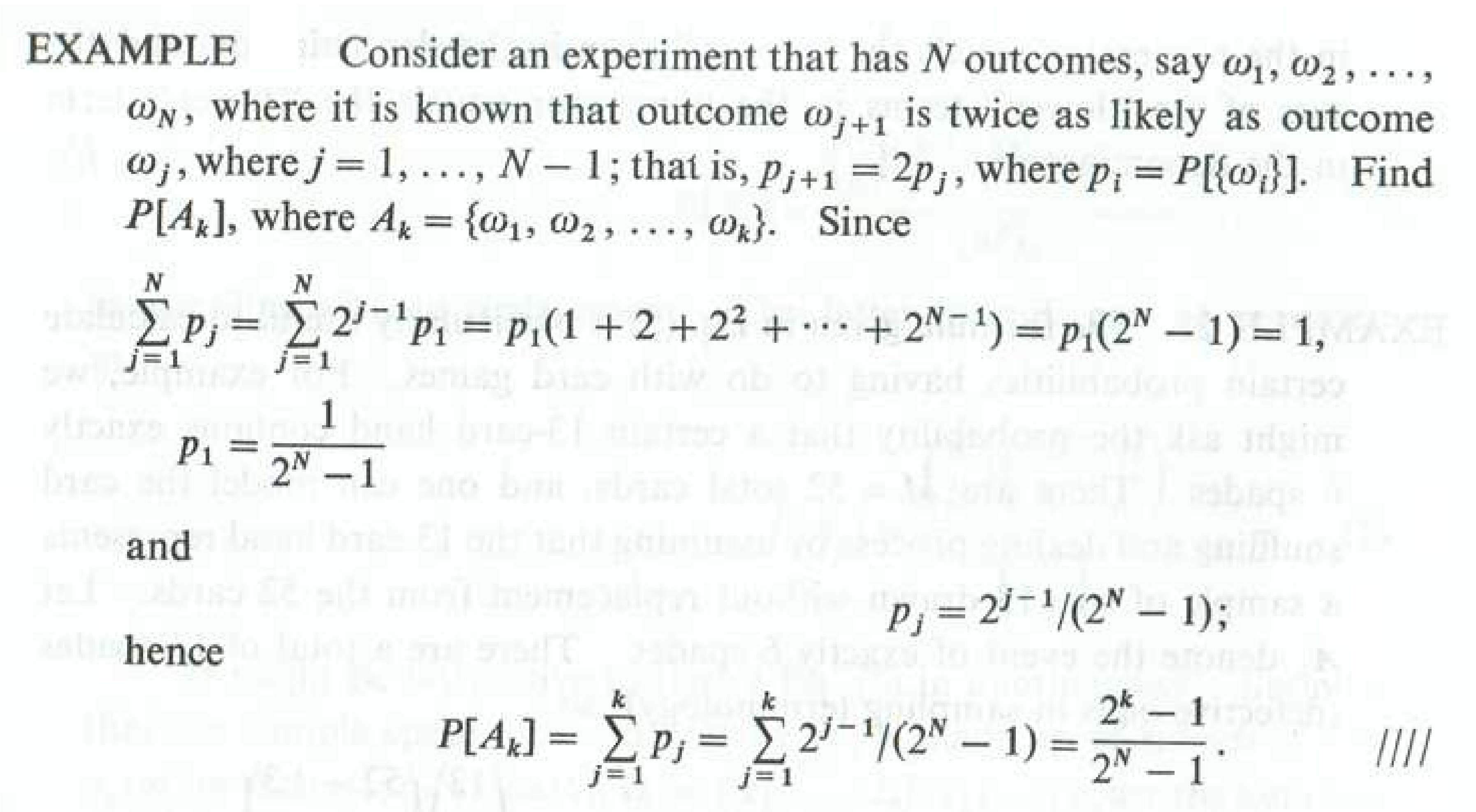


For finite sample spaces without equally likely sample points, things are not quite as simple, but we can completely define the values of $P[A]$ for each of the $2^{N(\Omega)}$ events A by specifying the value of $P[\cdot]$ for each of the $N = N(\Omega)$ elementary events. Let $\Omega = \{\omega_1, \ldots, \omega_N\}$, and assume $p_j = P[\{\omega_j\}]$ for $j = 1, \ldots, N$. Since

$$1 = P[\Omega] = P\left[\bigcup_{j=1}^{N} \{\omega_j\}\right] = \sum_{j=1}^{N} P[\{\omega_j\}],$$

$$\sum_{j=1}^{N} p_j = 1.$$

For any event A, define $P[A] = \Sigma p_j$, where the summation is over those ω_j belonging to A. It can be shown that $P[\cdot]$ so defined satisfies the three axioms and hence is a probability function.



EXAMPLE Consider an experiment that has N outcomes, say $\omega_1, \omega_2, \ldots, \omega_N$, where it is known that outcome ω_{j+1} is twice as likely as outcome ω_j, where $j = 1, \ldots, N-1$; that is, $p_{j+1} = 2p_j$, where $p_i = P[\{\omega_i\}]$. Find $P[A_k]$, where $A_k = \{\omega_1, \omega_2, \ldots, \omega_k\}$. Since

$$\sum_{j=1}^{N} p_j = \sum_{j=1}^{N} 2^{j-1} p_1 = p_1(1 + 2 + 2^2 + \cdots + 2^{N-1}) = p_1(2^N - 1) = 1,$$

$$p_1 = \frac{1}{2^N - 1}$$

and

$$p_j = 2^{j-1}/(2^N - 1);$$

hence

$$P[A_k] = \sum_{j=1}^{k} p_j = \sum_{j=1}^{k} 2^{j-1}/(2^N - 1) = \frac{2^k - 1}{2^N - 1}. \qquad ////$$

Axiomatic definition of probability: the formal way

- For S_e an algebra of events, a probability function P(.) is a set function with domain S_e and counterdomain the interval [0,1], which satisfies the following axiom:
  - o Axiom 1: $P(A) \geq 0$ (nonnegative), for every event A
  - o Axiom 2: $P(S_e) = 1$ (normed)
  - o Axiom 3: For a countable number of mutually exclusive events $A_1, A_2, ...$ in S_e, and <u>if the union of these events is itself an event</u>,

$$P(A_1 \cup A_2 \cup ...) = P(\sum_j A_j) = \sum_j P(A_j) \quad \text{(additive)}$$

- Why do we need this more general formulation?
  - If the **sample space** is sufficiently large, not all subsets of the sample space will be events …
    - recall: event = set of sample points, hence subset of sample space
    - recall: event space = class of all events associated with a given experiment
    - the **class of all events** can always be selected to be large enough so as to include all those subsets (events) whose probability we may want to talk about

- The triplet $\boxed{(\Omega,}\ \boxed{S_e,}\ P(.))$ is called a ***probability space***

- We are interested in events, mainly because we are interested in the probability an event or multiple events occur
  - o So we are interested in an event space that includes the sure event (i.e., sample space): $\Omega \in S_e$
  - o When we talk about the probability that an event occurs, we also want to talk about the probability that an event does not occur: If $A \in S_e$, then $\overline{A} \in S_e$
  - o Similarly, if A_1 and A_2 are events, then we also should $A_1 \cup A_2$ be an event: If $A_1 \in S_e$ and $A_2 \in S_e$, then $A_1 \cup A_2 \in S_e$
- Any collection of events with the aforementioned 3 properties is a Boolean algebra

Interludium:

- Let X be some set, and 2^X symbolically represent its power set. Then a subset $\Sigma \subset 2^X$ is called a *σ-algebra* if it satisfies the following three properties:

  1) Σ is non-empty
  2) Σ is closed under complementation: If A is in Σ, then so is its complement
  3) Σ is closed under countable unions: If A_1, A_2, A_3, ... are in Σ, then so is A = $A_1 \cup A_2 \cup A_3 \cup \ldots$.

From these axioms, it follows that the σ-algebra is also closed under countable intersections (by applying De Morgan's laws).

For a σ-algebra, the property “if A_1 and A_2 are events, then also the union is an event” for algebra’s, is replaced by 3) above

Rules of probability using set theory

1. Complement Rule

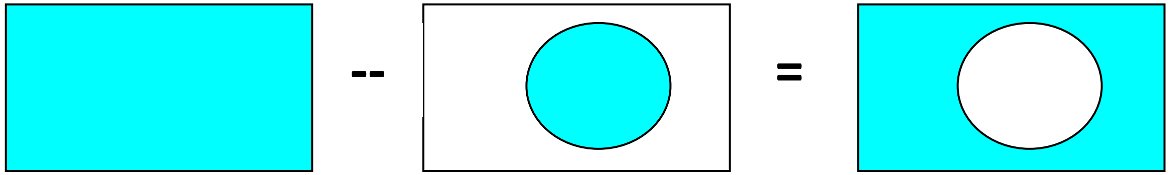
Denote “all events that are not A” as A^c. Since either A or not A must happen, $P(A) + P(A^c) = 1$. Hence

P(Event happens) = 1 - P(Event doesn't happen)

or

$$P(A) = 1 - P(A^c)$$
$$P(A^c) = 1 - P(A)$$

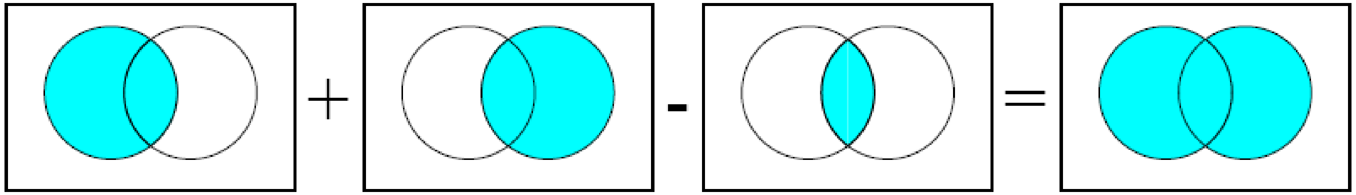
E.g. when throwing a fair die, P(not 6) = 1 – 1/6 = 5/6.



2. Addition Rule

For any two events A and B:

$$\begin{aligned} \text{P(A or B)} \quad &= \mathrm{P}(A \cup B) \\ &= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \text{ and } B) \\ &= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) \end{aligned}$$



Note: "A or B" = $A \cup B$ includes the possibility that both A and B occur.

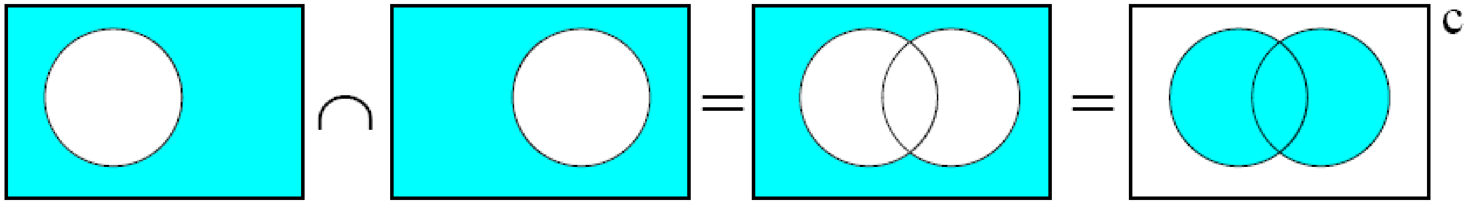
E.g. Throwing a fair die, let events be

A = get an odd number
B = get a 5 or 6

$$P(A \text{ or } B) = P(A \cup B) = P(\text{odd}) + P(5 \text{ or } 6) - P(5) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{4}{6} = \frac{2}{3}$$

This is consistent, since $P(A \cup B) = P(\{1,3,5,6\}) = \frac{4}{6} = \frac{2}{3}$

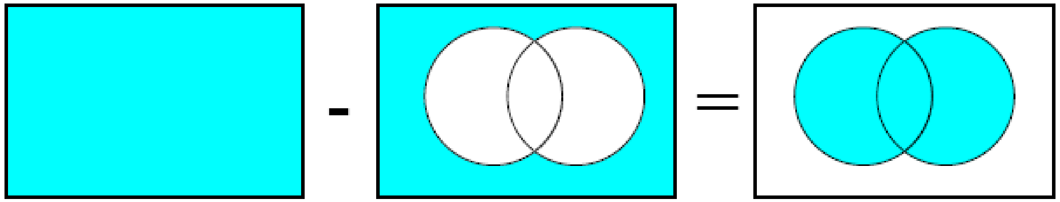
Alternative: Note that $A^c \cap B^c = (A \cup B)^c$



A^c B^c $A^c \cap B^c$

So we could also calculate $P(A \cup B)$ using

$$P(A \cup B) = 1 - P(A^c \cap B^c)$$



E.g. As before, throwing a fair die let results of interest be A = get an odd number, B = get a 5 or 6

Then A^c*={2,4,6},* B^c *= {1,2,3,4} so* $A^c \cap B^c$ *={2,4}. Hence*

$$P(A \text{ or } B) = 1 - P(A^c \cap B^c) = 1 - P(\{2,4\}) = 1 - \frac{1}{3} = \frac{2}{3}$$

This alternative form has the advantage of generalizing easily to lots of possible events:

$$P(A_1 \text{ or } A_2 \text{ or } \dots \text{ or } A_k) = 1 - P(A_1^c \cap A_2^c \cap \dots \cap A_k^c)$$

Special addition rule

- If $(A \cap B) = 0$, the events are mutually exclusive, so

$$P(A \,\mathrm{or}\, B) = P(A \cup B) = P(A) + P(B)$$

- We will often consider mutually exclusive sets of outcomes, in which case the addition rule is very simple to apply:
- In general, if several events $A_1, A_2, ..., A_k$ are mutually exclusive (i.e., at most one of them can happen in a single experiment), then

$$P(A_1 \,\mathrm{or}\, A_2 \,\mathrm{or} \ldots \mathrm{or}\, A_k) = P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + \ldots + P(A_k) = \sum_k P(A_k)$$

- E.g., throwing a fair die,

P(getting 4,5 or 6) = P(4)+P(5)+P(6)=1/6+1/6+1/6=1/2

Boole's inequality for events $A_1, A_2, ..., A_n$

$$P[A_1 \cup A_2 \cup \cdots \cup A_n] \leq P[A_1] + P[A_2] + \cdots + P[A_n].$$

PROOF $P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1A_2] \leq P[A_1] + P[A_2].$

The proof is completed using mathematical induction. ////

3. Multiplication Rule

We can re-arrange the definition of the conditional probability

$$P(A|B) = \frac{\boxed{P(A \cap B)}}{P(B)} \qquad P(B|A) = \frac{\boxed{P(A \cap B)}}{P(A)}$$

to obtain equivalent expressions for $P(A\text{and } B)$:

$$P(A \cap B) = \begin{cases} P(A|B)P(B) \\ P(B|A)P(A) \end{cases}$$

You can often think of $P(A \text{ and } B)$ as being the probability of first getting A with probability $P(A)$, and then getting B with probability $P(B|A)$. This is the same as first getting B with probability $P(B)$ and then getting A with probability $P(A|B)$.

E.g. Drawing two random cards from a pack without replacement, the probability of getting two hearts is

$$P(\text{first is a heart and second is a heart}) = P(\text{first is a heart}) \times P(\text{second is a heart} \mid \text{first is a heart})$$

$$= \frac{13}{52} \times \frac{12}{51} = \frac{1}{4} \times \frac{12}{51} = \frac{3}{51}$$

Special Multiplication Rule

If two events A and B are *independent* then P(A| B) = P(A) and P(B| A) = P(B): knowing that A has occurred does not affect the probability that B has occurred and vice versa. In that case

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

Probabilities for any number of independent events can be multiplied to get the joint probability. For example if you toss a fair coin twice, the outcome of the first throw shouldn't affect the outcome of the second throw, so the throws are independent.

E.g. A fair coin is tossed twice, the chance of getting a head and then a tail is
$P(H_1 \text{ and } T_2) = P(H_1)P(T_2) = ½ \text{ x } ½ = ¼$.

E.g. A die is thrown 3 times. The probability of getting the first six on the last throw is
P(not 6)P(not 6)P(6) = 5/6 x *5/6* x *1/6 = 25/216 = 0.116..*

3.4 A posteriori or frequency probability

Assignment of probability

- We have mentioned before that the axioms of probability define the properties of a probability measure but do not give leads on what values the probability function assigns to events: we will have to *model our random experiment* in some way in order to obtain values for the probability of events
- However, with our first definition of probability … :

Classical or "a priori" probability: If a random experiment can result in n mutually exclusive and equally likely outcomes and if n_A of these outcomes have an attribute A, then the probability of A is the fraction $n_A/ n.$

Limitations of the classical definition

- Limitation 1: The definition of probability must be modified somehow when the total number of possible outcomes is infinite

  - What is the probability that an integer drawn at random from the positive integers be even? Start with the first 2*N* integers… Your answer would be N/2N = ½
  - Can you make this argument under all circumstances?

    - Natural ordering: 1,2,3,4,5,6,… → 1/2
    - Different ordering 1,3, 2; 5,7, 4; 9,11, 6;… (first pair of odd integers, first even, etc) → 1/3
    - Oscillating sequence of integers → never attains definite value

- Limitation 2: Suppose that we toss a coin known to be biased in favor of heads (it is bent so that a head is more likely to appear than a tail).

    - What is the probability of a head?
        - The classical definition leaves us completely helpless…

- Limitation 3: Suppose notions of symmetry and equally likely do not apply?
    - What is the probability that a female will die before the age of 60?
    - What is the probability that a cookie bought at a certain bakery will have less than 3 raisins in it?
    - What is the probability that my boy (girl-) friend truly loves me?

A posteriori probabilities

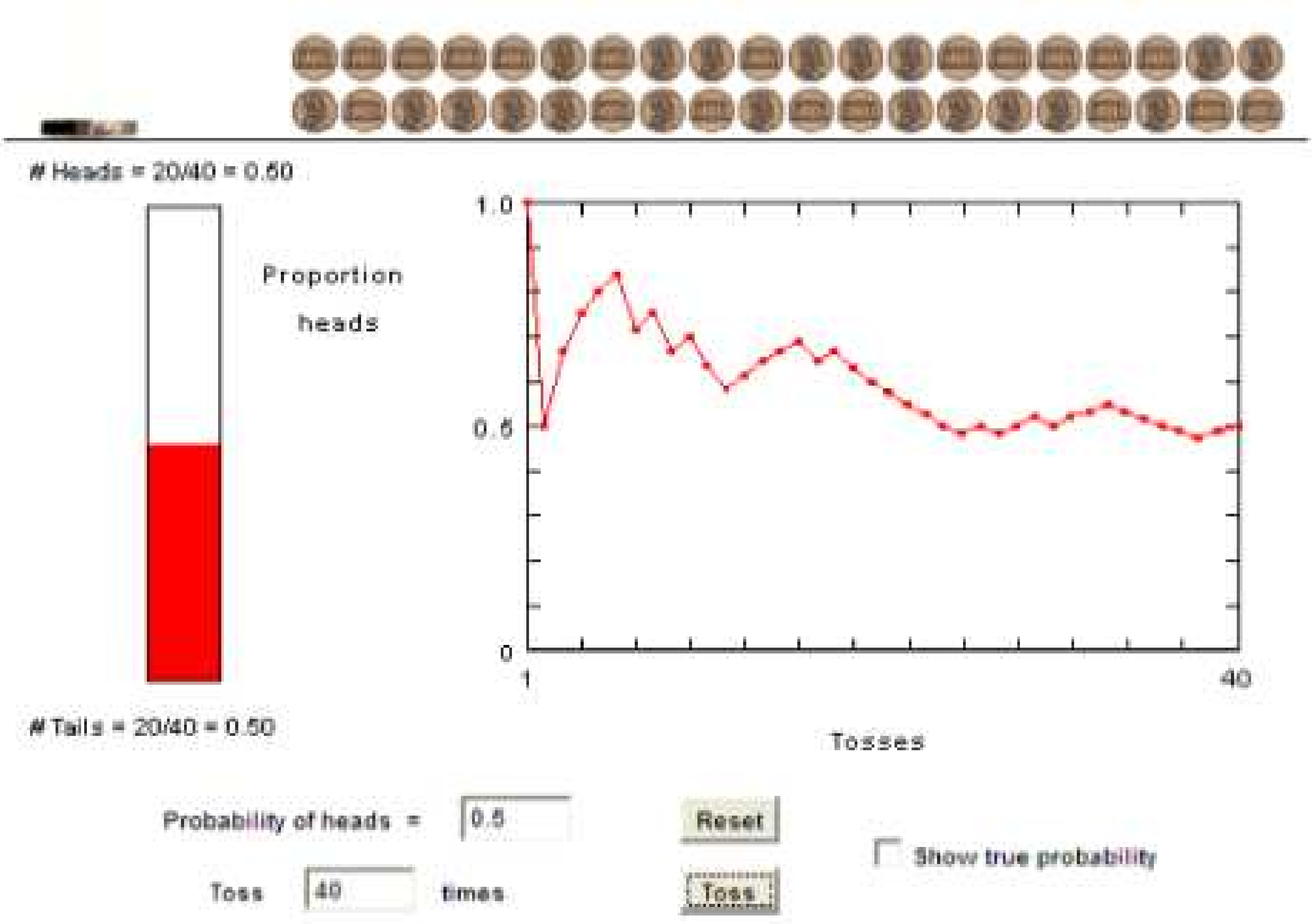
We assume that a series of observations (or experiments) can be made under quite uniform conditions:

- An observation of a random experiment is made
- Then the experiment is repeated under similar conditions, and another observation is taken
- This is repeated many times, and while conditions are similar each time, there is an uncontrollable variation which is haphazard or random so that the observations are individually unpredictable.

- In many cases the observations will fall into certain classes wherein the relative frequencies are quite stable. [Under stable or statistical regularity conditions, it is expected that this ratio will tend to a unique limit as the number of experiments becomes large.]
- This suggests that we postulate a number *p*, called the probability of the event, and approximate *p* by the *relative frequency* with which the repeated observations satisfy the event.

Frequency probability: Assuming that a random experiment is performed a large number of times, say n, then for any event A let n_A be the number of occurrences of A in the n trials and define the ratio n_A/ n as the relative frequency of A. The limiting value of the relative frequency is a probability measure of A.

Applet Probabilities



Heads = 20/40 = 0.50

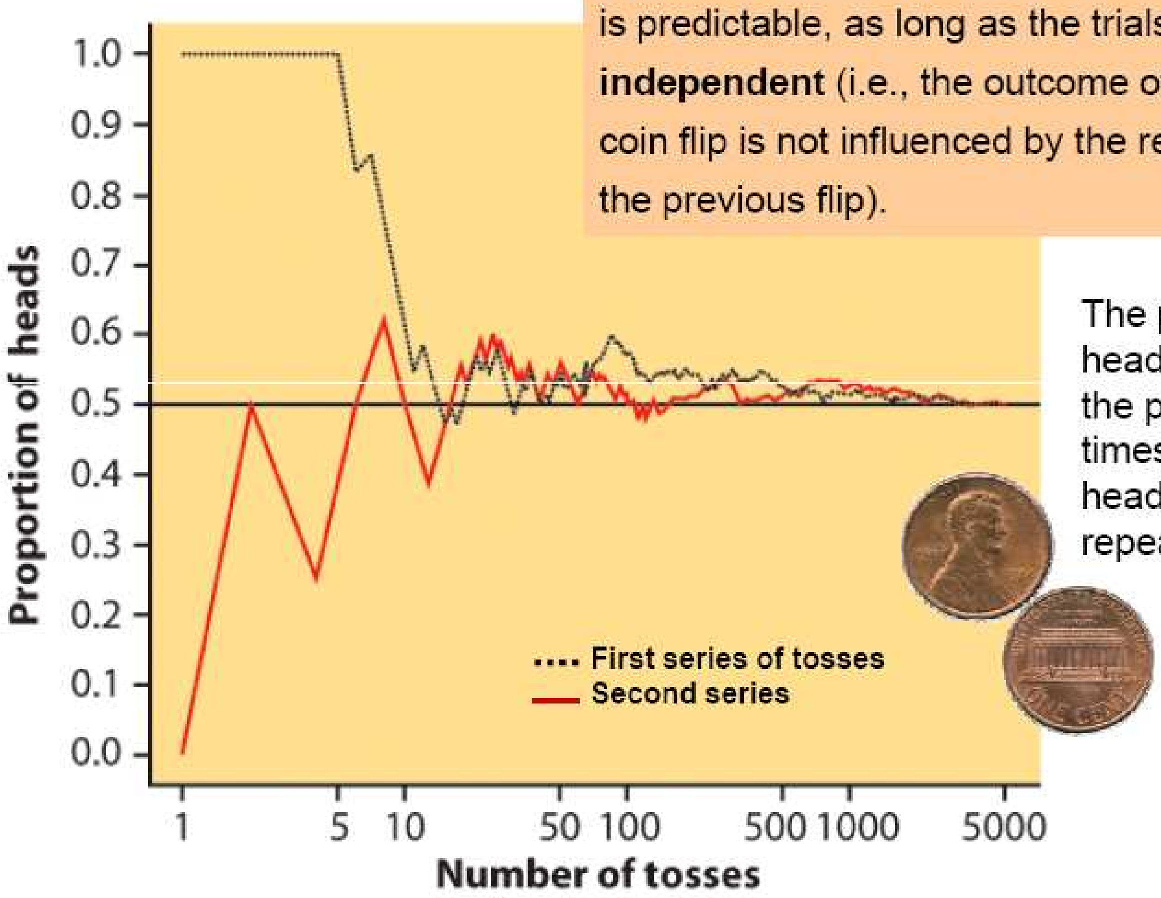
Tails = 20/40 = 0.50

Probability of heads = 0.5

Toss 40 times

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).

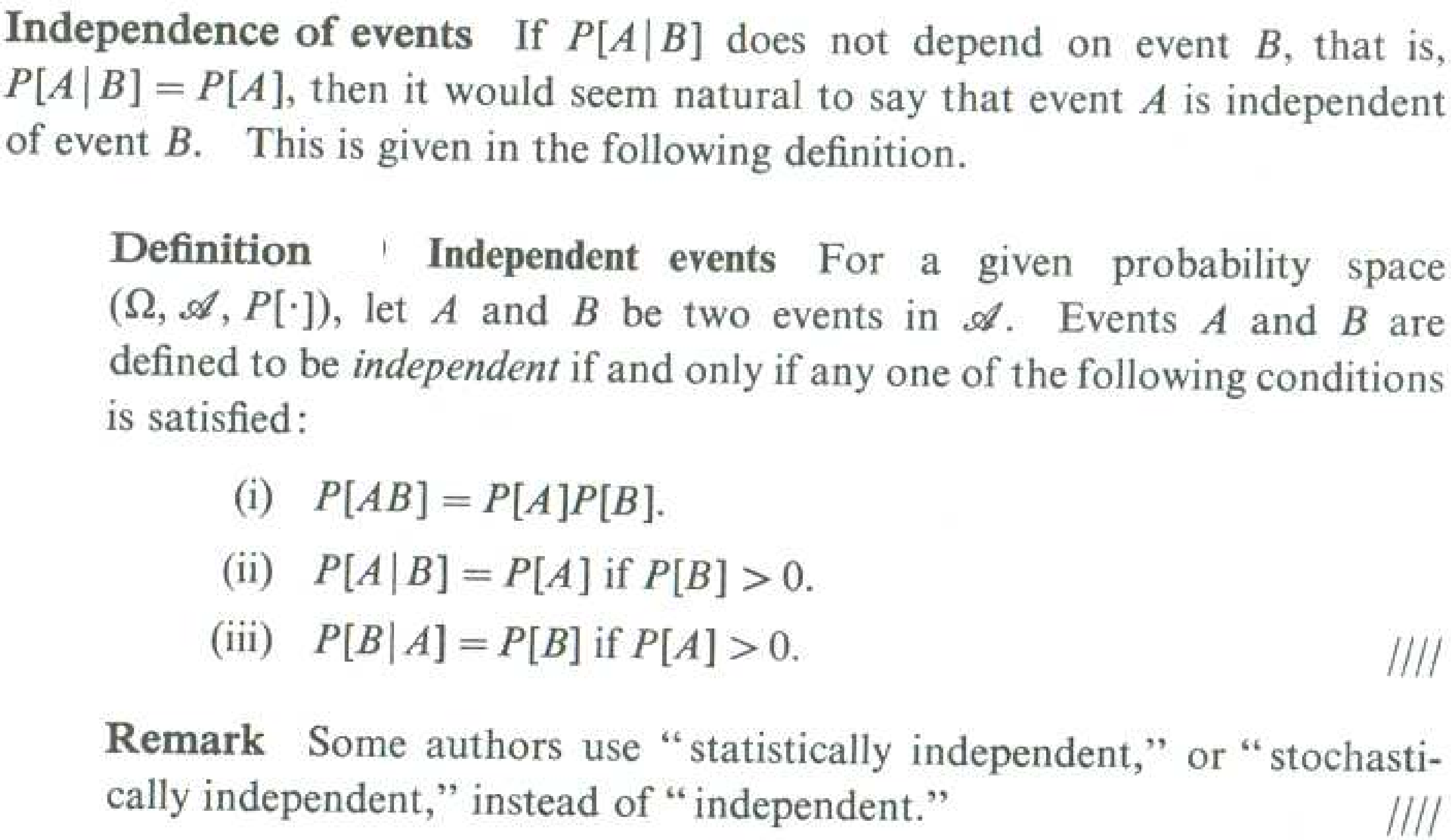


The probability of heads is 0.5 = the proportion of times you get heads in many repeated trials.

The long-run expected relative frequency of a balanced coin is 0.5

4 Statistical independence and conditional probability revisited

4.1 Independence



Independence of events If $P[A|B]$ does not depend on event B, that is, $P[A|B] = P[A]$, then it would seem natural to say that event A is independent of event B. This is given in the following definition.

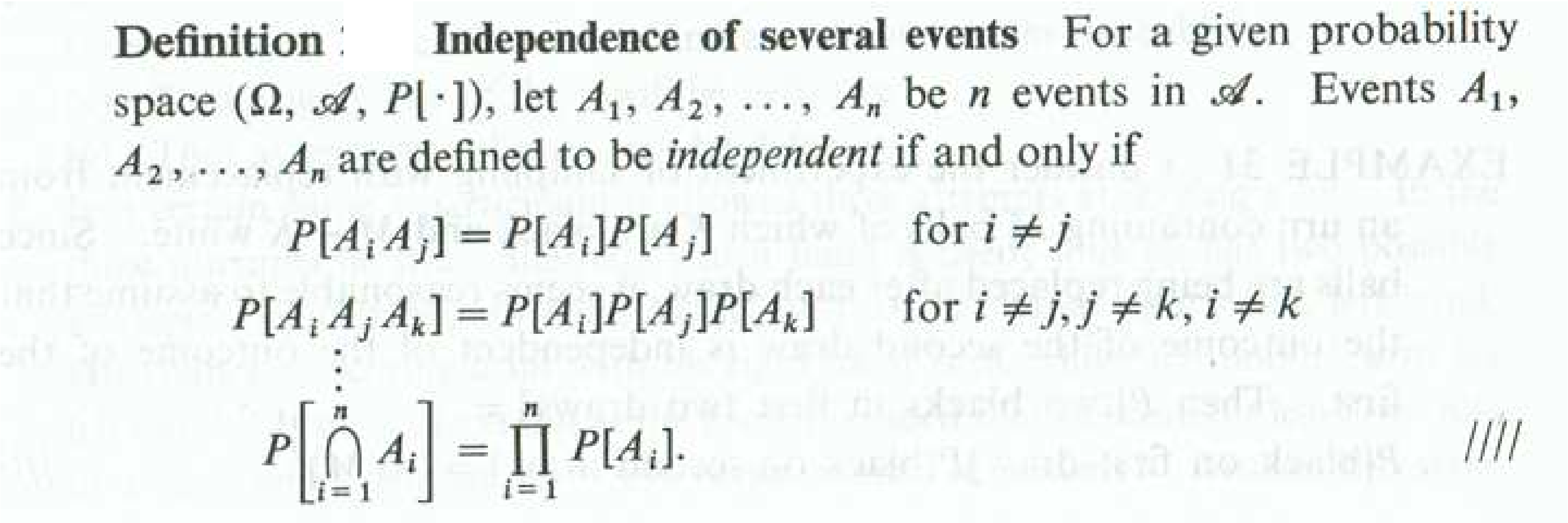
Definition **Independent events** For a given probability space $(\Omega, \mathscr{A}, P[\cdot])$, let A and B be two events in $\mathscr{A}$. Events A and B are defined to be *independent* if and only if any one of the following conditions is satisfied:

(i) $P[AB] = P[A]P[B]$.

(ii) $P[A|B] = P[A]$ if $P[B] > 0$.

(iii) $P[B|A] = P[B]$ if $P[A] > 0$. ////

Remark Some authors use “statistically independent,” or “stochastically independent,” instead of “independent.” ////



Definition : **Independence of several events** For a given probability space $(\Omega, \mathscr{A}, P[\cdot])$, let $A_1, A_2, \ldots, A_n$ be n events in $\mathscr{A}$. Events $A_1, A_2, \ldots, A_n$ are defined to be *independent* if and only if

$$P[A_i A_j] = P[A_i]P[A_j] \qquad \text{for } i \neq j$$

$$P[A_i A_j A_k] = P[A_i]P[A_j]P[A_k] \qquad \text{for } i \neq j, j \neq k, i \neq k$$

$$\vdots$$

$$P\left[\bigcap_{i=1}^{n} A_i\right] = \prod_{i=1}^{n} P[A_i]. \qquad ////$$

4.2 Conditional probability

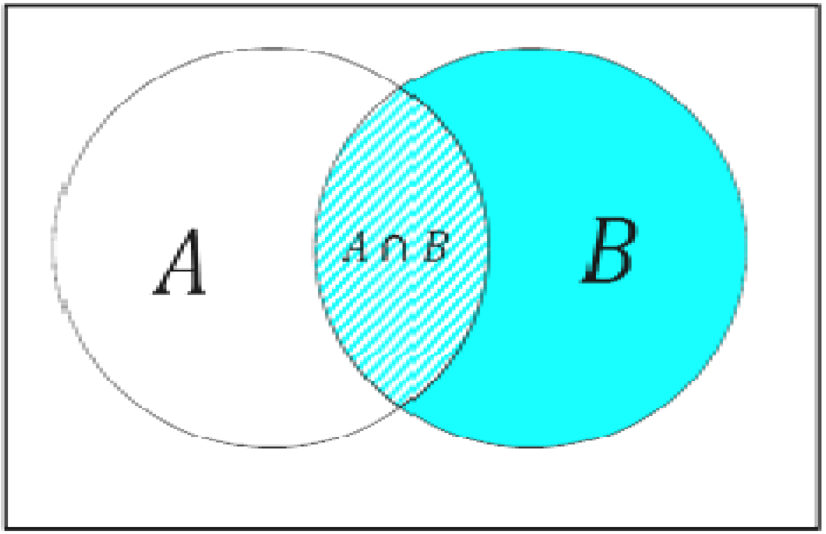
Conditional probability: P(A|B) means the probability of A given that B has happened or is true.
e.g. P(result of coin toss is heads | the coin is fair) =1/2
P(Tomorrow is Tuesday | it is Monday) = 1
P(card is a heart | it is a red suit) = 1/2

Probabilities are always conditional on something, for example prior knowledge, but often this is left implicit when it is irrelevant or assumed to be obvious from the context.

In terms of P(B) and P(A and B) we have

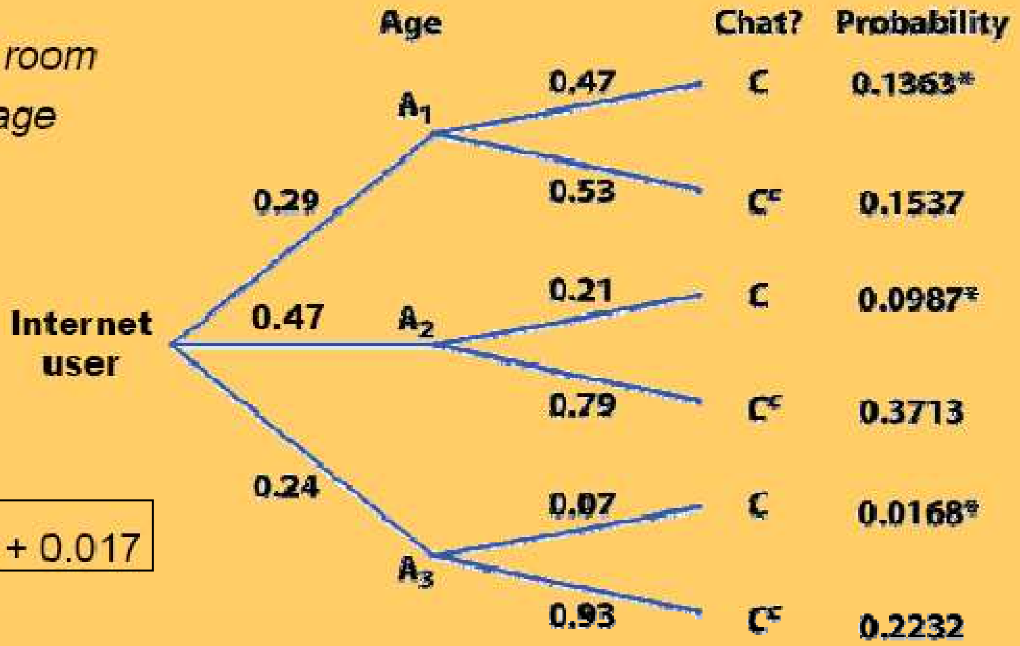
$$P(A|B) = \frac{P(A\cap B)}{P(B)}$$

$P(B)$ gives the probability of an event in the B set. Given that the event is in B, $P(A|B)$ is the probability of also being in A. It is the fraction of the B outcomes that are also in A:



Conditional probabilities can get complex, and it is often a good strategy to build a **probability tree** that represents all possible outcomes graphically and assigns conditional probabilities to subsets of events.

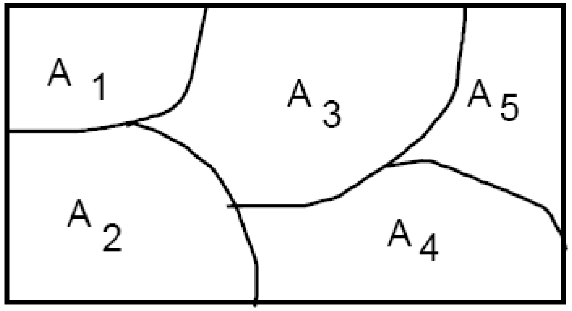
Tree diagram for chat room habits for three adult age groups



P(chatting) = 0.136 + 0.099 + 0.017
= 0.252

About 25% of all adult Internet users visit chat rooms.

The law of total probability: relating the prob of an event to cond probs



If $A_1, A_2, \ldots, A_k$ form a partition (a mutually exclusive list of all possible outcomes) and B is any event then

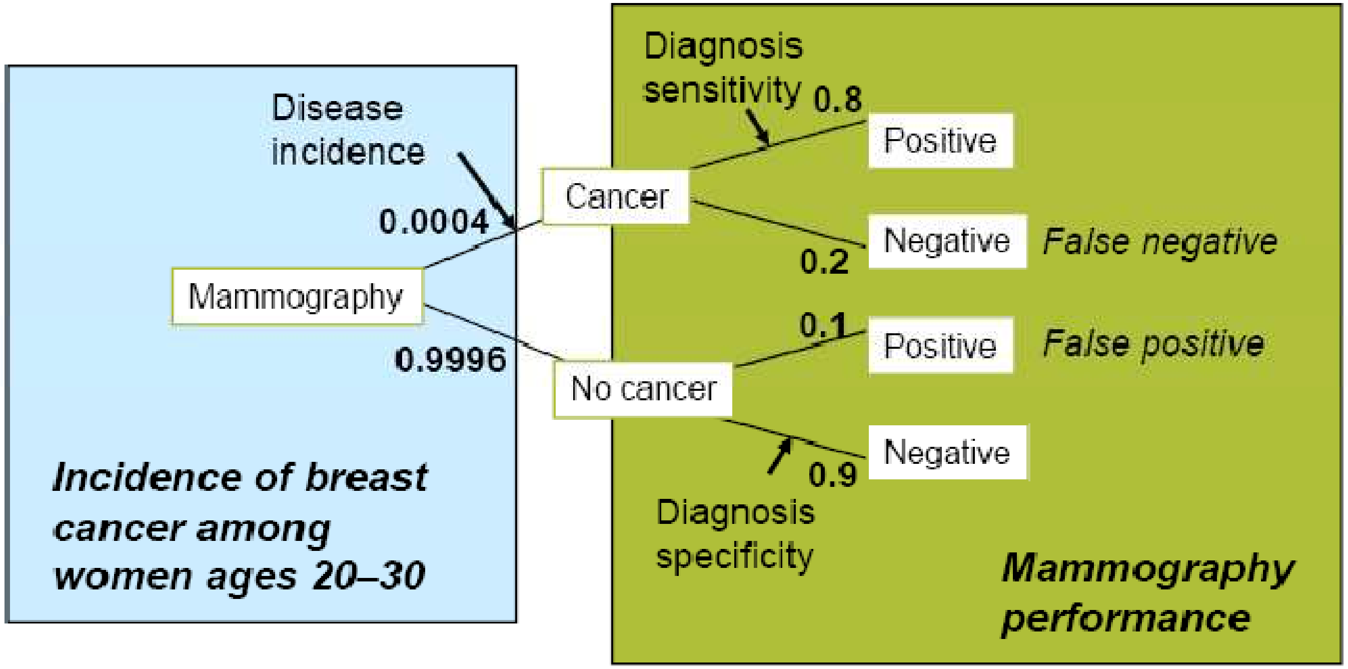
$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k) = \sum_k P(B|A_k)P(A_k)$$

Proof: This follows since

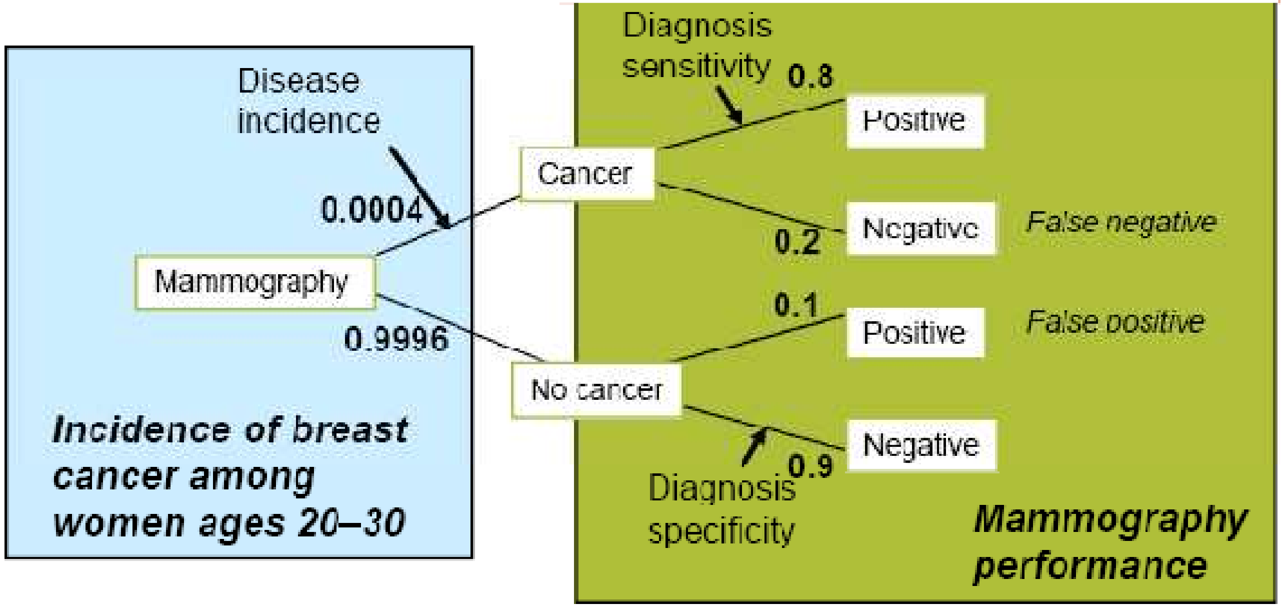
$$\begin{aligned} \mathrm{P}(B) &= \mathrm{P}(B \mid A_1)\mathrm{P}(A_1) + \mathrm{P}(B \mid A_2)\mathrm{P}(A_2) + \ldots + \mathrm{P}(B \mid A_k)\mathrm{P}(A_k) \\ &= \mathrm{P}(B \cap A_1) + \mathrm{P}(B \cap A_2) + \ldots + \mathrm{P}(B \cap A_k) \\ &= \mathrm{P}(B \cap A_1 \text{ or } B \cap A_2 \text{ or.. + or } B \cap A_k) \\ &= \mathrm{P}(B \cap (A_1 \text{ or } A_2 \text{ or } A_k)) \\ &= \mathrm{P}(B) \end{aligned}$$

Another example: breast cancer screening

If a woman in her 20s gets screened for breast cancer and receives a positive test result, what is the probability that she does have breast cancer?



She could either have a positive test and have breast cancer or have a positive test but not have cancer (false positive).



Possible outcomes given the positive diagnosis: positive test and breast cancer or positive test but no cancer (false positive).

$$P(cancer \mid pos) = \frac{P(cancer\ and\ pos)}{P(cancer\ and\ pos) + P(nocancer\ and\ pos)}$$

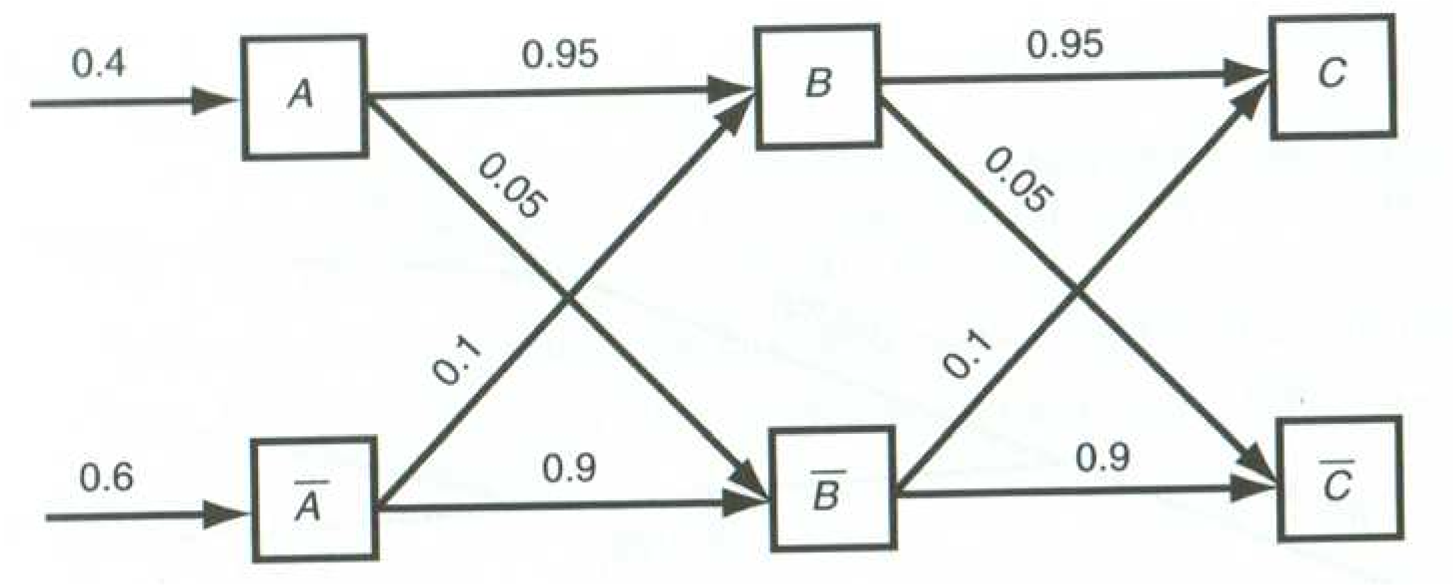
$$= \frac{0.0004 * 0.8}{0.0004 * 0.8 + 0.9996 * 0.1} \approx 0.3\%$$

This value is called the positive predictive value, or *PV+*. It is an important piece of information but, unfortunately, is rarely communicated to patients.

Example: two-stage binary channel system

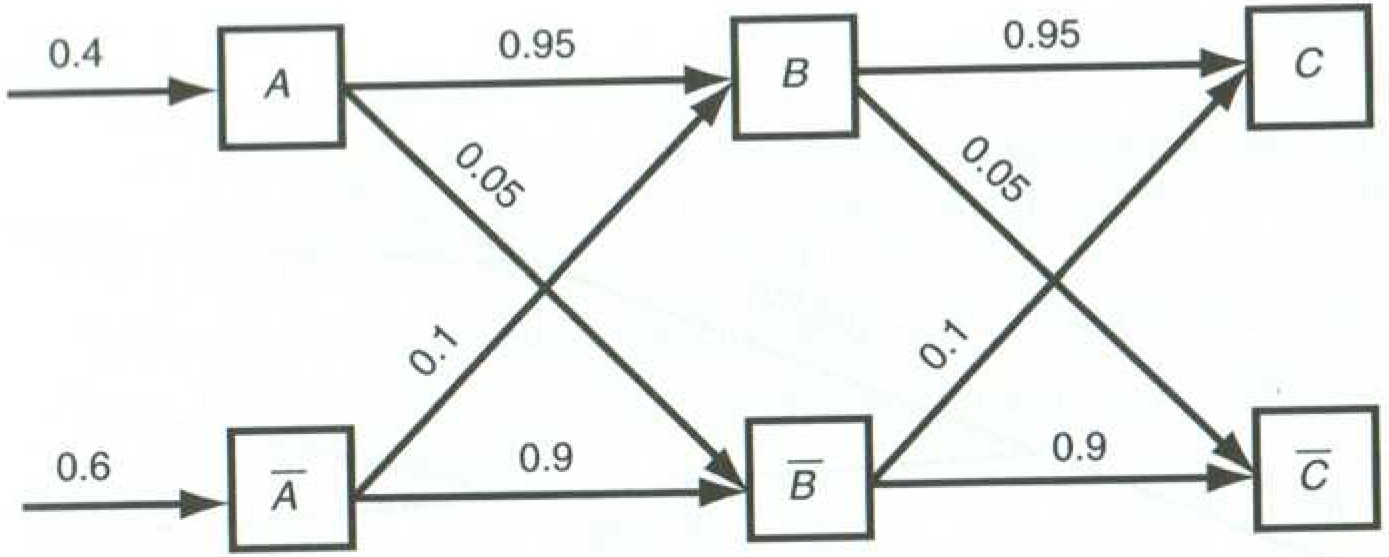
- Suppose the outcome at the second stage is dependent only on what happened at the first stage and not on outcomes at stages prior to the first:

$$P(C|BA) = P(C|B), P(\overline{C}|\overline{B}A) = P(\overline{C}|\overline{B}), ...$$

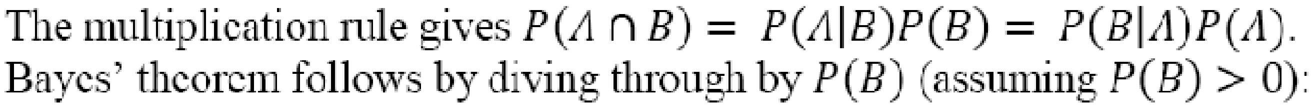


$$\boxed{P(ABC) = P(A)P(B|A)P(C|BA)}$$
$$= P(A)P(B|A)P(C|B)$$
$$= 0.4(0.95)(0.95) = 0.361.$$

$$P(C) = P(ABC) + P(A\overline{B}C) + P(\overline{A}BC) + P(\overline{A}\,\overline{B}C)$$
$$= 0.95(0.95)(0.4) + 0.1(0.05)(0.4) + 0.95(0.1)(0.6) + 0.1(0.9)(0.6)$$
$$= 0.472.$$

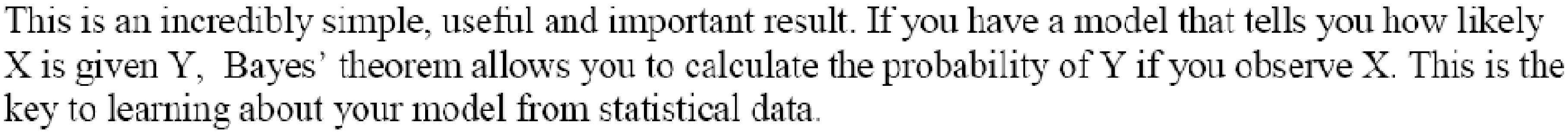


Bayes' Theorem

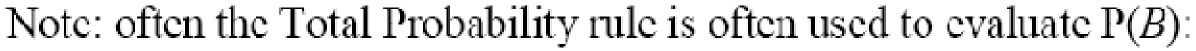


The multiplication rule gives $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$.
Bayes' theorem follows by diving through by $P(B)$ (assuming $P(B) > 0$):

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



This is an incredibly simple, useful and important result. If you have a model that tells you how likely X is given Y, Bayes' theorem allows you to calculate the probability of Y if you observe X. This is the key to learning about your model from statistical data.



Note: often the Total Probability rule is often used to evaluate P(B):

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

Principle of proportionality

- This is an immediate consequence of Bayes' Theorem.
- If various alternatives are equally likely, and then some event is observed, the updated probabilities for the alternatives are proportional to the probabilities that the observed event would have occurred under those alternatives.

The formal derivation is simple. Assume

(*) $P(A_1) = P(A_2) = ... = P(A_n) > 0$ and $P(B) > 0$.

Then $P(A_m|B) = K\ P(B|A_m)$, for all $m = 1, 2, ..., n$, where $K > 0$ does not depend on m.

Indeed, by Bayes' theorem,

$$\begin{aligned} P(A_m|B) &= P(A_m \cap B) / P(B) \\ &= P(A_m)\, P(B|A_m) / P(B) \\ &= (P(A_m) / P(B))\, P(B|A_m). \end{aligned}$$

The assertion holds, with $K = P(A_m) / P(B)$ - constant from (*) before.

The Bear cubs problem revisited

There are two bears - white and dark. Assume it is known that one of the bears is male. What is the probability that both are males?

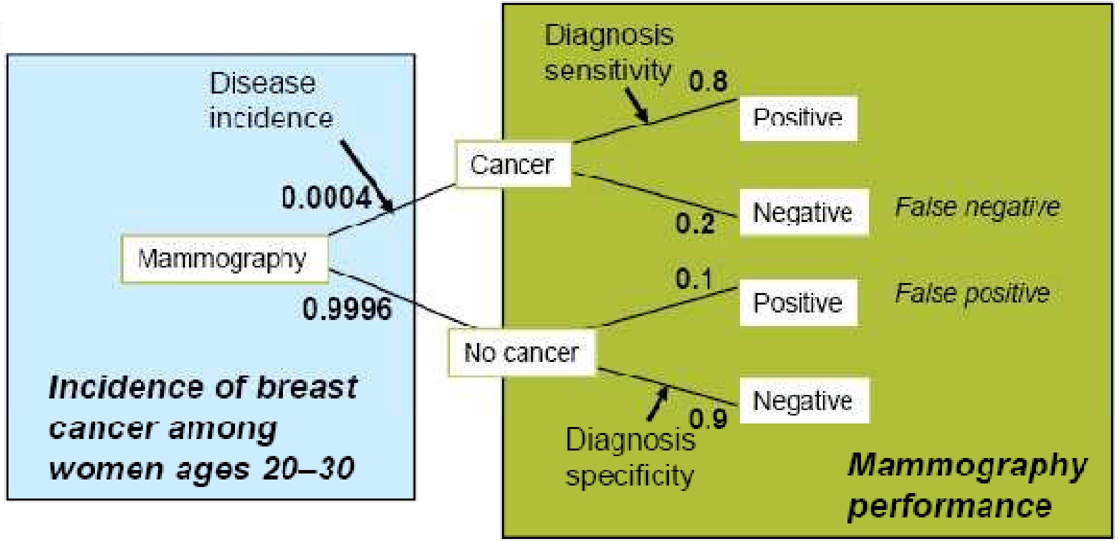
Solution: With the common assumption that sexes are evenly distributed among the bears as among the humans, at the outset, there are four equally probable variants: A_1 = (female/female), A_2 = (female/male), A_3 = (male/female), A_4 = (male/male). Event B is the acknowledgement that one of the bears is male. Conditional probabilities of B assuming one of the A's are as follows:

$$P(B|A_1) = 0,\ P(B|A_2) = 1,\ P(B|A_3) = 1,\ P(B|A_4) = 1.$$

The conditional probabilities of A's assuming B are proportional to the above but must add to 1. So they are 0, 1/3, 1/3, 1/3. Only in the last event the second bear happens to be male, thus the probability of the latter happening is 1/3. (http://www.cut-the-knot.org/)

Breast cancer screening example: application of Bayes' theorem

If a woman in her 20s gets screened for breast cancer and receives a positive test result, what is the probability that she does have breast cancer?



Disease incidence

0.0004

Mammography

0.9996

Incidence of breast cancer among women ages 20–30

Diagnosis sensitivity

Cancer

0.8

Positive

0.2

Negative

False negative

No cancer

0.1

Positive

False positive

0.9

Negative

Diagnosis specificity

Mammography performance

This time, we use Bayes's rule: $P(A_i \mid C) = \dfrac{P(C \mid A_i)P(A_i)}{P(C \mid A_1)P(A_1) + P(C \mid A_2)P(A_2) + \cdots + P(A_k)P(C \mid A_k)}$

A1 is cancer, A2 is no cancer, *C* is a positive test result.

$$P(cancer \mid pos) = \frac{P(pos \mid cancer)P(cancer)}{P(pos \mid cancer)P(cancer) + P(pos \mid nocancer)P(nocancer)}$$

$$= \frac{0.8 * 0.0004}{0.8 * 0.0004 + 0.1 * 0.9996} \approx 0.3\%$$

Bayesian odds

- On occasion when there are two events, say *A* and *B*, whose comparative **posterior probabilities** are of interest, it may be more advantageous to consider the ratios, i.e.:

$$\frac{\boxed{p(A|C)}}{p(B|C)} = \frac{p(C|A)}{p(C|B)} \cdot \frac{p(A)}{p(B)}.$$

- Ward Edwards gives a simple example where the latter formula comes in handy:
  There are two bags, one containing 700 red and 300 blue chips, the other containing 300 red and 700 blue chips. Flip a fair coin to determine which one of the bags to use. Chips are drawn with replacement. In 12 samples, 8 red and 4 blue chips showed up. What is the probability that it was the predominantly red bag?

.

Solution:

Author Edwards writes

Clearly the sought probability is higher than 0.5.

Is it?

Let *A* be the event of selecting the first bag. Let *B* be the event of selecting the second bag. Finally, let C be the result of the experiment, i.e., drawing 8 red and 4 blue chips from the selected bag. Clearly,

$$p(C|A) = (\frac{7}{10})^8(\frac{3}{10})^4$$
$$p(C|B) = (\frac{7}{10})^4(\frac{3}{10})^8$$

so that $\frac{p(C|A)}{p(C|B)} = \left(\frac{7}{3}\right)^4 \approx 29.642$.

Now, *p*(*A*)=*p*(*B*)=0.5, implying that

$$\frac{p(A|C)}{p(B|C)} = \frac{p(C|A)}{p(C|B)} \times 1 = 29.642$$

From *p*(*A*|*C*)+*p*(*B*|*C*)=1, it then follows that

$$\frac{p(A|C)}{1 - p(A|C)} = 29.642 \text{ [this is an odds!!!]}$$

and

$$p(A|C) \approx \frac{29.642}{1 + 29.642} = \frac{29.642}{30.642} \approx 0.967$$

(http://www.cut-the-knot.org/)

Odds

- Note that by our assumption of equal probabilities for the events A and B,

$$\frac{p(A|C)}{p(B|C)} = \frac{p(A|C)}{1 - p(A|C)}$$

  and is therefore a genuine *odds*.

- The experts on this issue live just south of here in a town called Peculiar, Missouri. The sign just outside city limits reads "Welcome to Peculiar, where the odds are with you." ☺ ☺ ☺

- Odds are just an alternative way of expressing the likelihood of an event such as catching the flu. Probability is the expected number of flu patients divided by the total number of patients. Odds would be the expected number of flu patients divided by the expected number of non-flu patients.

- During the flu season, you might see ten patients in a day. One would have the flu and the other nine would have something else.
    - So the probability of the flu in your patient pool would be one out of ten.
    - The odds would be one to nine.
- It's easy to convert a probability into an odds. Simply take the probability and divide it by one minus the probability:

odds = probability / (1-probability)

- If you know the odds in favor of an event, the probability is just the odds divided by one plus the odds.

probability = odds / (1+odds)

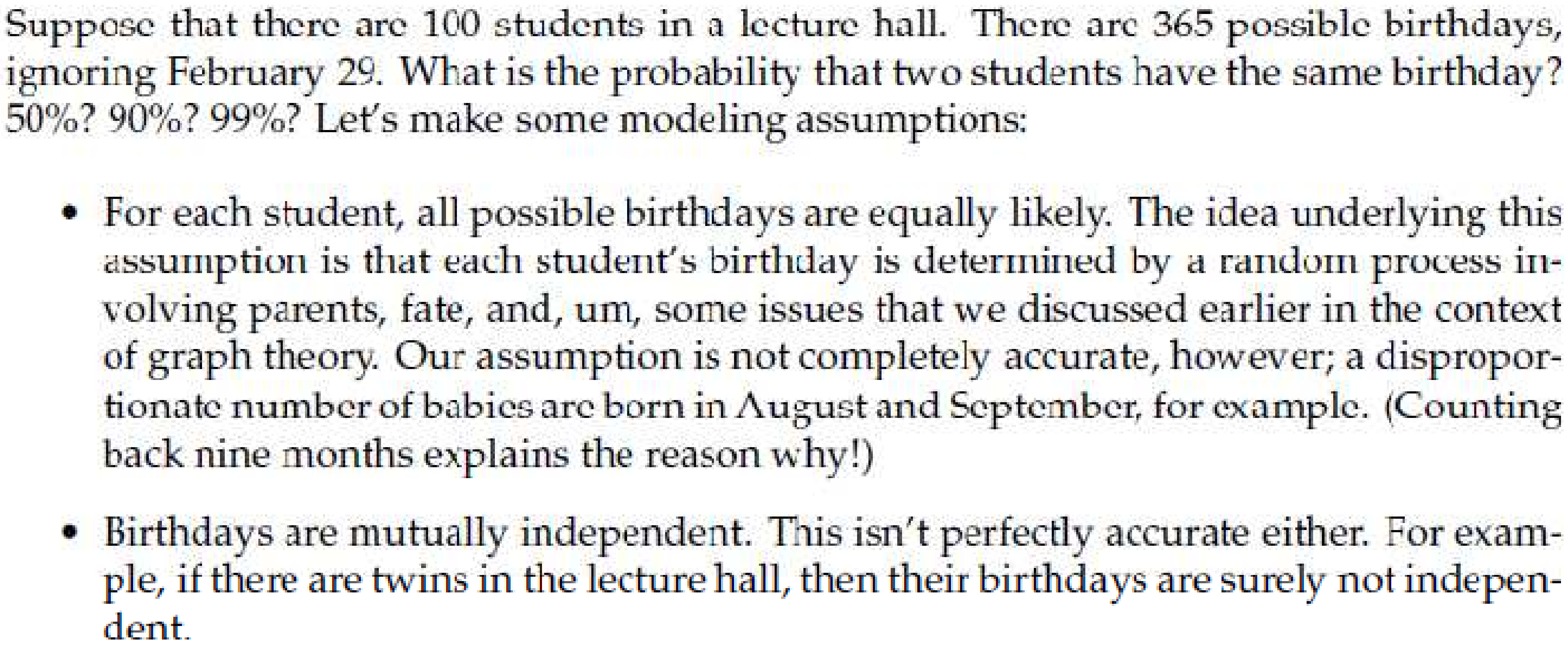
- You should get comfortable with converting probabilities to odds and vice versa. Both are useful depending on the situation.

5 In conclusion

5.1 Take-home messages

- We have introduced an axiomatic definition of probability and have offered a guideline on how to associate probabilities to an event.
- We have derived several useful properties to compute the probability of a set of events
- We have encountered two main widely application interpretations of a probability:
    - as the idealized value of a relative frequency from many independent repetitions of the same thing (frequentist)
    - as a measure of the belief that an event will occur (Bayesian)
- Whereas the first involves a so-called frequentist view, the second involves a so-called Bayesian view and is the subject of a more advanced course in statistics.

5.2 The birthday paradox



Suppose that there are 100 students in a lecture hall. There are 365 possible birthdays, ignoring February 29. What is the probability that two students have the same birthday? 50%? 90%? 99%? Let's make some modeling assumptions:

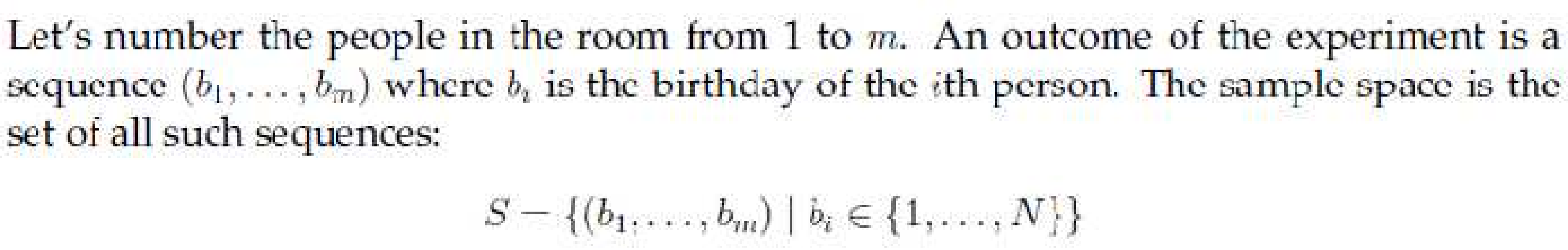
- For each student, all possible birthdays are equally likely. The idea underlying this assumption is that each student's birthday is determined by a random process involving parents, fate, and, um, some issues that we discussed earlier in the context of graph theory. Our assumption is not completely accurate, however; a disproportionate number of babies are born in August and September, for example. (Counting back nine months explains the reason why!)
- Birthdays are mutually independent. This isn't perfectly accurate either. For example, if there are twins in the lecture hall, then their birthdays are surely not independent.

The four-step method

Let us switch from specific numbers to variables. Let m be the number of people in the room and let N be the number of days in a year.

Step 1: *Find the sample space*

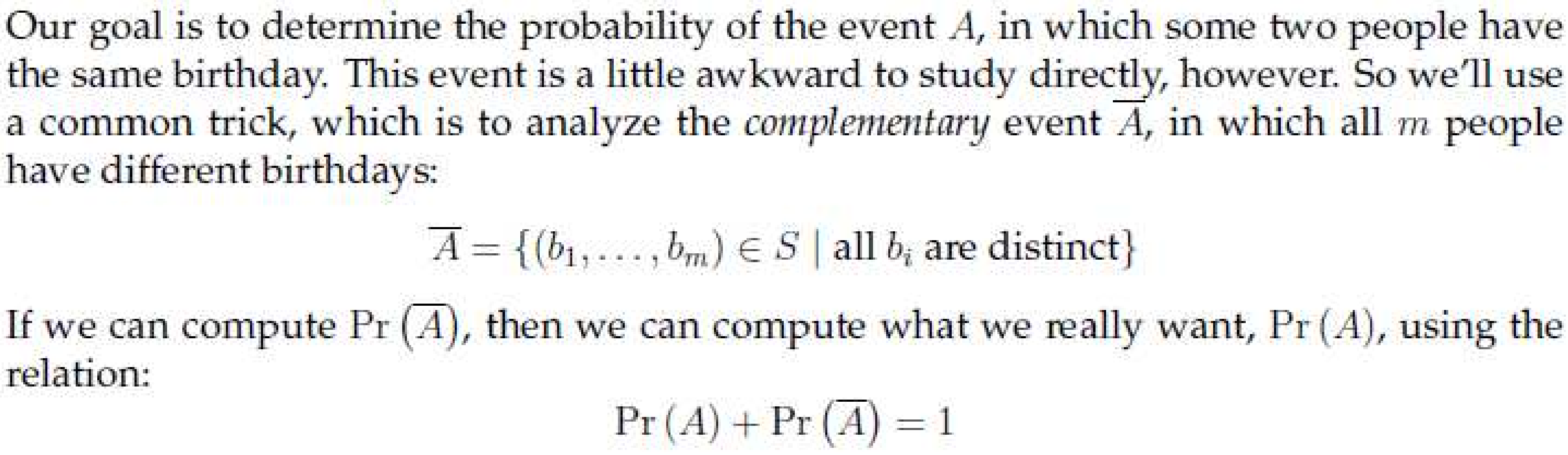
[When the sample space is not too large, it is feasible to use tree diagrams, as in the breast cancer example, to capture the sample space]



Let's number the people in the room from 1 to m. An outcome of the experiment is a sequence $(b_1, \ldots, b_m)$ where b_i is the birthday of the ith person. The sample space is the set of all such sequences:

$$S = \{(b_1, \ldots, b_m) \mid b_i \in \{1, \ldots, N\}\}$$

Step 2: *Define events of interest*



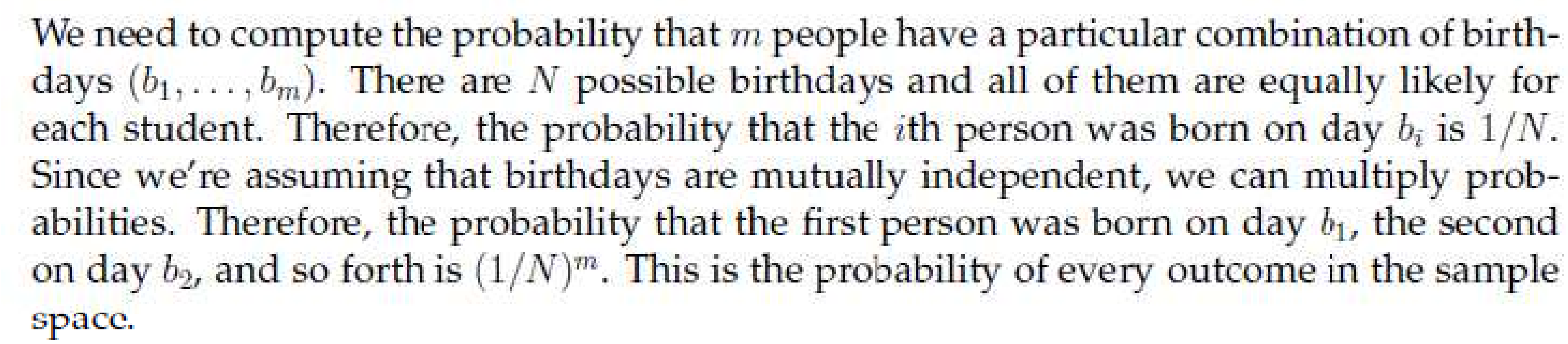
Our goal is to determine the probability of the event A, in which some two people have the same birthday. This event is a little awkward to study directly, however. So we'll use a common trick, which is to analyze the *complementary* event $\overline{A}$, in which all m people have different birthdays:

$$\overline{A} = \{(b_1, \ldots, b_m) \in S \mid \text{all } b_i \text{ are distinct}\}$$

If we can compute $\Pr\left(\overline{A}\right)$, then we can compute what we really want, $\Pr(A)$, using the relation:

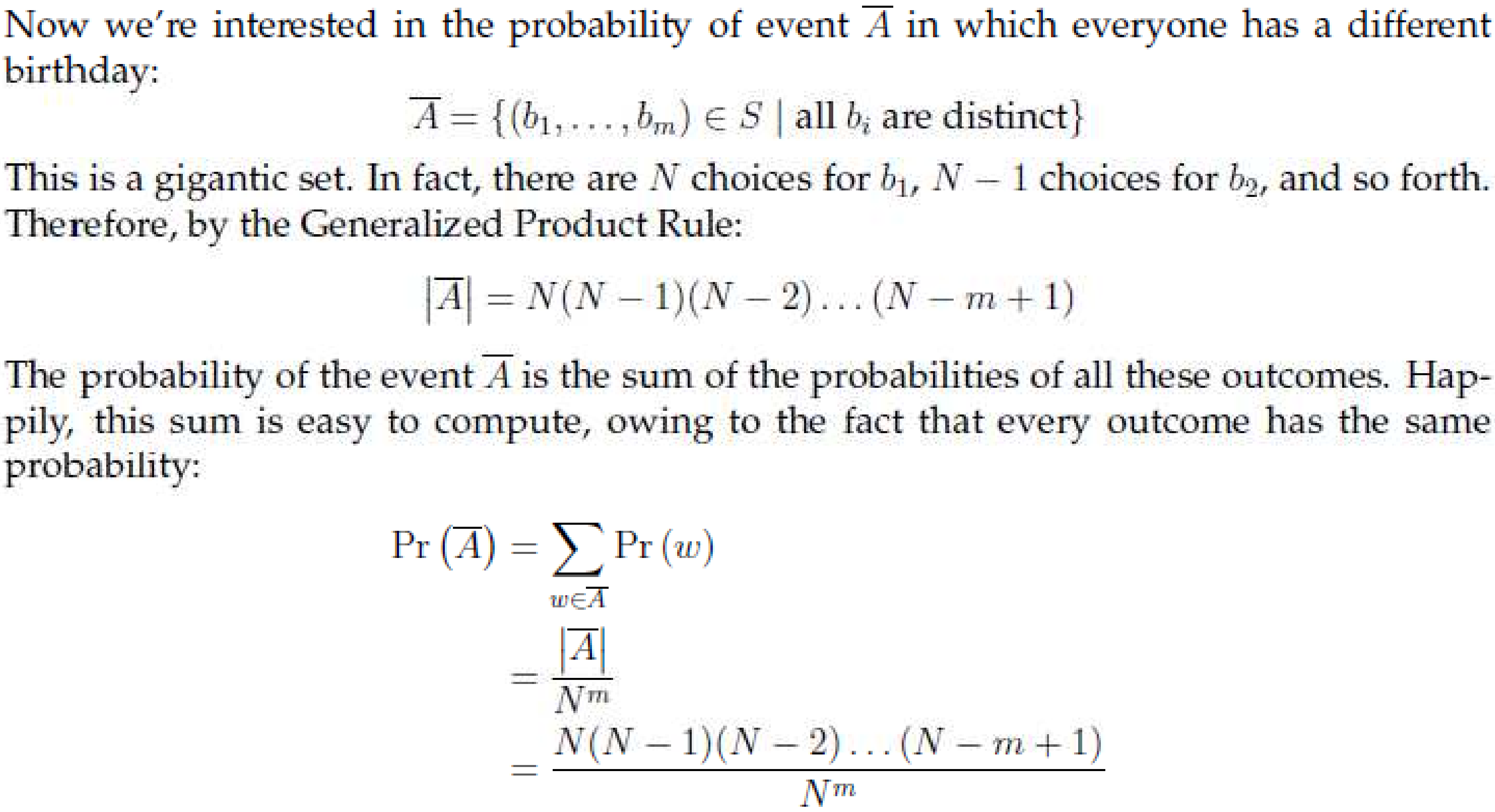
$$\Pr(A) + \Pr\left(\overline{A}\right) = 1$$

Step 3: *Assign outcome probabilities*



We need to compute the probability that m people have a particular combination of birthdays $(b_1, \ldots, b_m)$. There are N possible birthdays and all of them are equally likely for each student. Therefore, the probability that the ith person was born on day b_i is $1/N$. Since we're assuming that birthdays are mutually independent, we can multiply probabilities. Therefore, the probability that the first person was born on day b_1, the second on day b_2, and so forth is $(1/N)^m$. This is the probability of every outcome in the sample space.

***Step 4**: Compute event probabilities*



Now we're interested in the probability of event $\overline{A}$ in which everyone has a different birthday:

$$\overline{A} = \{(b_1, \ldots, b_m) \in S \mid \text{all } b_i \text{ are distinct}\}$$

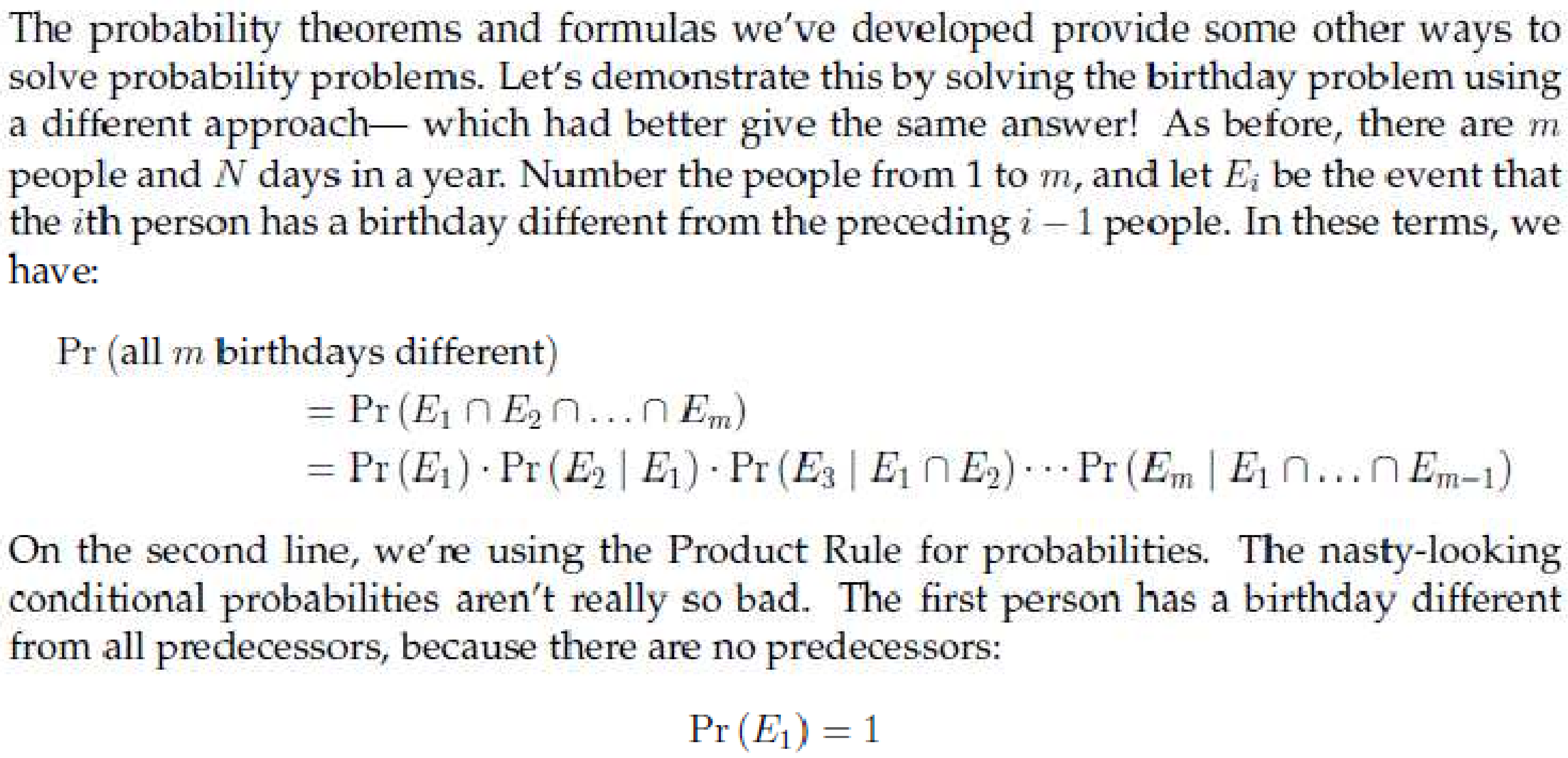
This is a gigantic set. In fact, there are N choices for b_1, $N-1$ choices for b_2, and so forth. Therefore, by the Generalized Product Rule:

$$|\overline{A}| = N(N-1)(N-2)\ldots(N-m+1)$$

The probability of the event $\overline{A}$ is the sum of the probabilities of all these outcomes. Happily, this sum is easy to compute, owing to the fact that every outcome has the same probability:

$$\begin{aligned}
\Pr\left(\overline{A}\right) &= \sum_{w \in \overline{A}} \Pr(w) \\
&= \frac{|\overline{A}|}{N^m} \\
&= \frac{N(N-1)(N-2)\ldots(N-m+1)}{N^m}
\end{aligned}$$

An alternative approach



The probability theorems and formulas we've developed provide some other ways to solve probability problems. Let's demonstrate this by solving the birthday problem using a different approach— which had better give the same answer! As before, there are m people and N days in a year. Number the people from 1 to m, and let E_i be the event that the ith person has a birthday different from the preceding $i-1$ people. In these terms, we have:

$$\begin{aligned}
&\Pr(\text{all } m \text{ birthdays different}) \\
&\qquad = \Pr(E_1 \cap E_2 \cap \ldots \cap E_m) \\
&\qquad = \Pr(E_1) \cdot \Pr(E_2 \mid E_1) \cdot \Pr(E_3 \mid E_1 \cap E_2) \cdots \Pr(E_m \mid E_1 \cap \ldots \cap E_{m-1})
\end{aligned}$$

On the second line, we're using the Product Rule for probabilities. The nasty-looking conditional probabilities aren't really so bad. The first person has a birthday different from all predecessors, because there are no predecessors:

$$\Pr(E_1) = 1$$

We're assuming that birthdates are equally probable and birthdays are independent, so the probability that the second person has the same birthday as the first is only $1/N$. Thus:

$$\Pr(E_2 \mid E_1) = 1 - \frac{1}{N}$$

Given that the first two people have different birthdays, the third person shares a birthday with one or the other with probability $2/N$, so:

$$\Pr(E_3 \mid E_1 \cap E_2) = 1 - \frac{2}{N}$$

Extending this reasoning gives:

$$\Pr(\text{all } m \text{ birthdays different}) = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{m-1}{N}\right)$$

We're done— again! This is our previous answer written in a different way.