An introduction to probability

Ben Lambert

University of Oxford

ben.c.lambert@gmail.com

August 31, 2021

Who am I?

- Statistician working mainly in epidemiology.
- Claim to probability fame: born in the same town where Thomas Bayes lived (Tunbridge Wells, UK).

Course outline

- 9am-10.30am: follow along lecture and problems
- 10.30am-10.45am: refreshments break
- 10.45am-midday: follow along lecture and problems

Note: problems will involve both pen and paper type questions *and* those using R.
Levels: each problem set will have at least one *advanced* question.
Time: if you don't finish the questions in time, don't worry. There are answers to these here:
`https://github.com/ben18785/introduction_to_probability`.

Resources

- Introduction to probability, Blitzstein and Hwang. Open source book available here:
  `https://projects.iq.harvard.edu/stat110/home`
- Seeing theory, Kunin et al. A beautiful online resource that has lots of creative ways to think about probability.
  `https://seeing-theory.brown.edu/`

Outline

1. What is probability and why do we need it?
2. Probability and counting
3. Conditional probability
4. Bayes' rule
5. Random variables and probability distributions
6. Expectations
7. Joint distributions
8. Continuous probability distributions

1 What is probability and why do we need it?

2 Probability and counting

3 Conditional probability

4 Bayes' rule

5 Random variables and probability distributions

6 Expectations

7 Joint distributions

8 Continuous probability distributions

What is probability?

Mathematics is the logic of certainty; probability [theory] is the [mathematical] logic of uncertainty

---

Bitzstein and Hwang, 2019

Why do we need probability theory?

Life and science are full of unknown things. We say these things are *uncertain*.

Faced with these, we can give up; *Probability theory* gives us a way to make assumptions about uncertain phenomena which allows us to make progress with understanding without having to know everything.

> There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.
>
> Donald Rumsfeld, 2002

Who uses probability?

It is used in:

- Statistics: probability is the foundational language of it
- Biology: e.g. inheritance of genes
- Meteorology: e.g. weather forecasts are generated using probabilities
- Epidemiology: e.g. analysing randomised clinical trials and fitting models to epidemiological data
- Physics: our current best explanation of the universe at small scales (quantum theory) is based on probability

The difficulties of probability

If we rely on our intuitions, we can easily get things wrong when dealing with uncertainty.

So, we need careful mathematical analysis.

Fortunately, *simulation* using R / Python / etc. can also really help us to understand.

1 What is probability and why do we need it?

2 Probability and counting

3 Conditional probability

4 Bayes' rule

5 Random variables and probability distributions

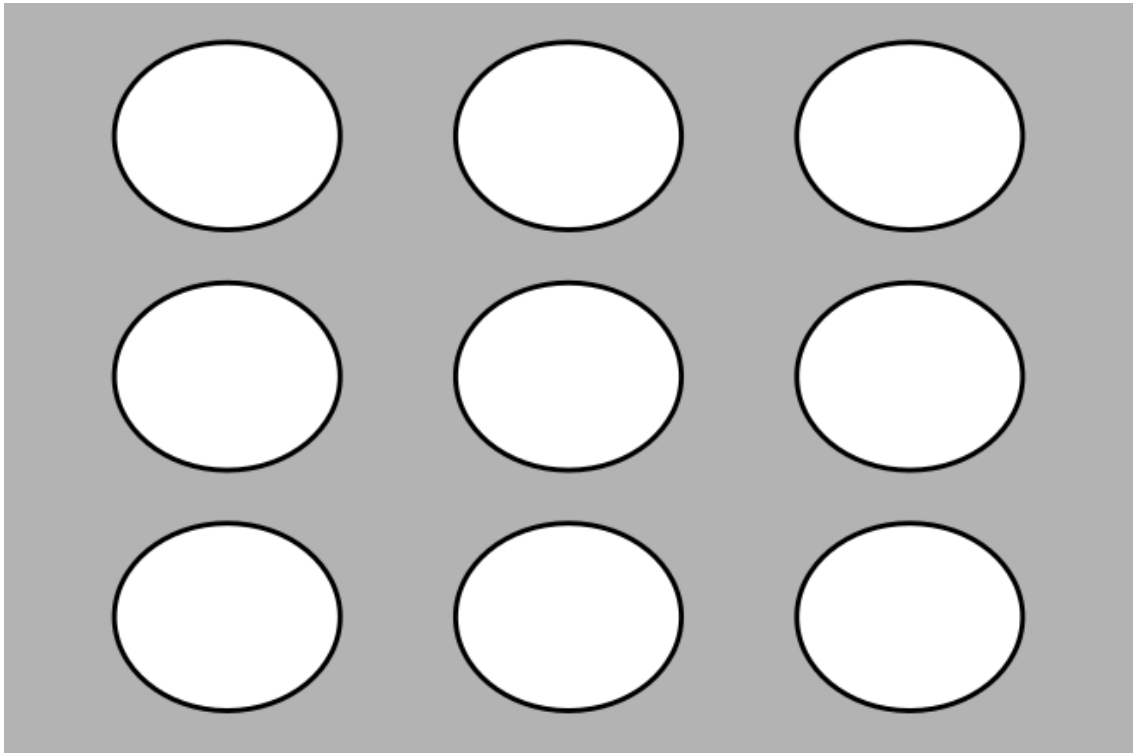
6 Expectations

7 Joint distributions

8 Continuous probability distributions

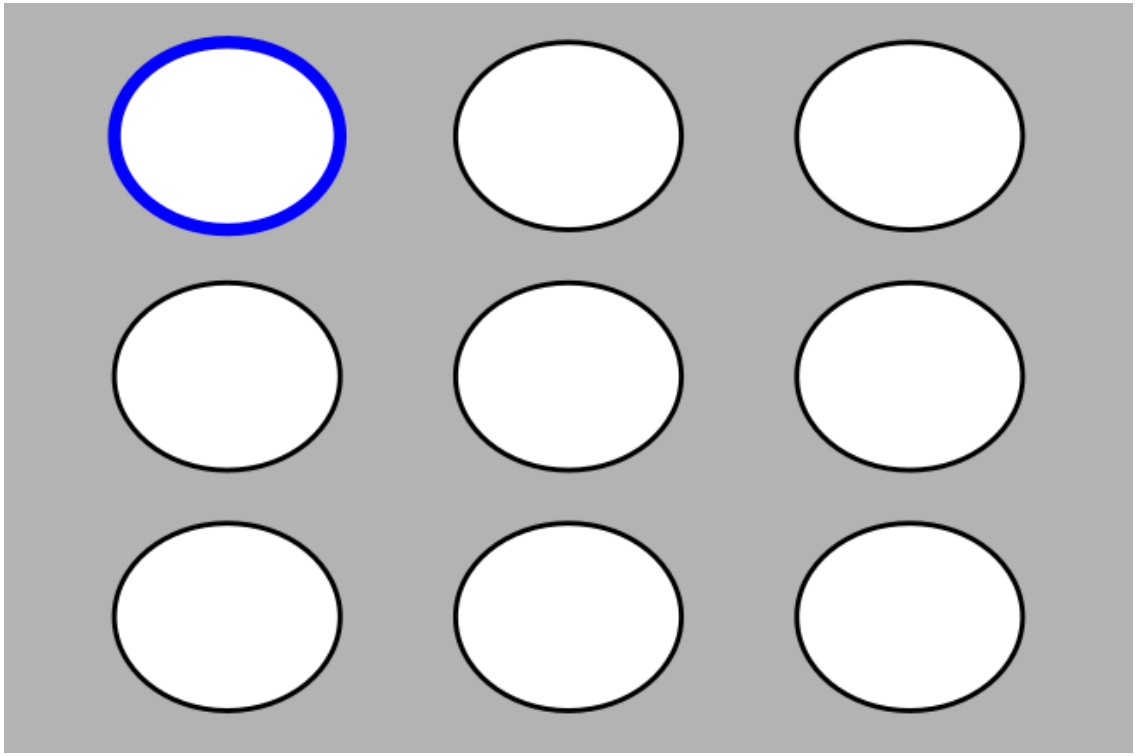
Blitzstein and Hwang's Pebble World

As an example, consider reaching into a bag to pull out one of nine pebbles: we call this *pebble world*.



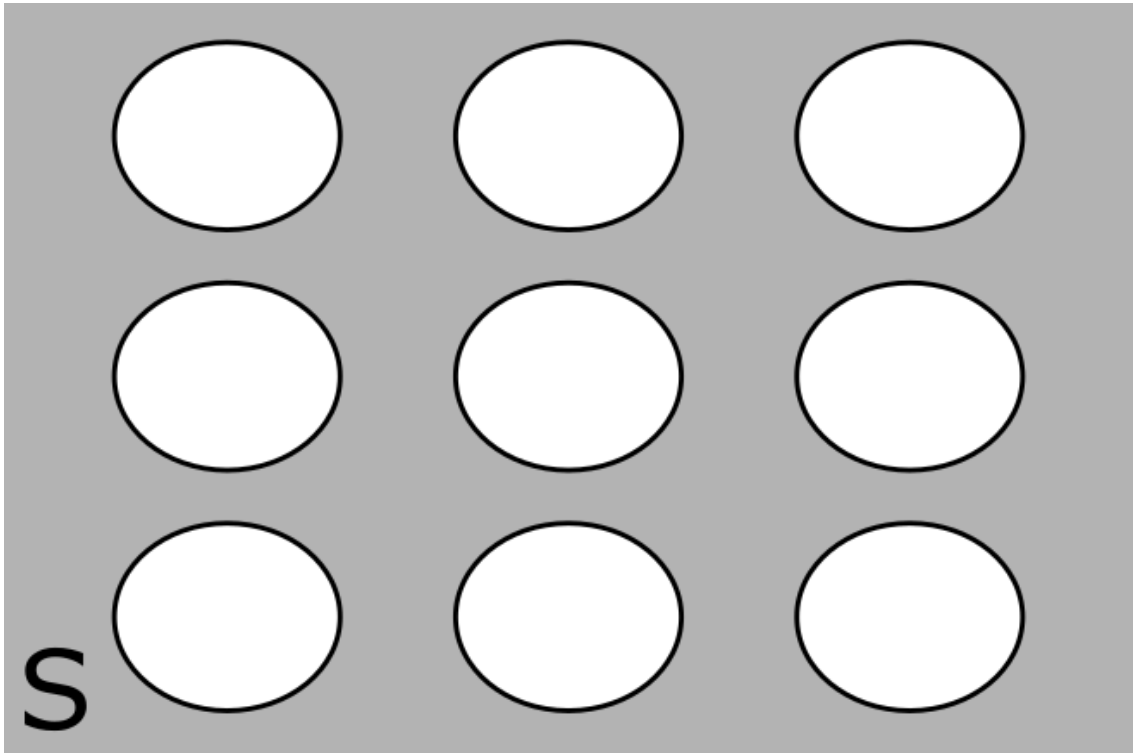
Outcomes

An *outcome* is a possible result of some activity. Here pulling one particular pebble out of the bag would be an outcome.



Sample spaces

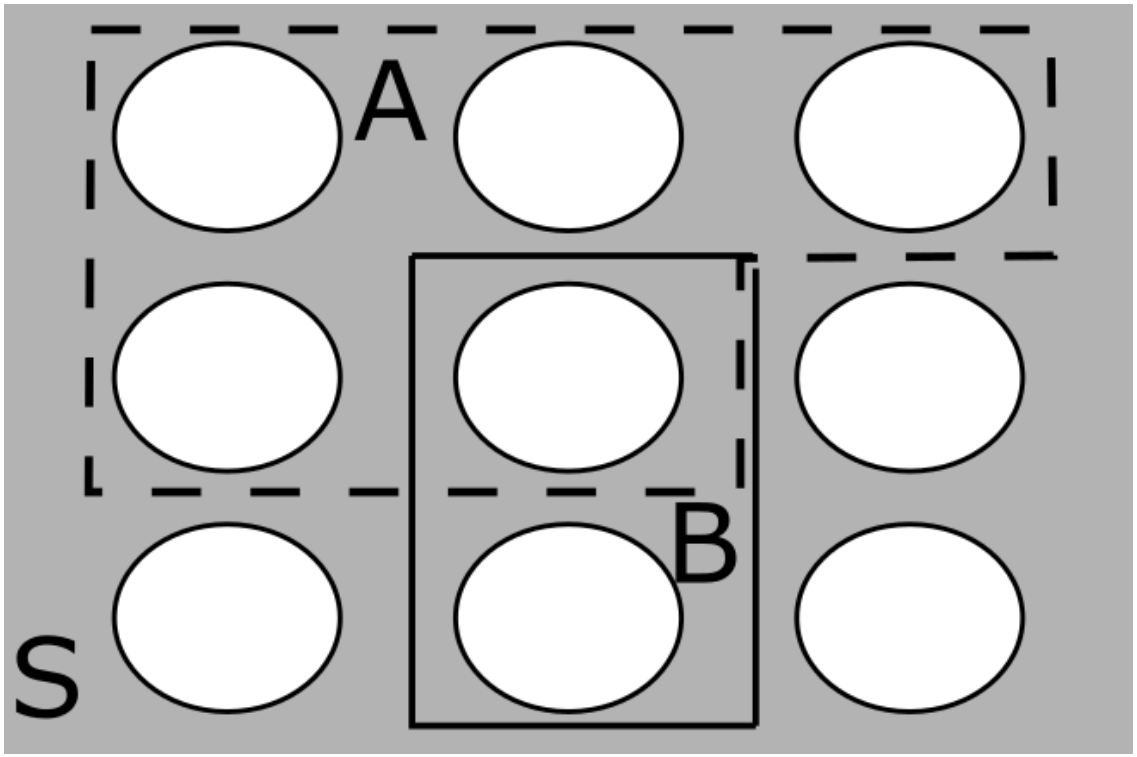
The *sample space*, S, is the *set* of all possible outcomes of an experiment. Here, it is the set of all pebbles.



Events

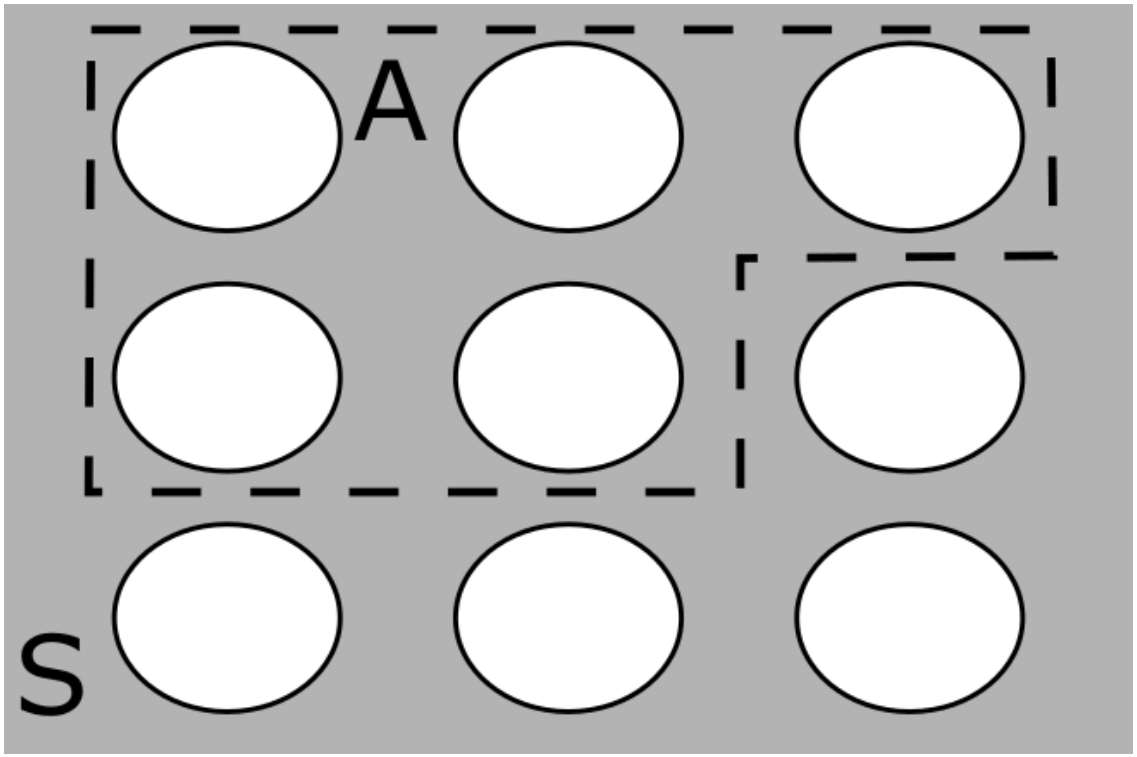
An *event* is a *set* of possible outcomes. For example, below event A corresponds to selecting one of five pebbles; event B to selecting one of two.

As we can see, two or more events can happen at once.



Probabilities of events

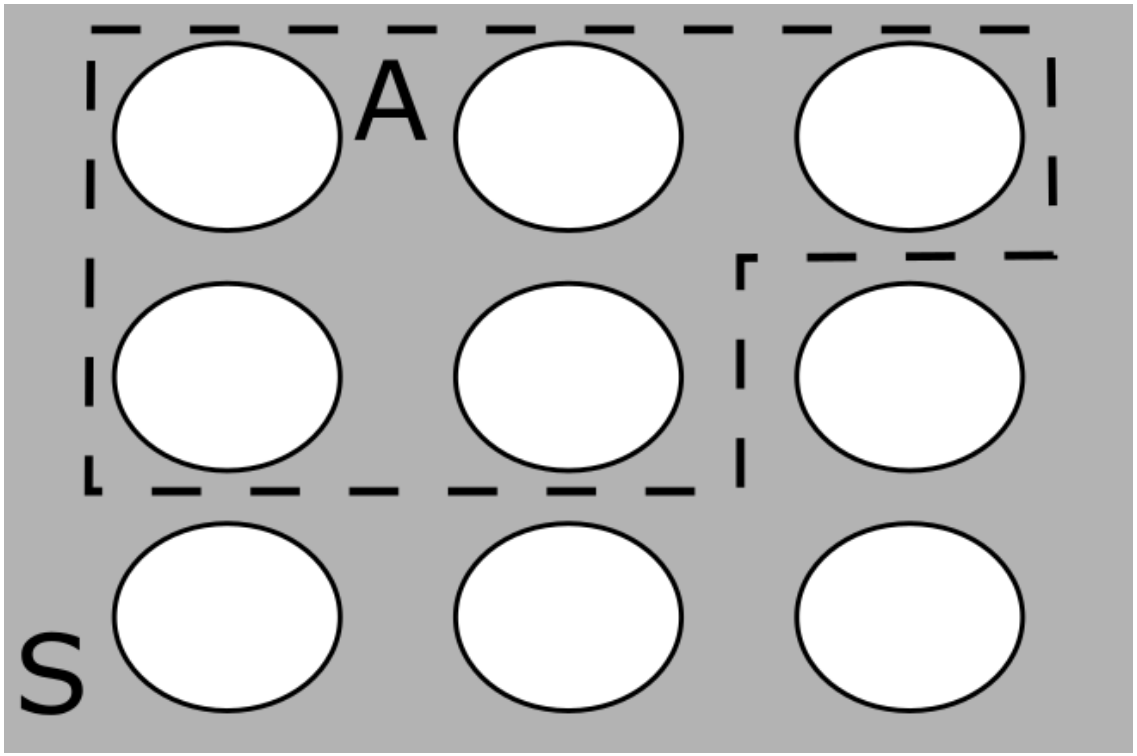
What do we mean by the probability of A occurring? We write this as: $\mathbb{P}(A)$.



Probabilities from counting

If all pebbles equally likely to be drawn:

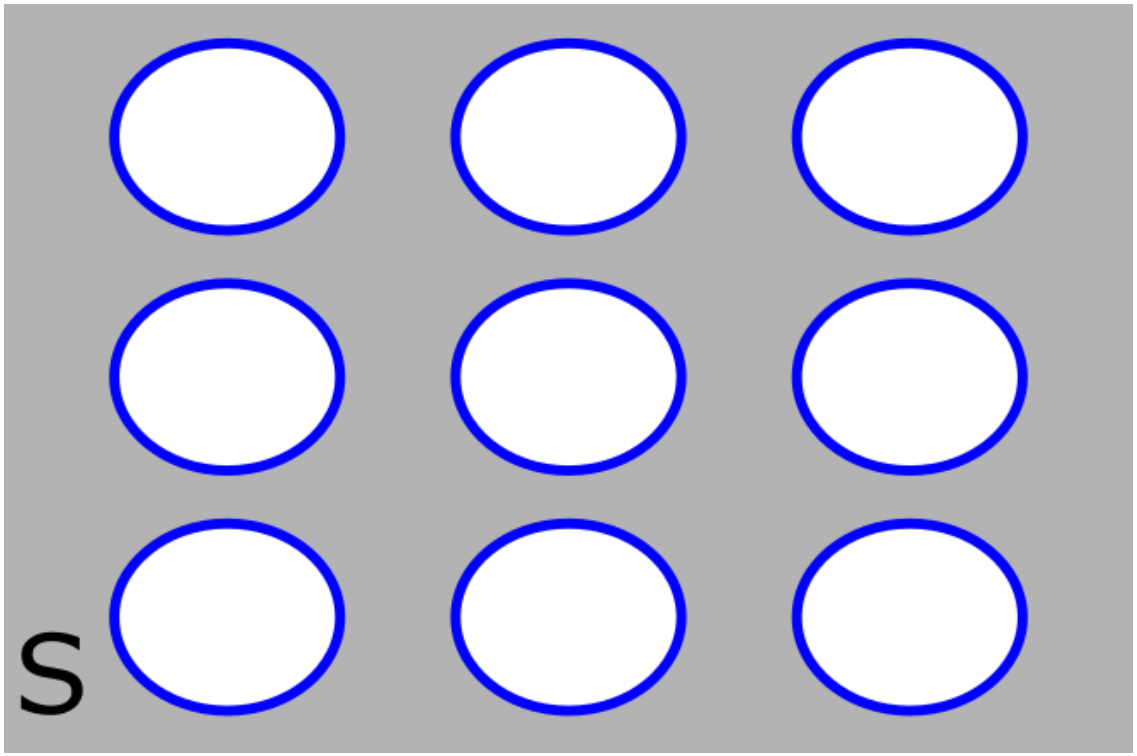
$$\mathbb{P}(A) = \frac{\text{number of pebbles in } A}{\text{number of pebbles in } S} = \frac{5}{9} \tag{1}$$



Probability of an event in S

Consider the probability that some event in S occurs:

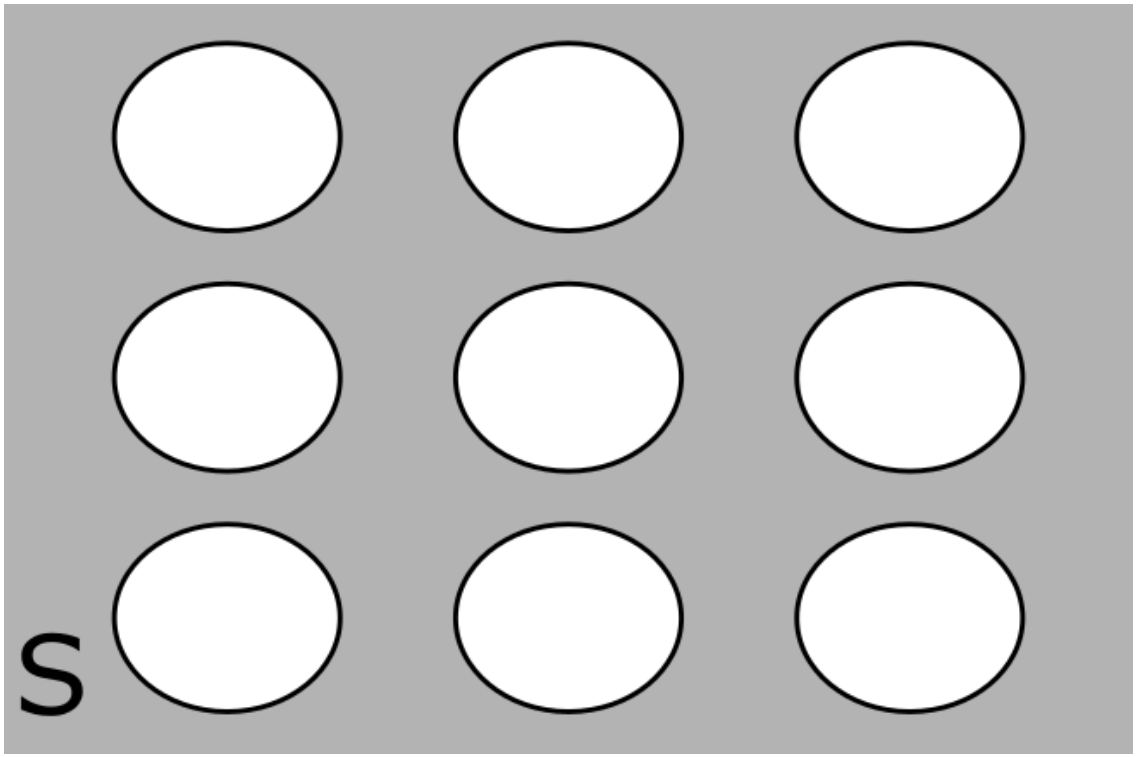
$$\mathbb{P}(S) = \frac{\text{number of pebbles in } S}{\text{number of pebbles in } S} = \frac{9}{9} = 1 \tag{2}$$



Probability of null event

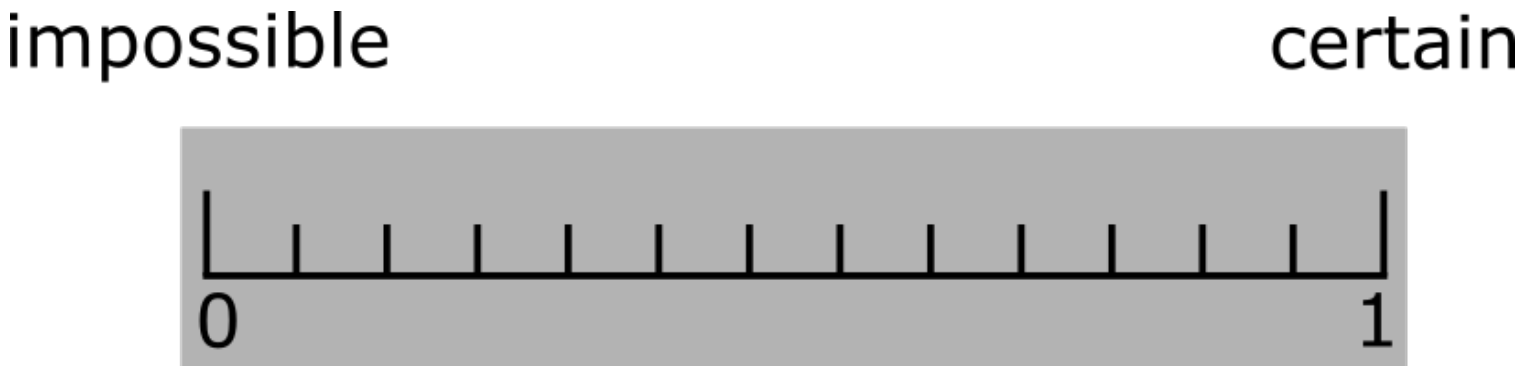
What about the probability that no event in S occurs?

$$\mathbb{P}(\text{not } S) = \frac{0}{9} = 0 \tag{3}$$



Defining probability

A probability of an event must be bounded between 0 and 1[1].



[1]Note: "impossible" isn't 100% accurate here but you can start out by thinking of probabilities this way.

Interpreting probabilities

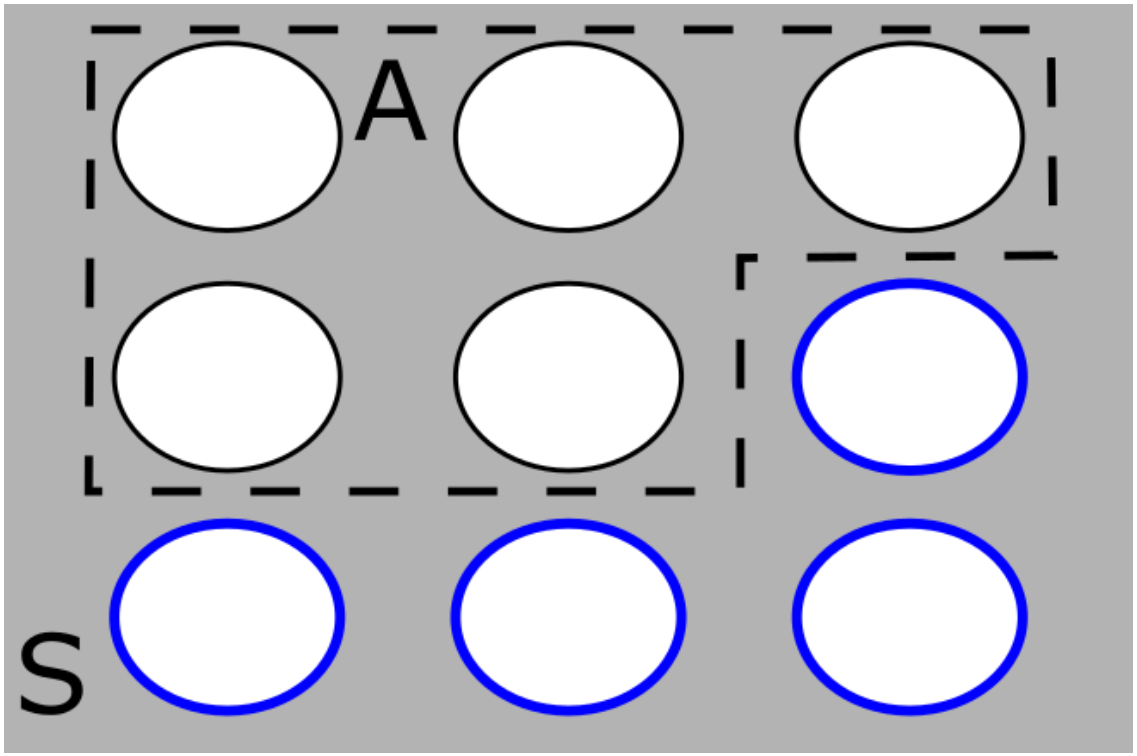
There are many schools of thought for what probabilities mean. Two common ones are:

- *Frequentist*. Think of probabilities as frequencies that would be obtained under many (actually infinite) repetitions of an experiment. E.g. flipping a coin a large number of times and using fraction of heads as $\mathbb{P}(H)$.
- *Bayesian*. Suppose probabilities reflect an underlying subjective belief about the chance of events occurring.

An event not occurring

We can also determine the probability that A does not occur:

$$\mathbb{P}(\text{not } A) = \frac{\text{number of pebbles not in } A}{\text{number of pebbles in } S} = \frac{4}{9} \tag{4}$$



Question

Can anyone think of an alternative way of determining the probability that A does not occur?

Question

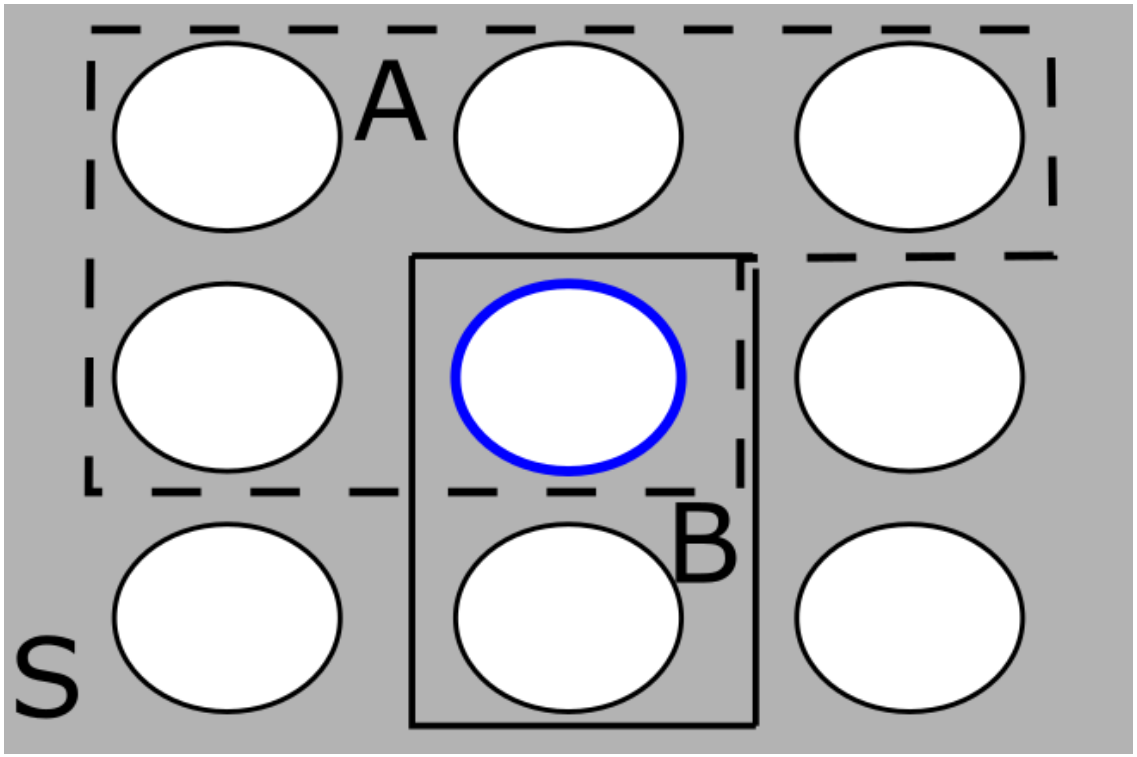
Can anyone think of an alternative way of determining the probability that A does not occur?

$$\mathbb{P}(\text{not } A) = \mathbb{P}(S) - \mathbb{P}(A) = 1 - \mathbb{P}(A) = 1 - \frac{5}{9} \qquad (5)$$

Combinations of events: intersection

We can determine the probability of A *and* B occurring by determining the overlap between these two events:

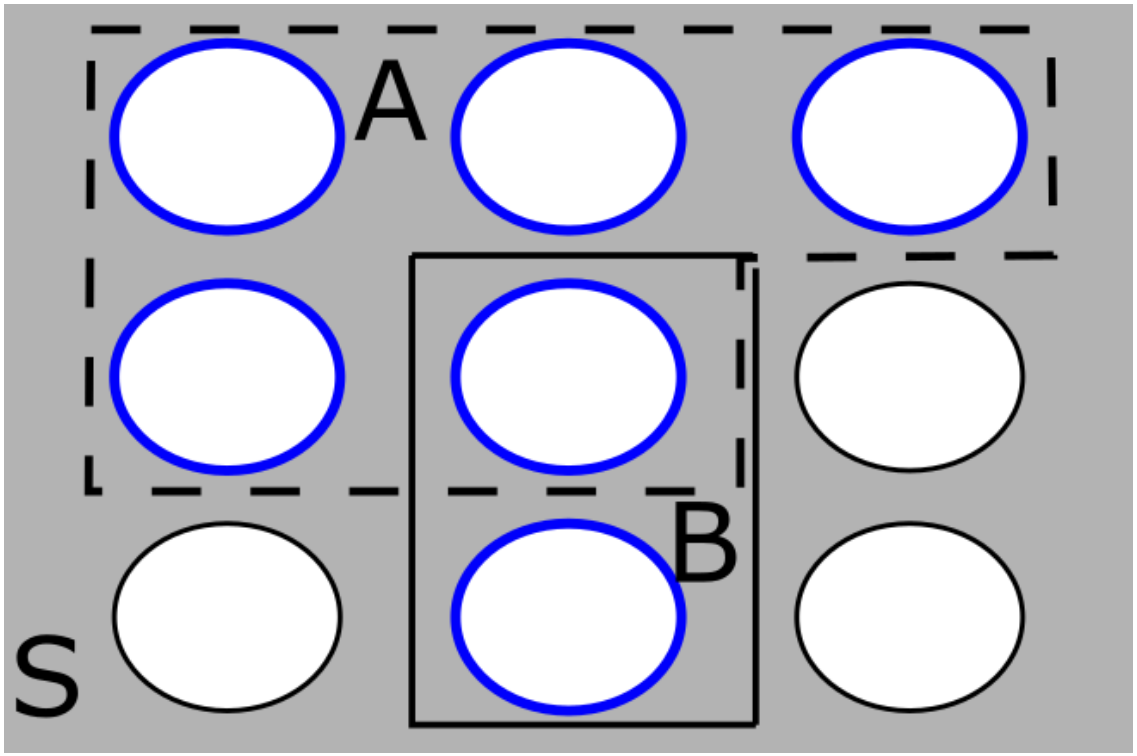
$$\mathbb{P}(A \cap B) = \mathbb{P}(A, B) = \frac{\text{number of pebbles in both } A \text{ and } B}{\text{number of pebbles in } S} = \frac{1}{9} \tag{6}$$



Combinations of events: union

We can determine the probability of A *and/or* B occurring by:

$$\mathbb{P}(A \cup B) = \frac{\text{number of pebbles in either } A \text{ or } B \text{ or both}}{\text{number of pebbles in } S} = \frac{6}{9} \tag{7}$$



Question

Can anyone think of an alternative way of determining the probability of the union of A and B?

Question

Can anyone think of an alternative way of determining the probability of the union of A and B?

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \tag{8}$$

$$= \frac{5}{9} + \frac{2}{9} - \frac{1}{9} = \frac{6}{9} \tag{9}$$

Questions?

Problems: dice

Consider a six-sided die with numbers 1-6 on each face that is thrown once.

1. What's the probability that a six is thrown?
2. What's the probability that an even number is thrown?
3. Suppose two dice are thrown, what's the probability that the their sum adds up to 11 or less?
4. *Advanced*: if six dice are thrown, what's the probability that one of each of the numbers is obtained? If you like, use R to check your result.

Example: a random sweet

Suppose you visit a sweet shop. The shop produces both ice cream and cake. It also offers three sauces: strawberry, vanilla and chocolate.

Unlike most shops, you don't get a choice. The way the shopkeeper allocates food is they randomly choose either a ice cream or cake (selecting either with equal probability). They then randomly select a sauce from the three available (again with equal probability of each).

Suppose you like anything with chocolate sauce. What's the probability that you obtain this?

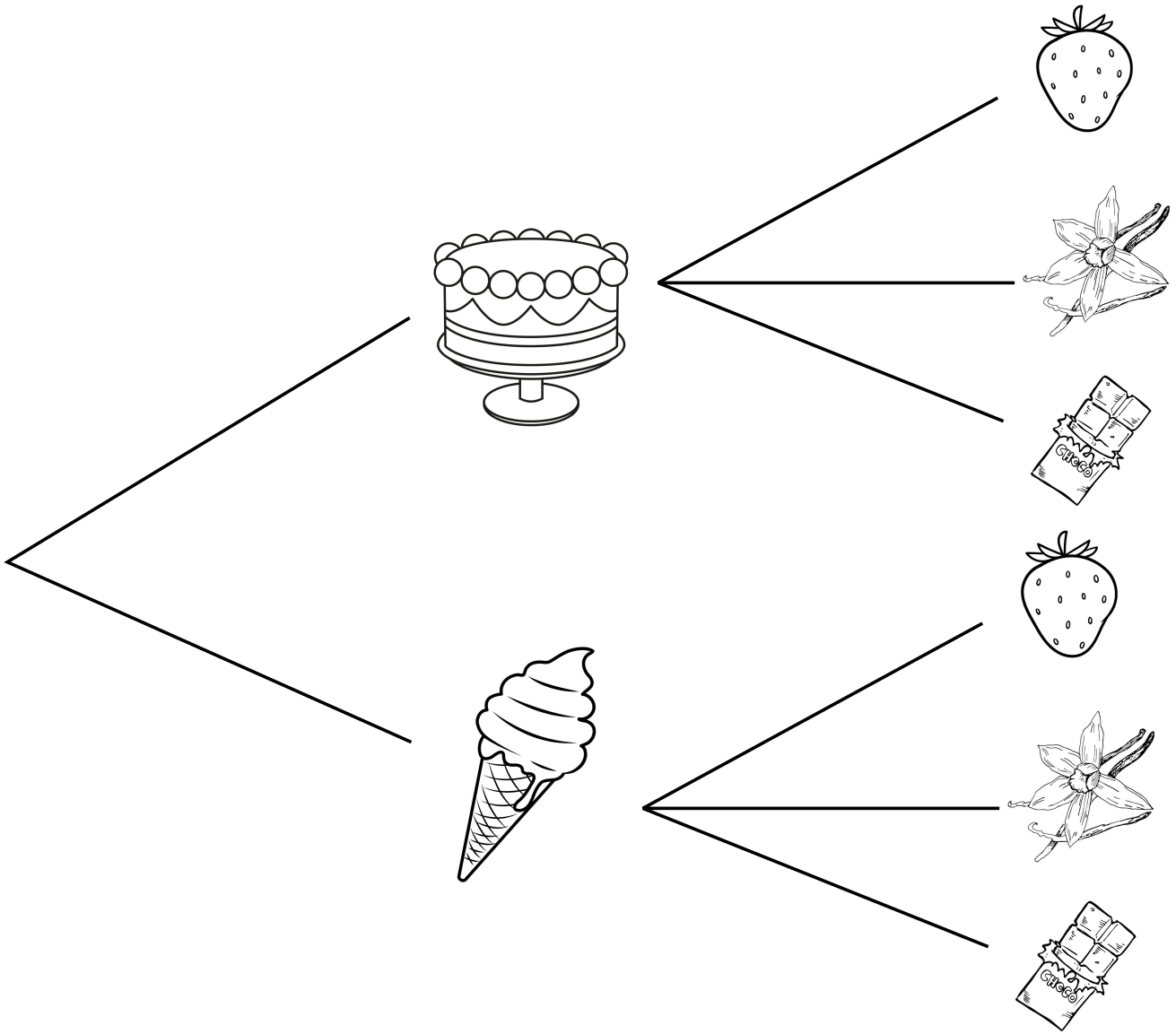
A naive way to count

How many outcomes are there? Cake with strawberry, cake with vanilla, cake with chocolate, ice cream with strawberry, ice cream with vanilla, ice cream with chocolate. So there are *six* outcomes.

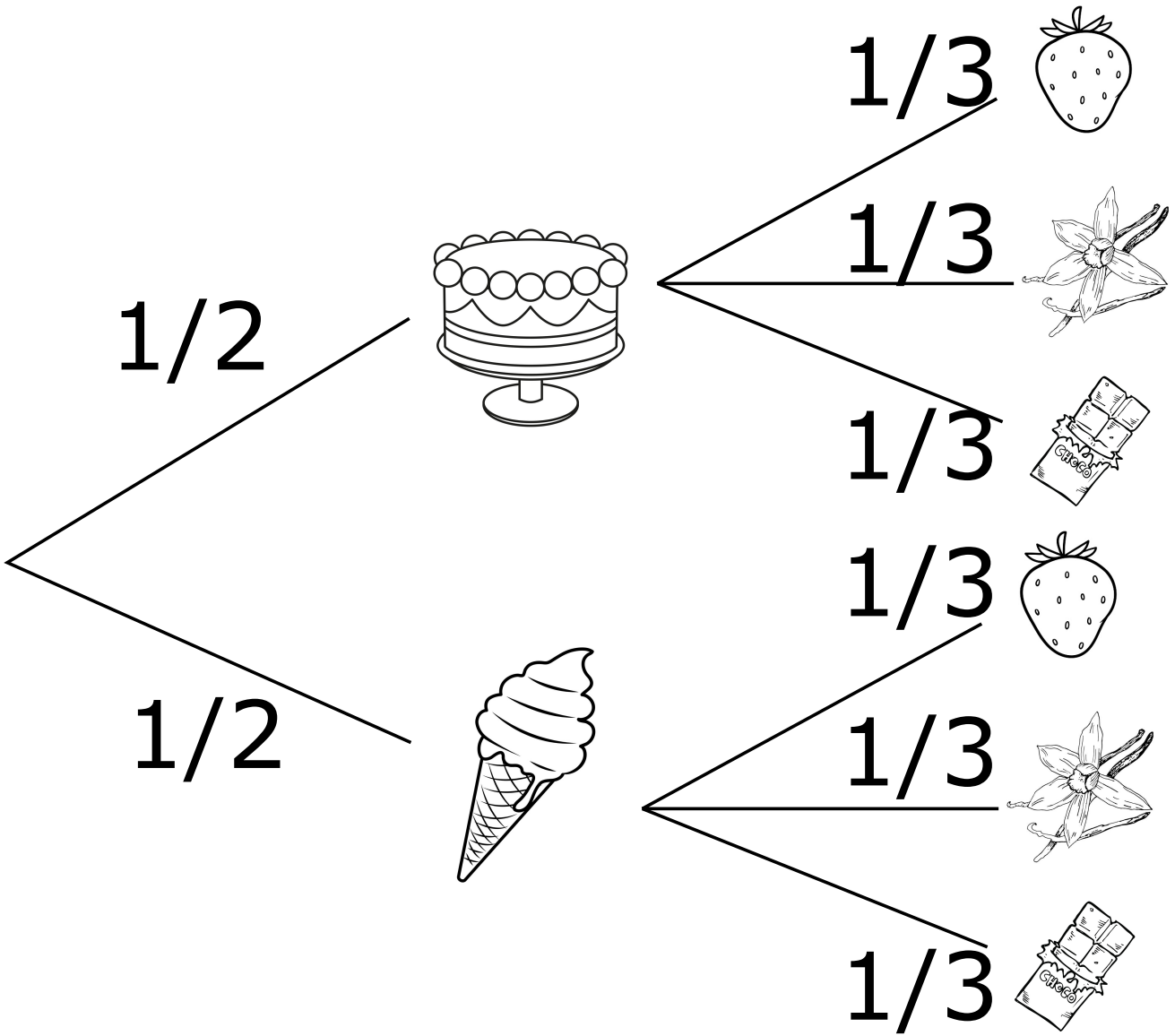
How many of these have chocolate sauce: two.

$$\mathbb{P}(\text{chocolate}) = \frac{2}{6} \tag{10}$$

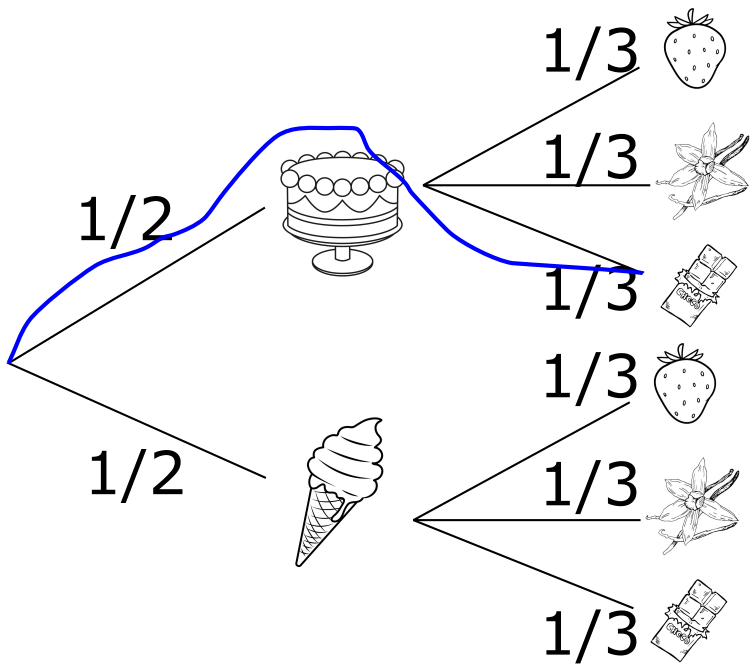
A sweet tree: making counting easier



Using trees to determine probabilities



What's the probability of cake with chocolate sauce?

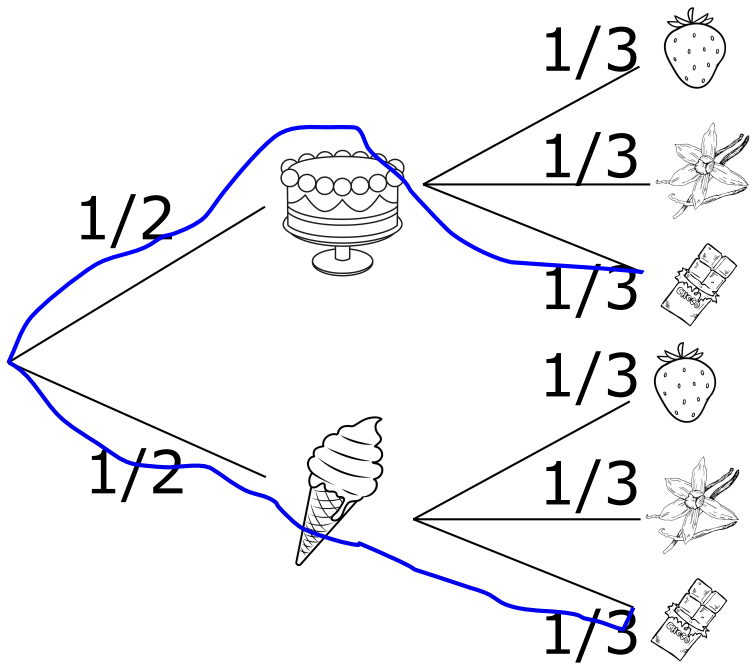


To obtain this probability, we take the probabilities obtained along the path and multiply:

$$\mathbb{P}(\text{chocolate cake}) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6} \tag{11}$$

This is equivalent to counting possibilities (if all choices equally likely).

Chocolate probability



Sum up probabilities:

$$\mathbb{P}(\text{chocolate}) = \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} = \frac{2}{6} \tag{12}$$

A particular shopkeeper

Now suppose we visit another shop. Here, the shopkeeper allocates cake or ice cream as before (i.e. with equal probability). The differ in terms of how they offer sauces:

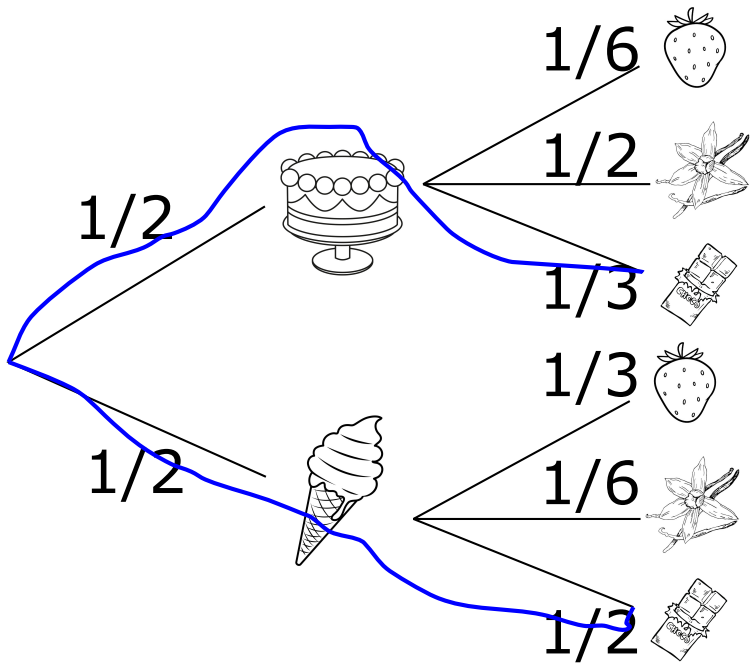
If a cake is chosen, they randomly select sauces: vanilla with probability $1/2$, chocolate with probability $1/3$ and strawberry with probability $1/6$.

If an ice cream is chosen, they randomly select sauces: vanilla with probability $1/6$, chocolate with probability $1/2$ and strawberry with probability $1/3$.

Now what is the probability an item with chocolate sauce is obtained?

The problem with counting

Now different outcomes have different weights, so can't use counting. But tree still works.



Sum up probabilities:

$$\mathbb{P}(\text{chocolate}) = \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{2} = \frac{5}{12} \tag{13}$$

1 What is probability and why do we need it?

2 Probability and counting

3 Conditional probability

4 Bayes' rule

5 Random variables and probability distributions

6 Expectations

7 Joint distributions

8 Continuous probability distributions

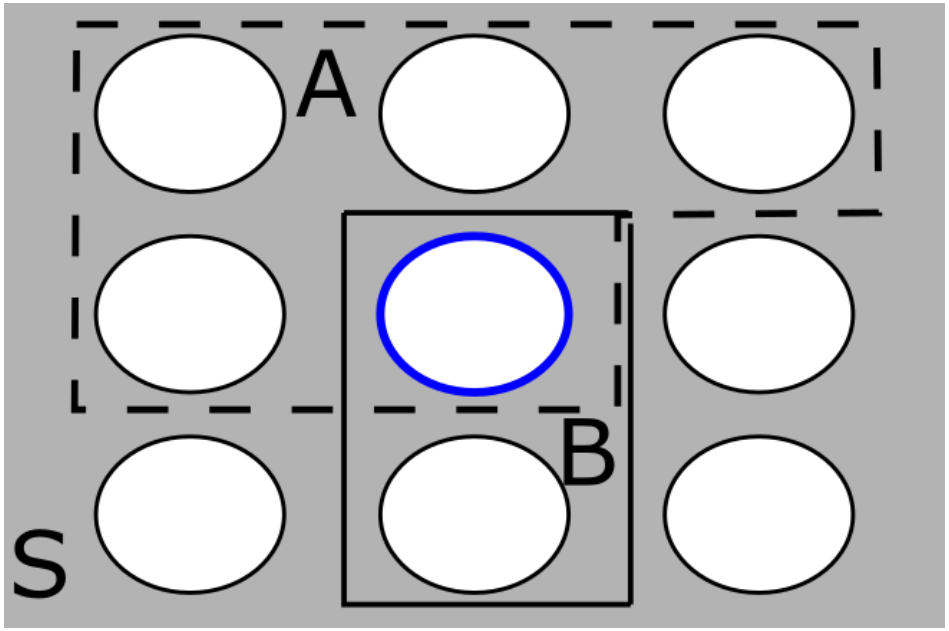
What is conditioning?

When we receive new information, we want to take it into account to make better predictions.

Effectively, learning something about the world (typically) helps us to reduce our own uncertainty.

Conditioning is how this is handled in statistics.

Back to pebble world



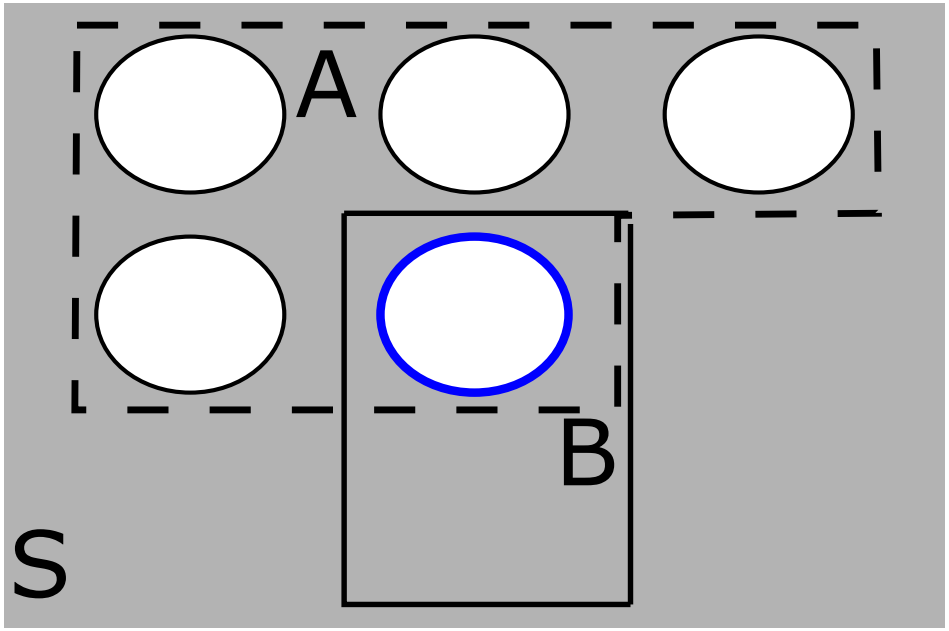
If we know that A has occurred, what's the probability that B has occurred? We write this *conditional* probability as:

$$\mathbb{P}(B|A) \tag{14}$$

which reads, "The probability that B occurs given that A has."

Shrinking sample space

If we know A has occurred, our sample space shrinks.



So now we can just count:

$$\mathbb{P}(B|A) = \frac{1}{5} \tag{15}$$

Law of conditional probability

The below rule effectively renormalises the sample space to calculate updated probabilities:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \tag{16}$$

In our example:

$$\mathbb{P}(B|A) = \frac{1/9}{5/9} = \frac{1}{5} \tag{17}$$

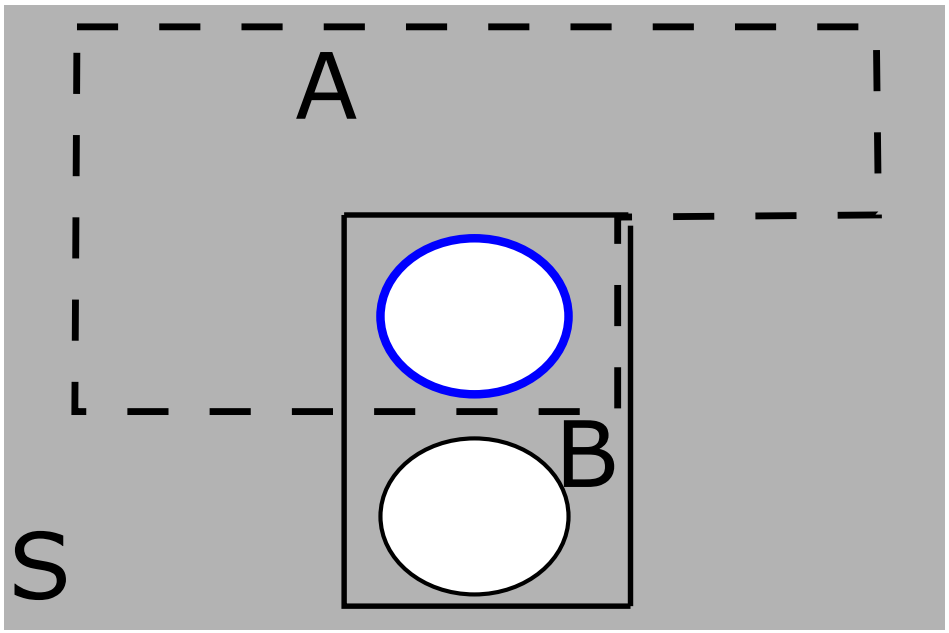
Question

Is $\mathbb{P}(A|B)$ equal to $\mathbb{P}(B|A)$?

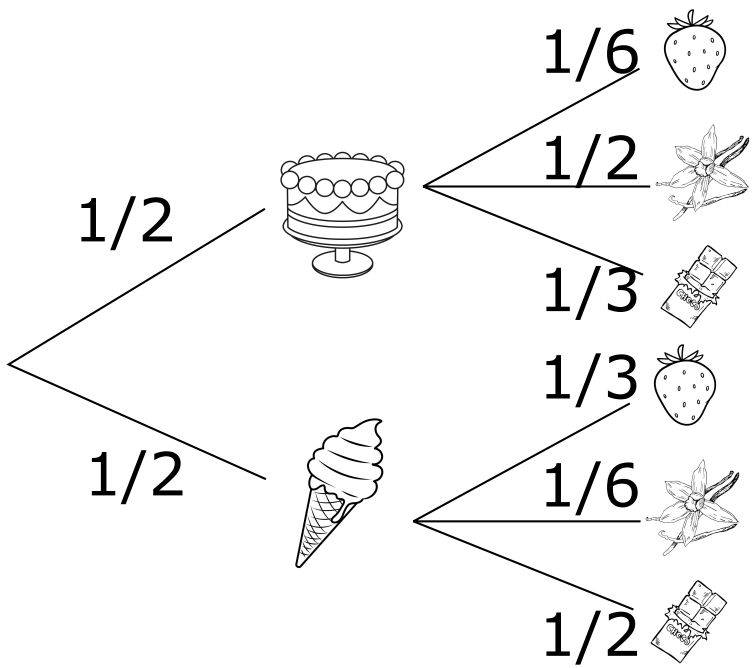
Inverse conditional

No due to different shrunk sample spaces.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/9}{2/9} = \frac{1}{2} \tag{18}$$

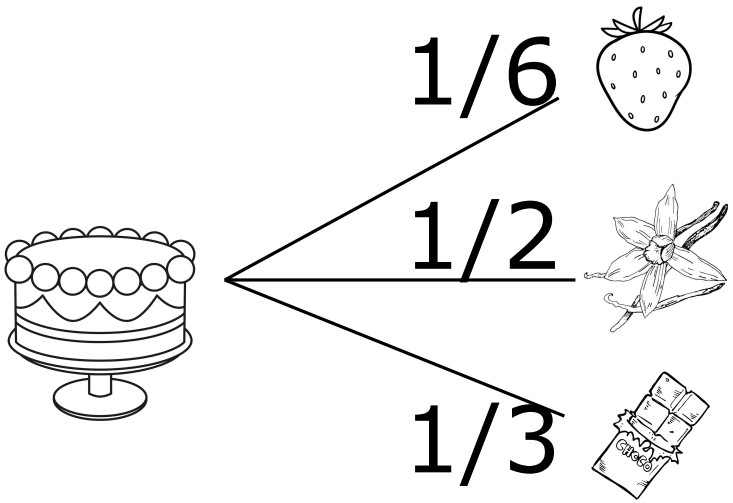


Trees and conditional probability: back to the shop



Question: What's the probability that we get given strawberry sauce given that we receive a cake?

Shrunk sample space

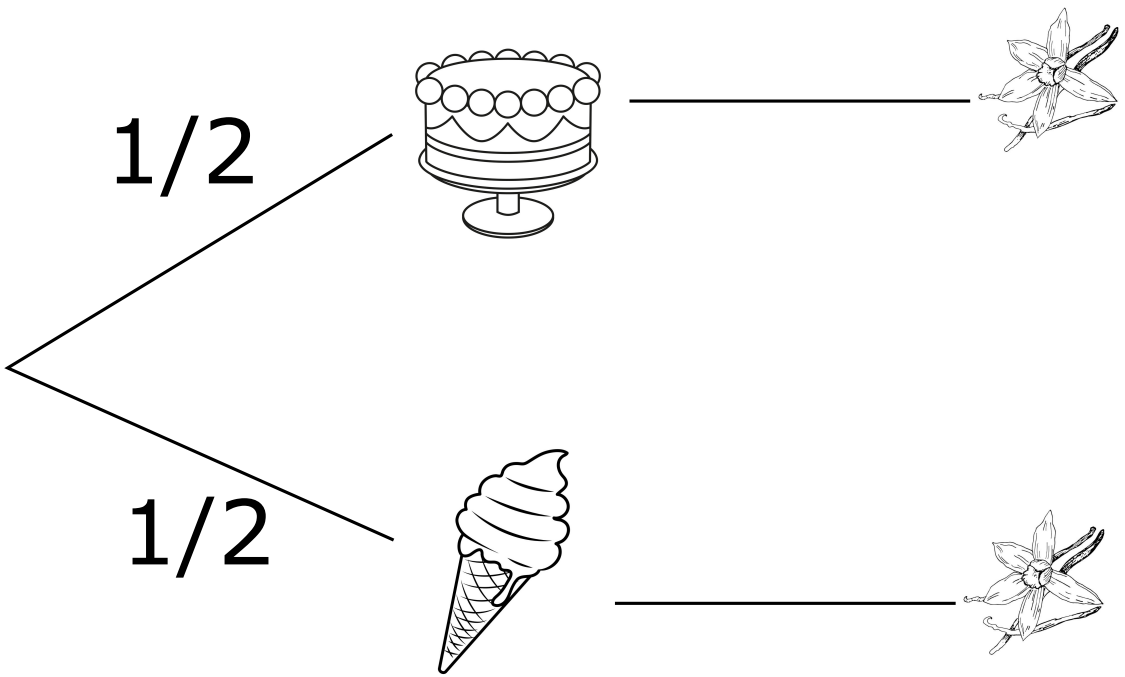


So, we simply read off 1/6.

Question

What's the probability that we received an ice-cream given that we got vanilla sauce?

Why is this not correct?



Meaning that:

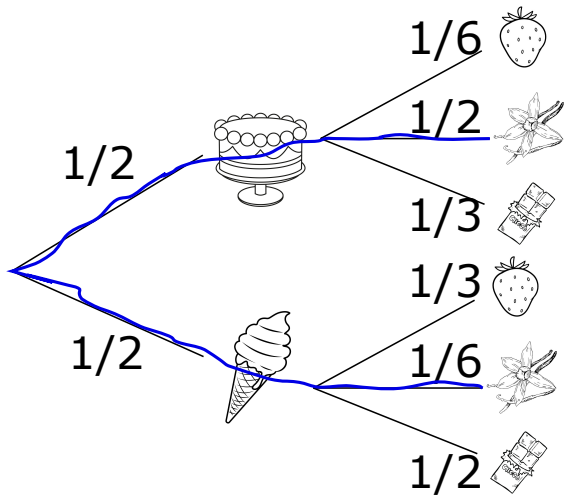
$$\mathbb{P}(\text{ice cream}|\text{vanilla}) = \frac{1}{2} \tag{19}$$

Apply law of conditional probability

$$\mathbb{P}(\text{ice cream}|\text{vanilla}) = \frac{\mathbb{P}(\text{ice cream} \cap \text{vanilla})}{\mathbb{P}(\text{vanilla})} = \frac{1/2 \times 1/6}{\mathbb{P}(\text{vanilla})} \tag{20}$$

Question: how to calculate the denominator $\mathbb{P}(\text{vanilla})$?

Vanilla overall



$$\mathbb{P}(\text{vanilla}) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{6} = \frac{1}{3} \tag{21}$$

Meaning:

$$\mathbb{P}(\text{ice cream}|\text{vanilla}) = \frac{1/2 \times 1/6}{\mathbb{P}(\text{vanilla})} = \frac{1/12}{1/3} = \frac{1}{4} \tag{22}$$

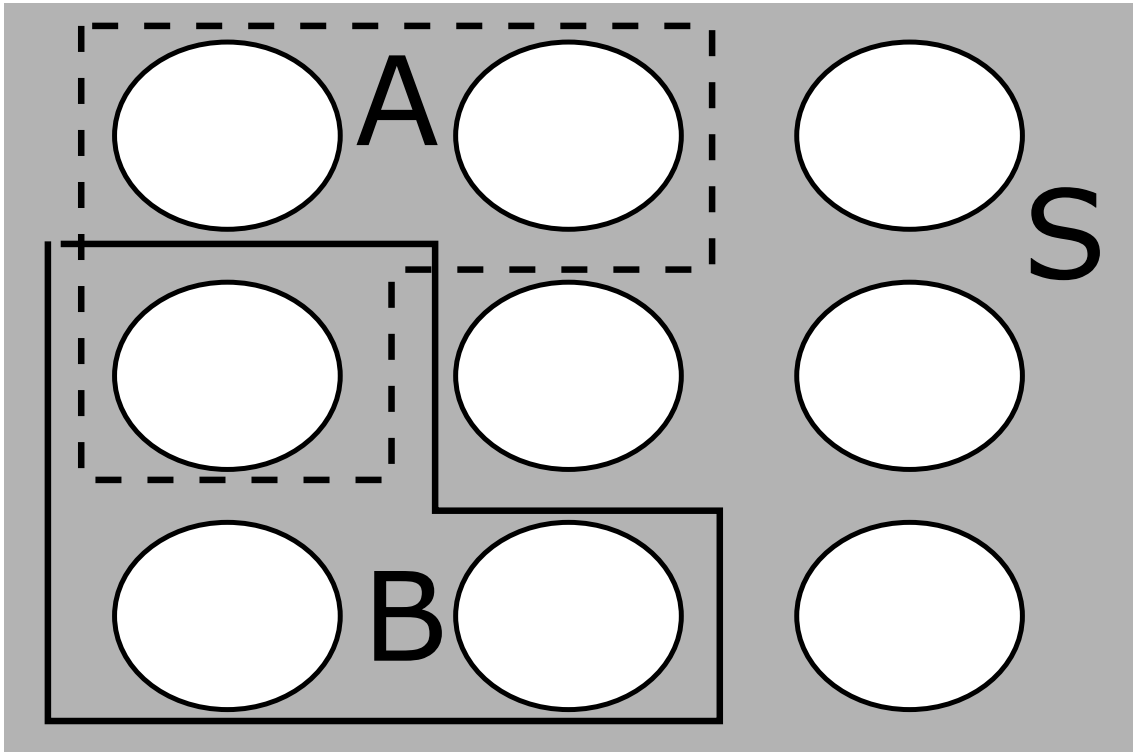
Independence

We say that events A and B are *independent* if:

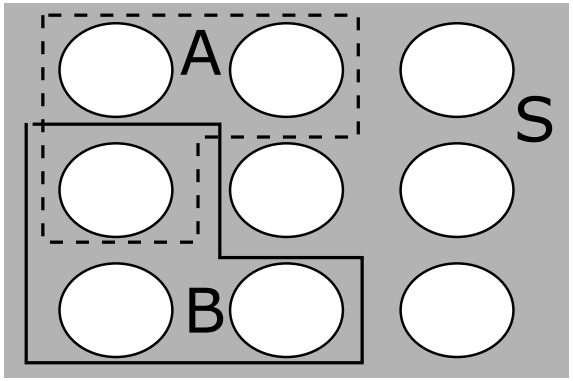
$$\mathbb{P}(A|B) = \mathbb{P}(A) \tag{23}$$

in other words, knowing that B has occurred gives us no additional information about whether A has occurred.

Are A and B independent?



Are A and B independent?



Yes!

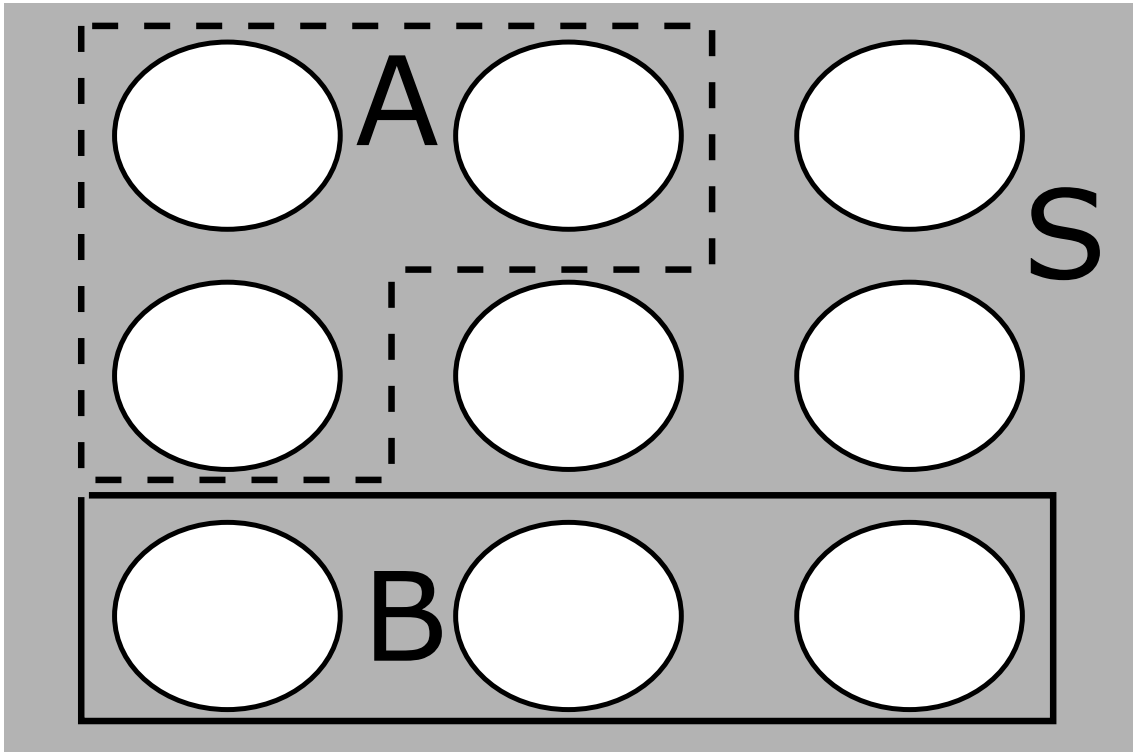
$$\mathbb{P}(A) = \frac{3}{9} = \frac{1}{3} \tag{24}$$

and:

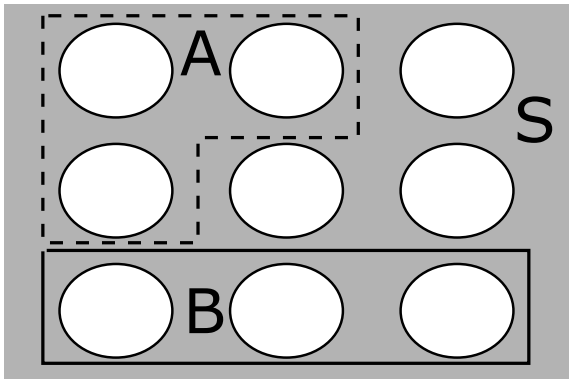
$$\mathbb{P}(A|B) = \frac{1}{3} \tag{25}$$

So $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Redux: Are A and B independent?



Redux: Are A and B independent?



No. Knowing that B occurs tells me that A could not have occurred:

$$\mathbb{P}(A) = \frac{1}{3} \tag{26}$$

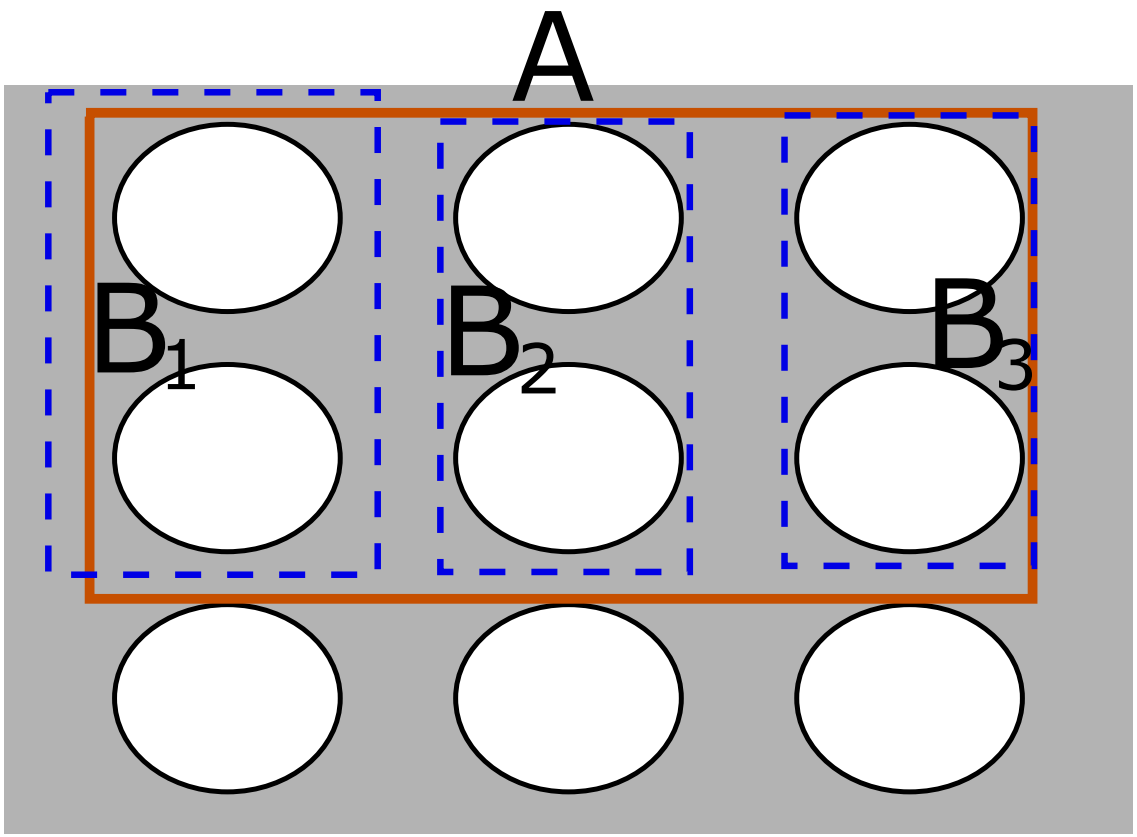
and:

$$\mathbb{P}(A|B) = 0 \tag{27}$$

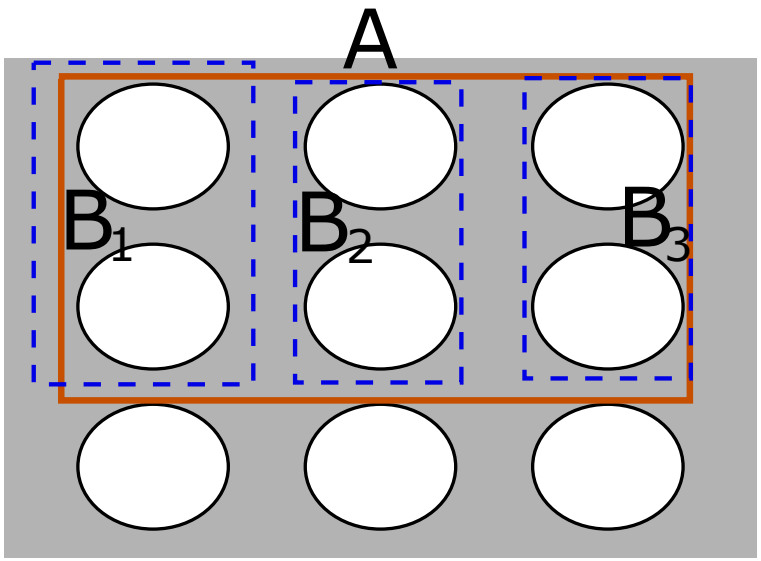
Events A and B are known as *disjoint*.

The law of total probability

How to determine $\mathbb{P}(A)$?



The law of total probability

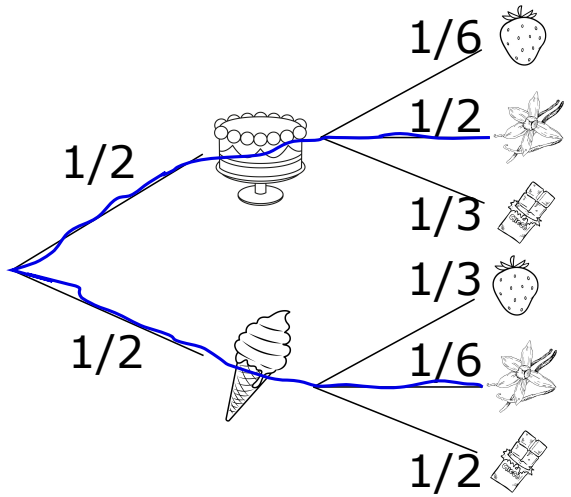


$$\mathbb{P}(A) = \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2) + \mathbb{P}(A|B_3)\mathbb{P}(B_3) \tag{28}$$

More generally (if B_i are disjoint events):

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A|B_i)\mathbb{P}(B_i) \tag{29}$$

Example recap: Law of total probability in action



$$\mathbb{P}(\text{vanilla}) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{6} = \frac{1}{3} \tag{30}$$

Because:

$$\mathbb{P}(\text{vanilla}) = \mathbb{P}(\text{cake}) \times \mathbb{P}(\text{vanilla}|\text{cake}) + \mathbb{P}(\text{ice cream}) \times \mathbb{P}(\text{vanilla}|\text{ice cream}) \tag{31}$$

Questions?

Problems: cards

Suppose you have a standard deck of 52 cards across four suits – two of which are red; two black. The cards in order are: 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A. Imagine in all cases that cards are drawn from a shuffled deck.

1. Consider the event A that a card is a Four and B that the card is black. Are these events A and B independent?
2. Consider the event A that the card is red and B that it is black. Are these events A and B independent?
3. What's the probability that, given the card has a value above 8, that it is an Ace?
4. *Advanced*: suppose two cards are drawn. What's the probability that the second card is an Ace given that the first card is an Ace? What's the probability that second card is an Ace given that at least one is an Ace?

1 What is probability and why do we need it?

2 Probability and counting

3 Conditional probability

4 Bayes' rule

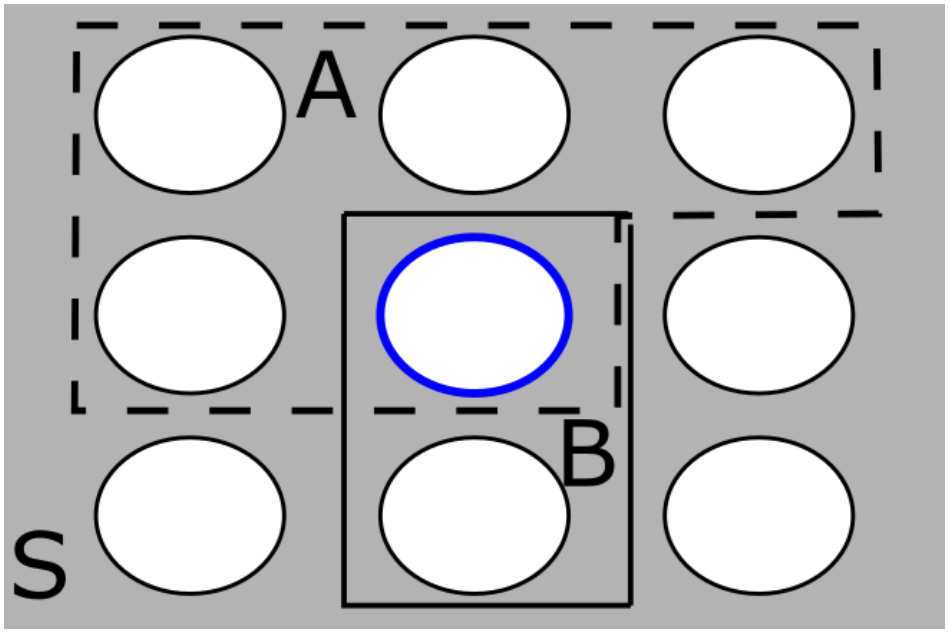
5 Random variables and probability distributions

6 Expectations

7 Joint distributions

8 Continuous probability distributions

Two ways to arrive at same intersection



How to calculate $\mathbb{P}(A \cap B)$? Law of conditional probability:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B|A) = \mathbb{P}(B) \times \mathbb{P}(A|B) \tag{32}$$

Bayes' rule

Combining:

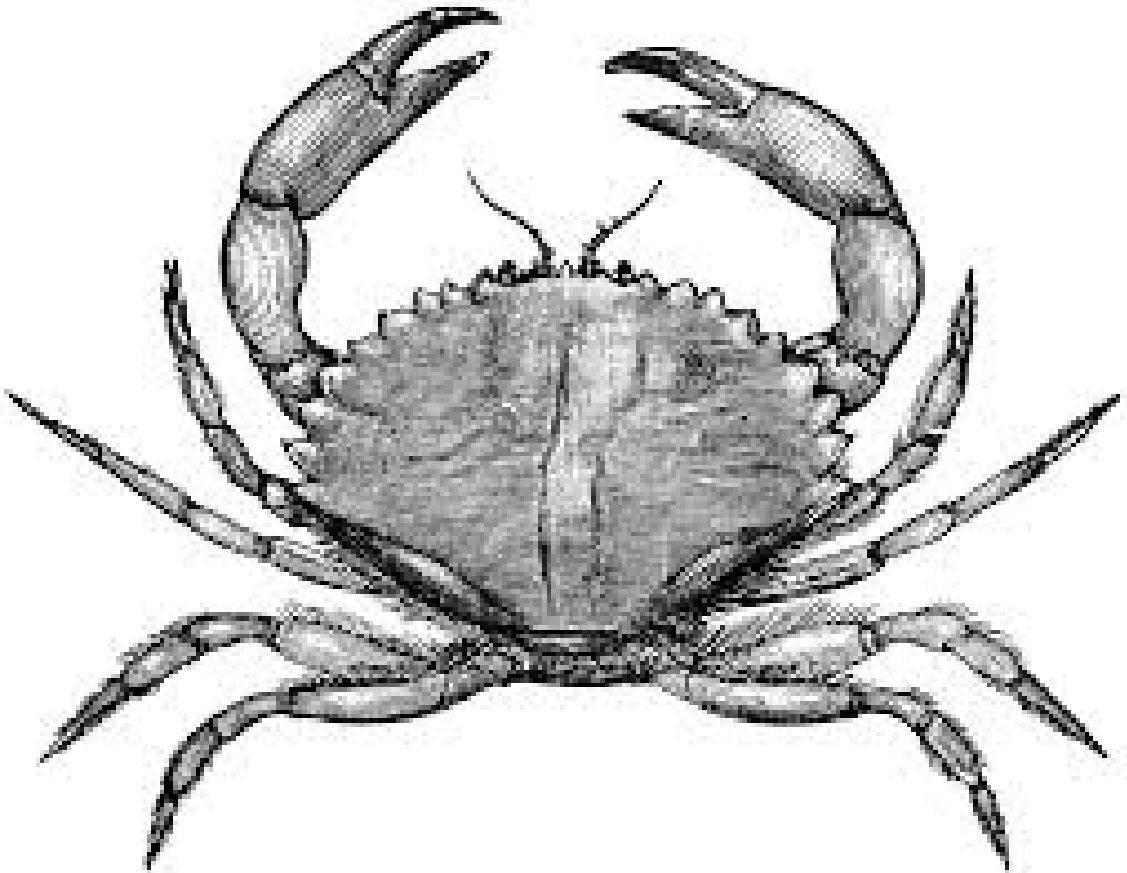
$$\mathbb{P}(A) \times \mathbb{P}(B|A) = \mathbb{P}(B) \times \mathbb{P}(A|B) \tag{33}$$

which we can rearrange to:

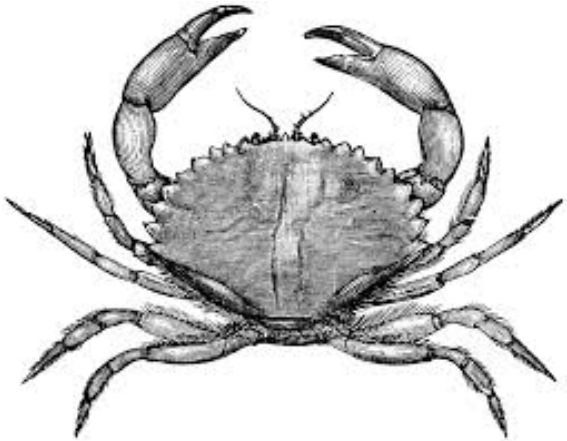
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A) \times \mathbb{P}(B|A)}{\mathbb{P}(B)} \tag{34}$$

this is known as *Bayes' rule* and is the foundation of *Bayesian inference*.

Example use of Bayes' rule: breast cancer screening



Screening probabilities



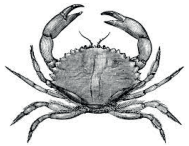
Suppose:

- The probability that a randomly chosen 40 year old woman has breast cancer is approximately $\frac{1}{100}$.
- If a woman has breast cancer the probability they will test positive in a mammography is about 90%.
- However there is a risk of about 8% of a false positive result of the test.

Question: given that a woman tests positive, what is the probability that she has breast cancer?

Bayes' rule in action: breast cancer screening

Answer: we want to find the probability the woman has cancer *given* she has tested positive, which we can do via Bayes' rule (it's the same for pmfs as it was for pdfs):

$$\Pr(\text{  } \mid +) = \frac{\Pr(+ \mid \text{  }) \times \Pr(\text{  })}{\Pr(+)}$$

Bayes' rule in action: breast cancer screening

$$\Pr(\text{crab} \mid +) = \frac{\overbrace{\Pr(+ \mid \text{crab})}^{0.9} \times \overbrace{\Pr(\text{crab})}^{0.01}}{\underbrace{\Pr(+)}_{?}}$$

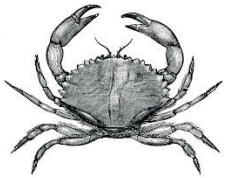
Calculate denominator using the law of total probability:

$$\Pr(+) = \underbrace{\Pr(+ \mid \text{crab}) \times \Pr(\text{crab})}_{0.9 \times 0.01} + \underbrace{\Pr(+ \mid \text{crab}) \times \Pr(\text{crab})}_{0.08 \times 0.99}$$

$$\approx 0.09$$

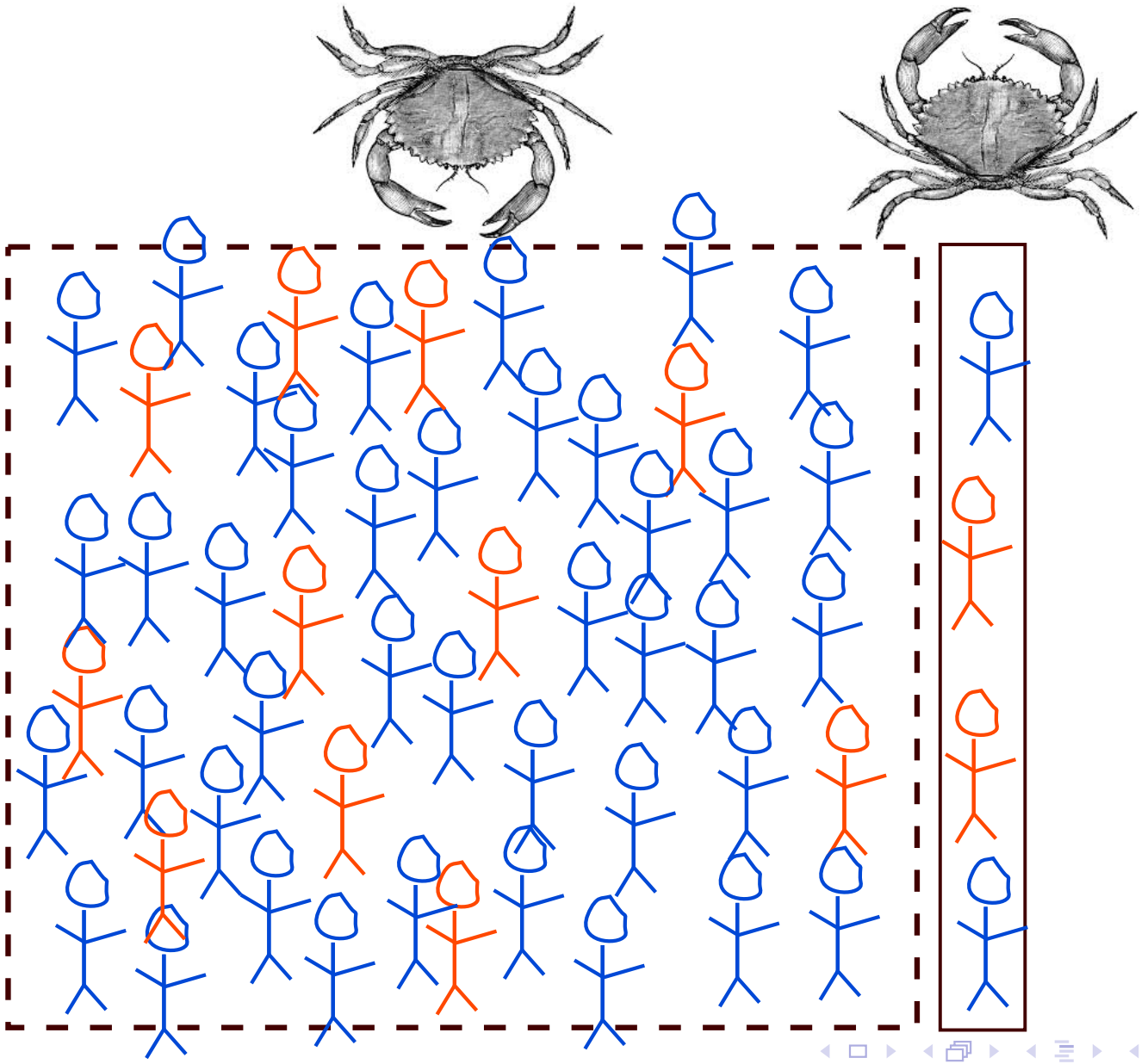
Bayes' rule in action: breast cancer screening

Putting this into Bayes' rule:

$$\Pr(\text{  } \mid +) = \frac{0.9 \times 0.01}{0.09}$$

$$\approx 0.1$$

Intuition: false positives dwarf true ones



Questions?

1 What is probability and why do we need it?

2 Probability and counting

3 Conditional probability

4 Bayes' rule

5 Random variables and probability distributions

6 Expectations

7 Joint distributions

8 Continuous probability distributions

Events and random variables

Up until now, we have used *events* to describe outcomes.

This notation is clunky and makes things unnecessarily tricky.

Random variable notation makes things easier. Definition:
"A random variable associates outcomes of an experiment with a number."

What is a random variable?

Insert pebble world with pebbles annotated with numbers. Mention that we typically use capital letters for rvs.

Example: flipping two coins

Suppose you flip two coins. There are four possible outcomes: $\{H, H\}, \{H, T\}, \{T, H\}, \{T, T\}$.

We create a random variable X so that it equals the number of heads. So, here:

- $\{T, T\} \to X = 0$
- $\{H, T\}$ and $\{T, H\} \to X = 1$.
- $\{H, H\} \to X = 2$

Probability distributions

There are two types of random variables:

- Discrete: e.g. $X = 0, 1, 2, 3, ...$
- Continuous: e.g. $X = 0.545, -4.124, 100.123.$

For now, we only consider discrete random variables.

A *probability distribution* specifies the probabilities of all events for a random variable.

Coin flip probability distribution

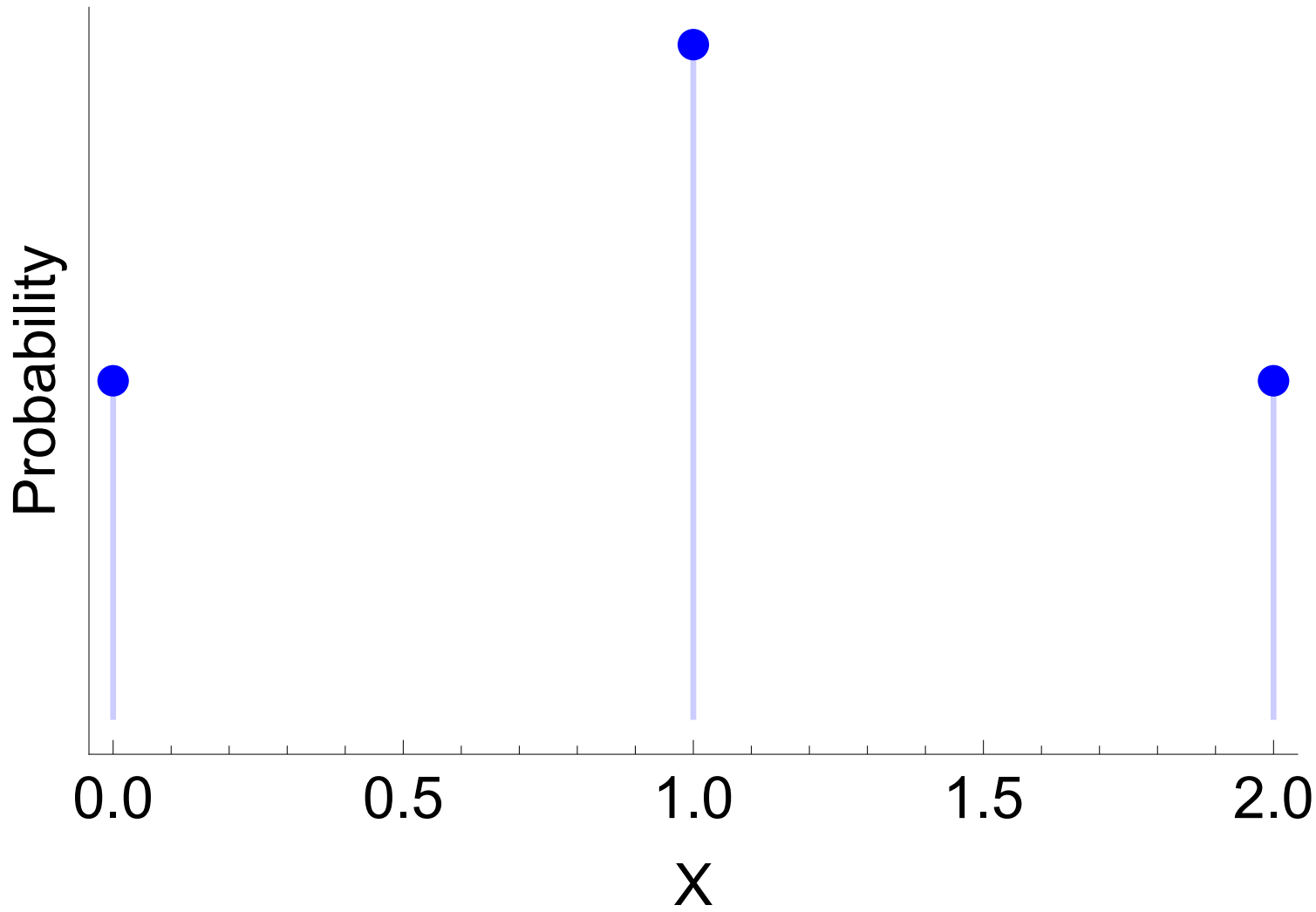
Since this random variable X is discrete, our probability distribution is known as a *probability mass function* or *p.m.f.*. How do we work out what it is?

Let's suppose the coin is fair, so that $Pr(H) = 0.5$. In that case, we can determine the probability of all possible values of X:

- $\{T, T\} \rightarrow X = 0$: $Pr(X = 0) = 0.5^2 = 0.25$.
- $\{H, T\}$ and $\{T, H\} \rightarrow X = 1$: $Pr(X = 1) = 2 \times 0.5^2 = 0.5$
- $\{H, H\} \rightarrow X = 2$: $Pr(X = 2) = 0.5^2 = 0.25$.

Important: note that there are *twice* as many ways to get $X = 1$.

Visualising the p.m.f.



Conditions for a valid p.m.f.

So, we have a *p.m.f.* $Pr(X = x)$ where x denotes the specific values that the random variable takes. How can I tell that it's valid:

- $Pr(X = x) \geq 0$ for all feasible x values: this just ensures we can't have a negative probability.
- $\sum_x Pr(X = x) = 1$: this just means that the outcome of the experiment must be in the set of feasible x values.

Are these satisfied for our coin example: all probabilities are non-negative, so ok there. In terms of the sum:

$$Pr(X = 0) + Pr(X = 1) + Pr(X = 2) = 0.25 + 0.5 + 0.25 = 1. \quad (35)$$

The binomial distribution

The coin flipping case is an example of a particular distribution known as the *binomial*. This sort of distribution rises when you have:

- A fixed number of *trials* (e.g. flips) with a binary outcome: e.g. heads or tails
- The outcome of a given trial does not depend on the previous outcome. I.e. *independence*

Like all probability distributions, the binomial can be used in many different situations. For example: COVID-19 testing, drug resistance and so on.

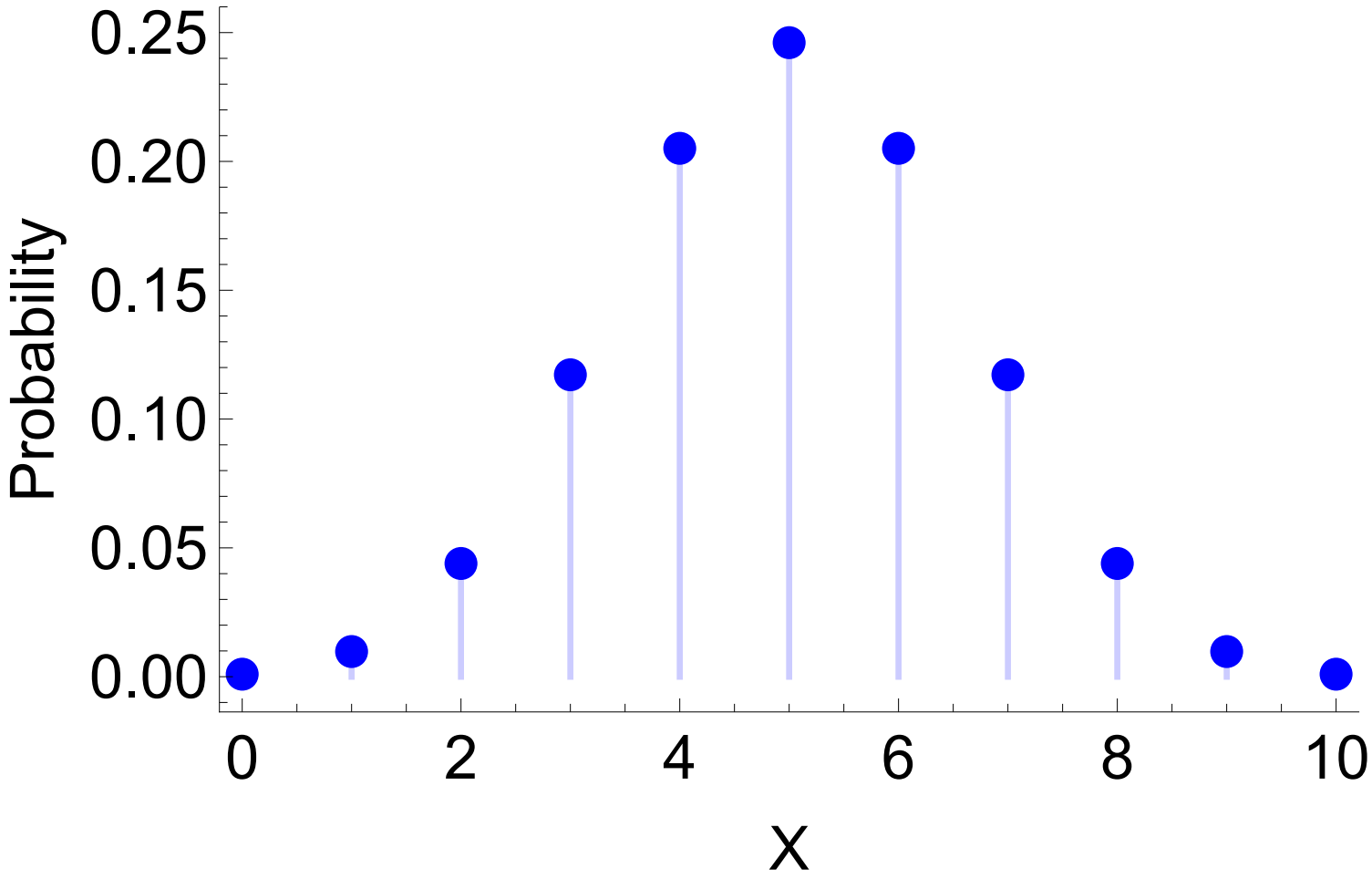
Many coin flips

Suppose now we flip our fair coin 10 times: the coin can now land heads up $X = 0, 1, 2, ..., 9, 10$ times. What is the probability of each outcome?

$$Pr(X = x) = \binom{10}{x} 0.5^x (1 - 0.5)^{10-x} \tag{36}$$

- $\binom{10}{x}$ accounts for the number of ways to achieve a particular number of heads x. E.g there is 1 way to obtain all heads and 252 ways to obtain 5 heads
- 0.5^x multiplies individual probabilities of heads (assumes independence)
- $(1 - 0.5)^{10-x}$ multiplies individual probabilities of tails (assumes independence)

Many coin flips: visualised



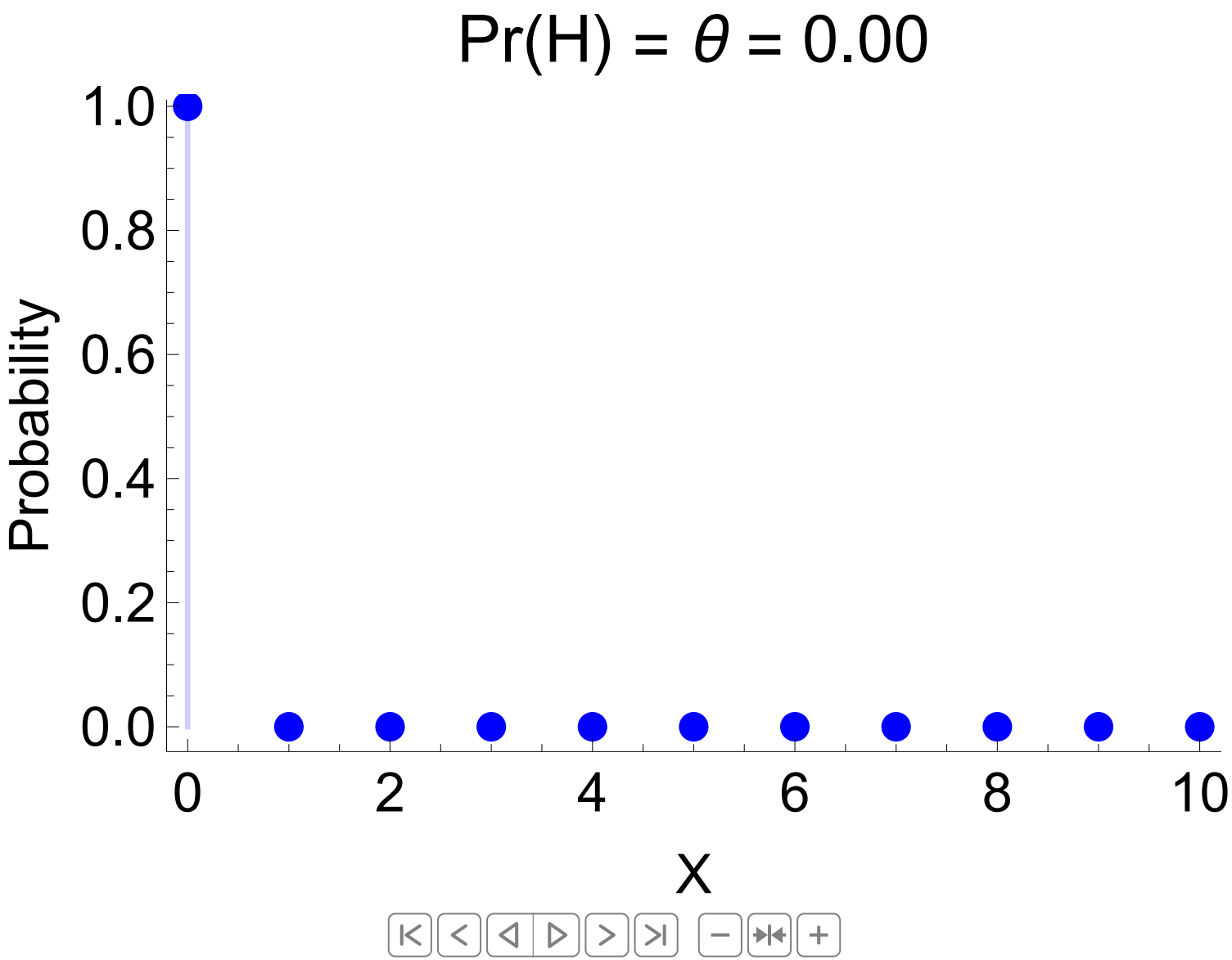
A biased coin

Suppose instead of our fair coin, we have a coin with an arbitrary probability of heads: $Pr(H) = \theta \in [0, 1]$. Probability distribution becomes:

$$Pr(X = x) = \binom{10}{x} \theta^x (1 - \theta)^{10-x} \tag{37}$$

How does changing θ change how the probability distribution looks?

A biased coin: visualised



Poisson distribution

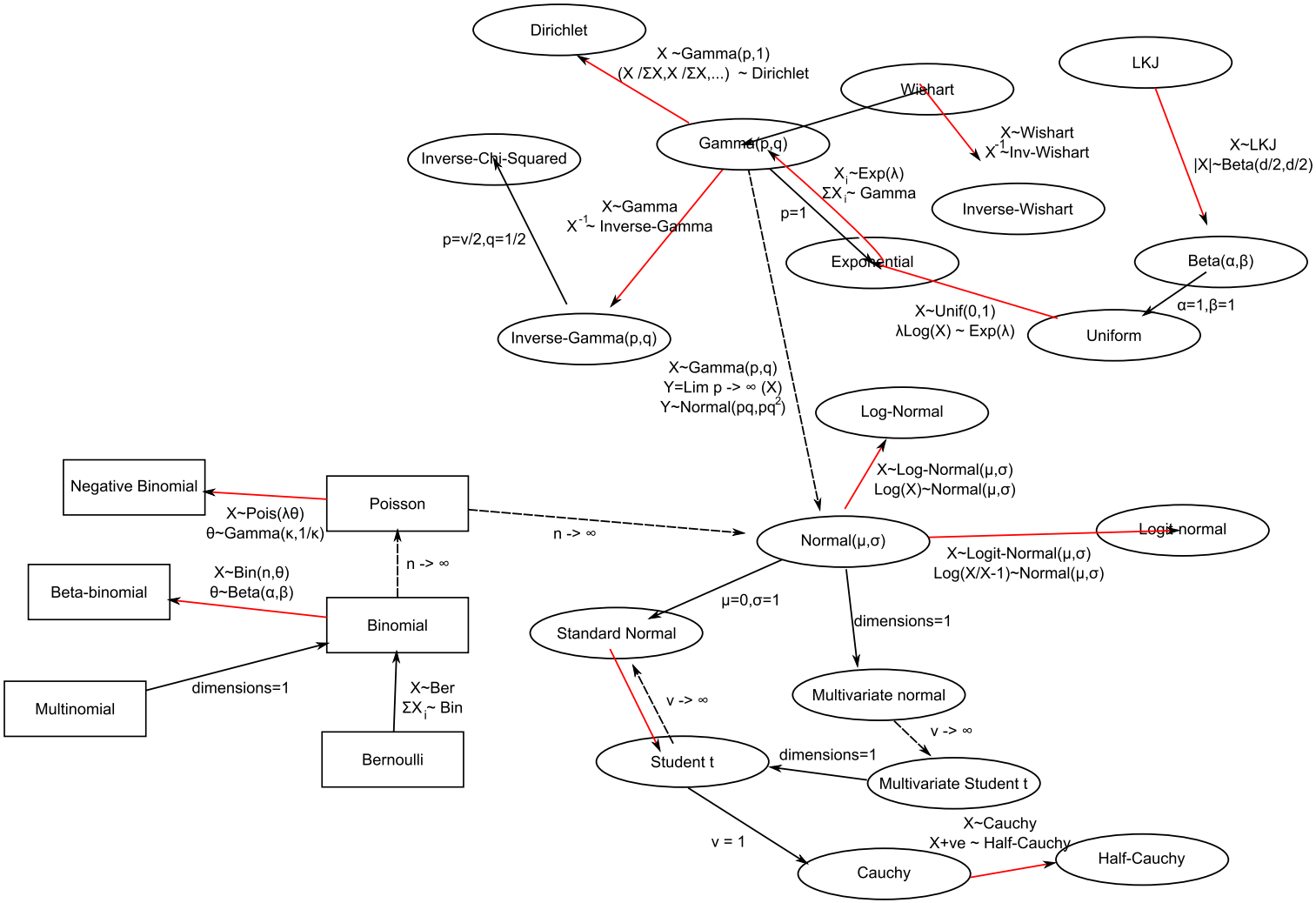
There are many other discrete distributions: each of these is appropriate under a different set of circumstances. One such distribution is the *Poisson*. This distribution is used to count the occurrence of something over a fixed interval in time or space.

For example, this distribution could be used to model outcome of counting cars passing in 15 minute intervals.

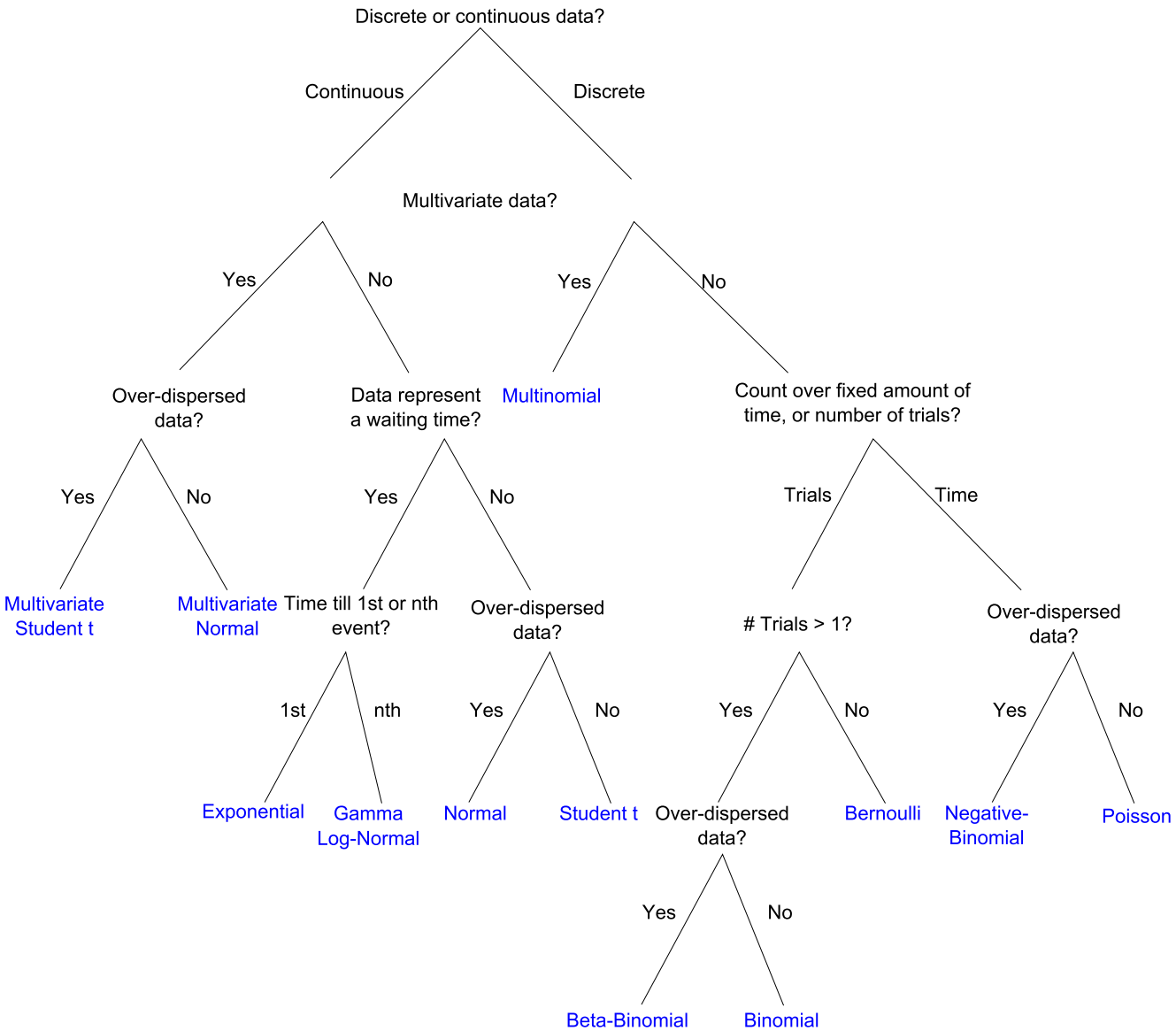
The assumptions underpinning this distribution is:

- Outcome can be $0, 1, 2, 3, 4, ..., \infty$
- Each individual event is not influenced by the occurrence of other events.

Lots of probability distributions



How to choose one



1 What is probability and why do we need it?

2 Probability and counting

3 Conditional probability

4 Bayes' rule

5 Random variables and probability distributions

6 Expectations

7 Joint distributions

8 Continuous probability distributions

1 What is probability and why do we need it?

2 Probability and counting

3 Conditional probability

4 Bayes' rule

5 Random variables and probability distributions

6 Expectations

7 Joint distributions

8 Continuous probability distributions

1 What is probability and why do we need it?

2 Probability and counting

3 Conditional probability

4 Bayes' rule

5 Random variables and probability distributions

6 Expectations

7 Joint distributions

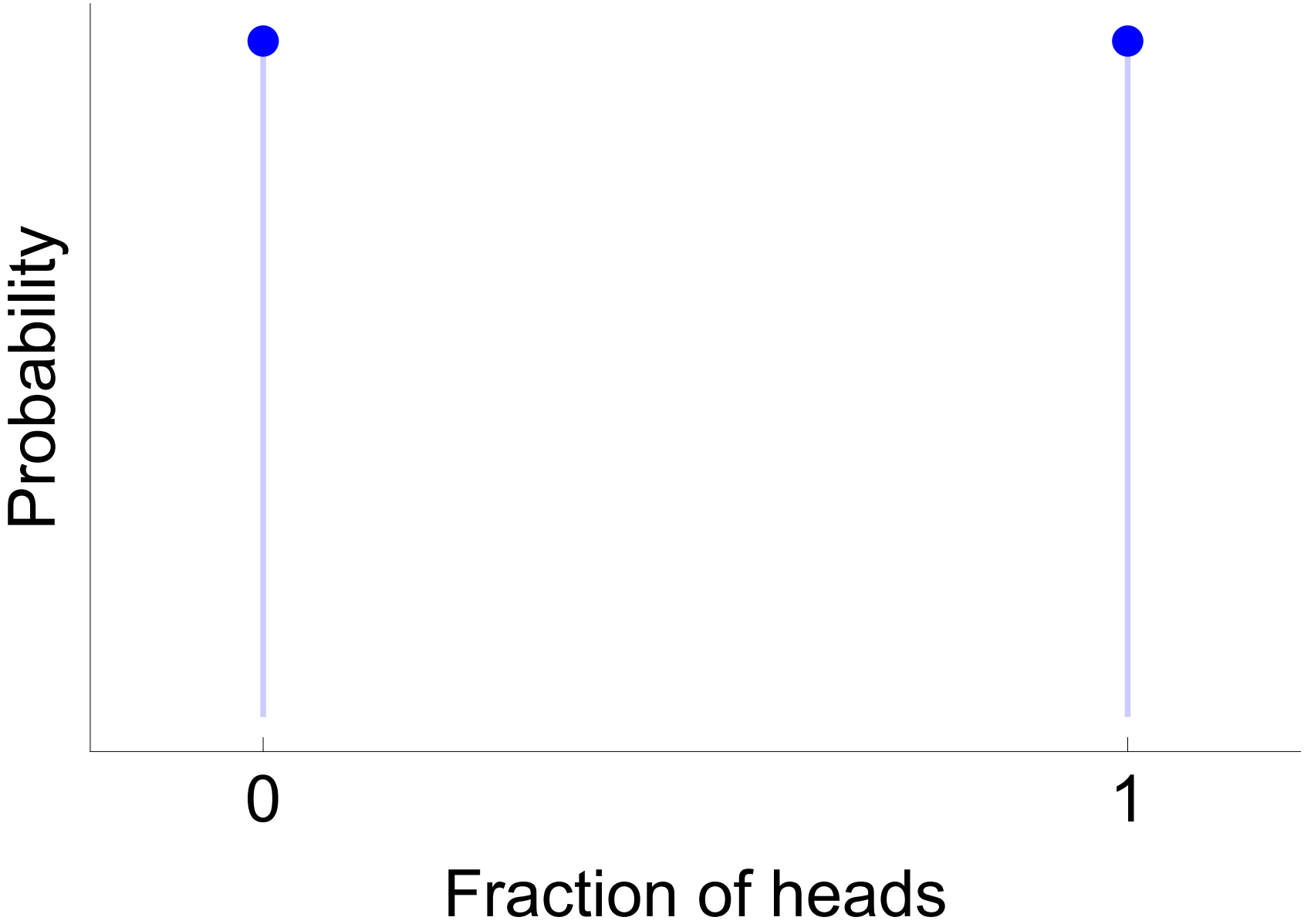
8 Continuous probability distributions

Isn't a normal distribution arbitrary?

No! Why? The central limit theorem.

Suppose we flip a fair coin once. What does its probability distribution look like?

One coin flip

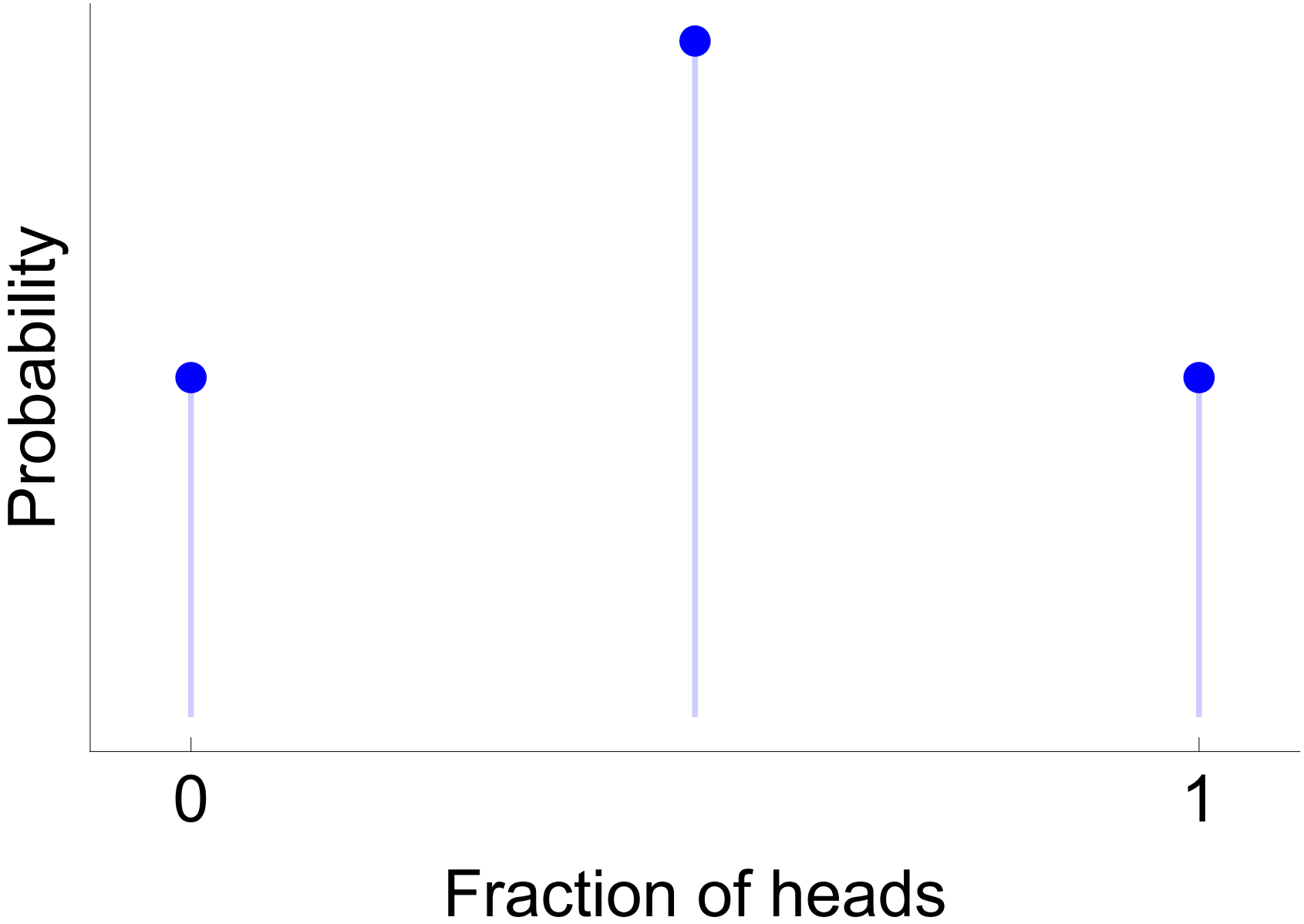


Two flips

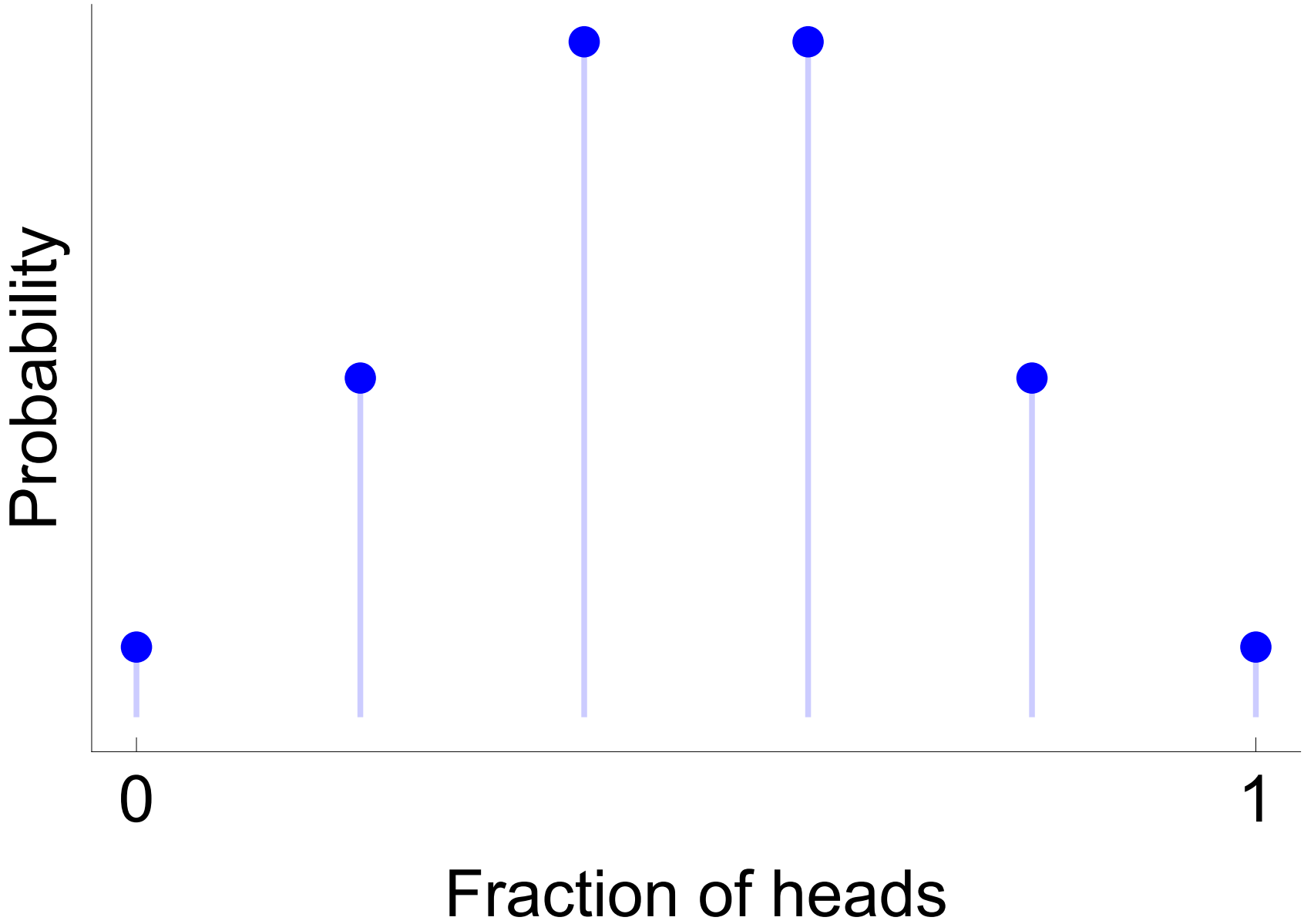
I now flip the coin 2 times.

Question: What does the distribution look like now?

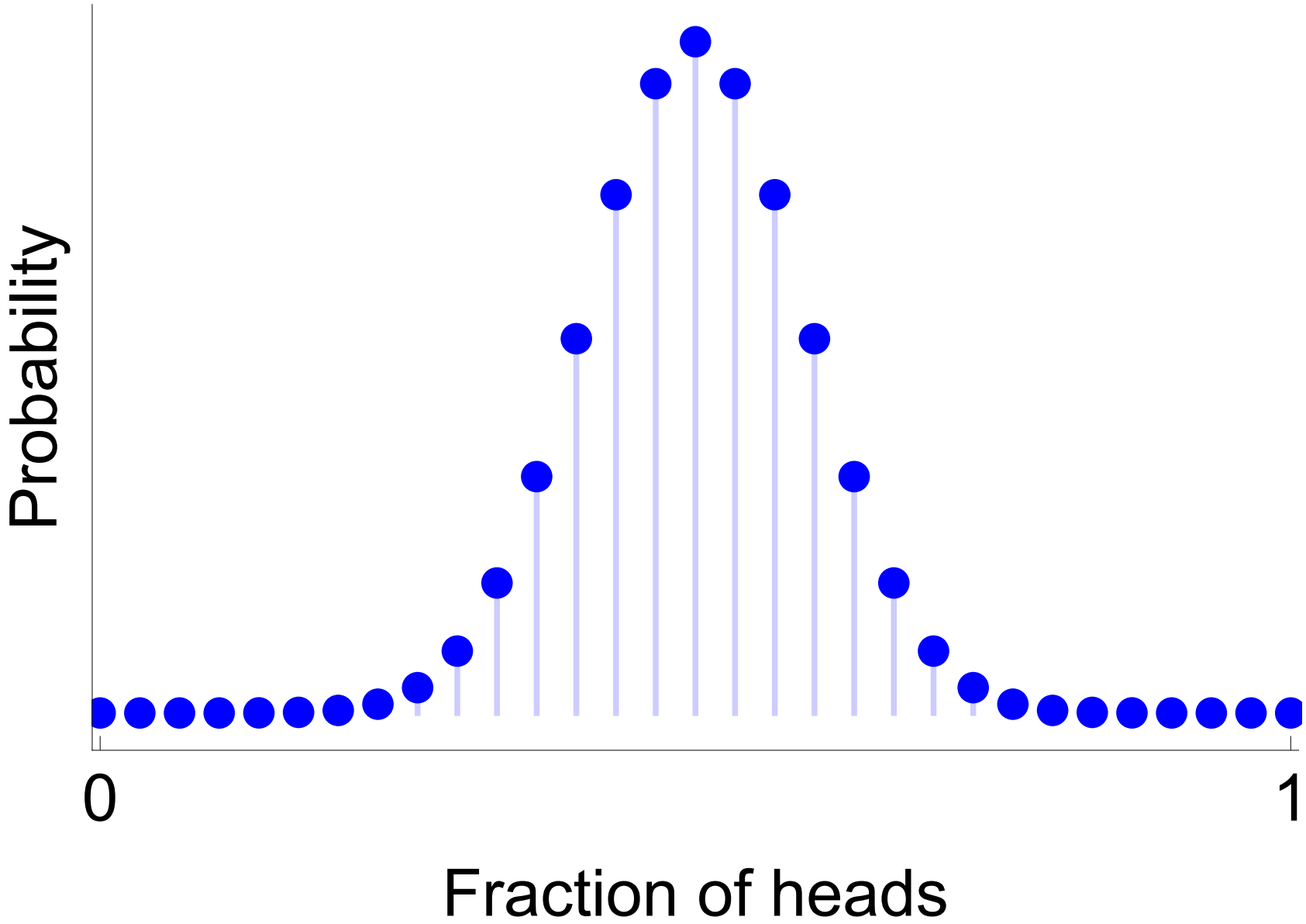
Two flips flips



What about five flips?



What about 30 flips?



What's going on?

The Central Limit Theorem (CLT) says that under general conditions:

"The distribution of the average of a large number of weakly dependent random variables is approximately normal."

In the coin flipping case, we effectively calculated the average number of independent coin flips landing heads up $\implies$ CLT applies.