Last time we saw that in many cases logic helps to make a good decision. It helped us many times with the Wumpus puzzle. Logical thinking also helps to solve the puzzles from Simon Tathems collection or `https://www.brainbashers.com`. However, sometimes logic is not enough like we mentioned last time. There are cases where the future cannot be inferred/calculated. But even then we have to find a way to make a decision. Were bringing another branch of math to facilitate this: probability theory.

This may not sound too good at first. But the success of artificial intelligence from the 1990s is based on probability. The meaningful search about a mistyped query, voice-controlled mobile phones, smart speakers, smartclocks, the translation of subtitles into a foreign language use probabilistic inference. All of these are based on probability theory, so at first, we will look at some definitions and theorems.

One-third of the AIMA book fits into this semester. The parts with more complex maths are just touched on, giving an insight into whats under the hood. For those interested in the topic, I suggest you take a look at the remaining chapters of the book!
Today's schedule is as follows:

- First, we address the concept of uncertainty: why it is needed and how can we deal with it.
- Then, as one such tool, we present the probability theory.
- This can also be thought of as some formal language, so it has a formalism (syntax) and has an underlying meaning (semantics).
- As we have already indicated, for us the goal is probabilistic inference, so we also deal with the concept of consequence.
- Next time we will talk about Bayesian networks, to substantiate this we will look at the concept of independence (conditional independence) of probability variables and the related Bayesian rule.

Uncertainty

- Let action $A_t$ = leave for airport $t$ minutes before flight.
- Will $A_t$ get me there on time?
- Problems:
  1. partial observability (road state, other drivers' plans, etc.)
  2. noisy sensors (KCBS traffic reports)
  3. uncertainty in action outcomes (flat tire, etc.)
  4. immense complexity of modelling and predicting traffic
- Hence a purely logical approach either
  1. risks falsehood: "$A_{25}$ will get me there on time" or
  2. leads to conclusions that are too weak for decision making: "$A_{25}$ will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."
- ($A_{1440}$ might reasonably be said to get me there on time but I'd have to stay overnight in the airport ...)

Lets see a real life example. You want to go home by plane from the Debrecen Airport. How much earlier do you have to leave? This depends on your travel method: normal bus, airport shuttle or by taxi, and it depend on time: early morning or rush hour. If the radio or an application that provides you with live road updates indicates that an accident has occurred on the road, your personal experience cannot be reliable used, although an alternative route may help. If the accident happens when you are already on the road, we may be in more trouble. Similarly, the same can happen if a friend offers to drive us, but then it turns out that his car wont start, or it has a flat tire. Then again, we may miss the plane. The problem is that transport has a lot of participants, whose goals we dont know, and as such we cant anticipate them.

Uncertainty
- Let action $A_t$ = leave for airport $t$ minutes before flight.
- Will $A_t$ get me there on time?
- Problems:
  1. partial observability (road state, other drivers' plans, etc.)
  2. noisy sensors (KCBS traffic reports)
  3. uncertainty in action outcomes (flat tire, etc.)
  4. immense complexity of modelling and predicting traffic
- Hence a purely logical approach either
  1. risks falsehood: "$A_{25}$ will get me there on time" or
  2. leads to conclusions that are too weak for decision making: "$A_{25}$ will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."
- ($A_{1440}$ might reasonably be said to get me there on time but I'd have to stay overnight in the airport ...)

How to deal with this?

- According to the blindfold solution, we do not anticipate any accidental events and start at the planned time.
- The other takes into account all the conditions that occur, but several hypothesis are not known, so this approach cannot be used in practice.

To be sure of catching the flight we may need to start much earlier, but who wants to spend their time at the airport?

Methods for handling uncertainty

**Default or nonmonotonic logic:**

- Assume my car does not have a flat tire
- Assume $A_{25}$ works unless contradicted by evidence

Issues: What assumptions are reasonable? How to handle contradiction?

**Rules with fudge factors**

- $A_{25} \mapsto_{0.3} AtAirportOnTime$
- $Sprinkler \mapsto_{0.99} WetGrass$
- $WetGrass \mapsto_{0.7} Rain$

Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*?

**Probability**

Given the available evidence, $A_{25}$ will get me there on time with probability 0.04

- Mahaviracarya (9th C.), Cardamo (1565) theory of gambling

What tools can we use to work with these uncertainties, join them, and draw conclusions?

The first is a variant of the logic you know. Here we use our our assumptions in case of missing information. For example, if we know that *Tux is a bird*, we are almost certain that *Tux can fly* since it is a bird. (The basic assumption is that *every bird can fly*.) Then if we find out that *this bird is a penguin*, we can already say that *it cant fly* because *penguins dont fly*. Similarly in the example above we assume that a tire of a car cannot be flat by itself. There are many, many questions that may arise when using this logic. What we take as a rule, if – as before – a contradiction arises, which rule do we consider to be stronger, and so on.

Methods for handling uncertainty
Default or nonmonotonic logic:
- Assume my car does not have a flat tire
- Assume $A_{25}$ works unless contradicted by evidence

Issues: What assumptions are reasonable? How to handle contradiction?

Rules with fudge factors
- $A_{25} \mapsto_{0.3} AtAirportOnTime$
- $Sprinkler \mapsto_{0.99} WetGrass$
- $WetGrass \mapsto_{0.7} Rain$

Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*?

Probability
Given the available evidence, $A_{25}$ will get me there on time with probability 0.04
- Mahaviracarya (9th C.), Cardamo (1565) theory of gambling

Alternatively, we can use frequencies based on experience. For example, if for the last ten times we left the our flat 25 minutes before the gate closed and caught the plane only three times, then a 30 percent success rate can be considered a factor of 0.3. It is also possible to put together such factors, we just need to know when and what can be combined. If we look at the results of the morning waterings around the faculty building, ten minutes later we can almost always see its results. If there is a crazy drought, the soil may absorb all the water in a matter of seconds, hence the 99 percent value here. On the other hand, if we look at why the grass is wet, – by creating a statistics on it, we get that rain is the reason 70 percent of the time. (If not necessary, why sprinkle?)

# AI #9

## Methods for handling uncertainty

Methods for handling uncertainty

**Default or nonmonotonic logic:**

- Assume my car does not have a flat tire
- Assume $A_{25}$ works unless contradicted by evidence

Issues: What assumptions are reasonable? How to handle contradiction?

**Rules with fudge factors**

- $A_{25} \mapsto_{0.3} AtAirportOnTime$
- $Sprinkler \mapsto_{0.99} WetGrass$
- $WetGrass \mapsto_{0.7} Rain$

Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*?

**Probability**

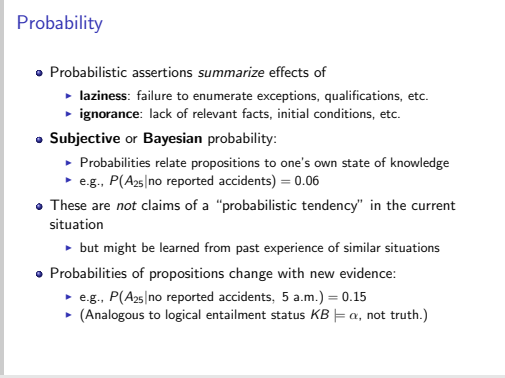
Given the available evidence, $A_{25}$ will get me there on time with probability 0.04

- Mahaviracarya (9th C.), Cardamo (1565) theory of gambling

If we consider the relationships ($A \supset B$, $B \supset C$) as implications, then it arises that we connect them here as well, similarly to the rule of resolution? Because in our case, it gives a very strange result (which the garden owners swear by the way), if we water, it will rain. But how often should we assign it? (We could similarly replace watering with window cleaning!)A Probability calculation does not give a frequency, but a chance, and when we perform enough tests, change gets close to the frequency. Probability arose from desire of winning in gambling. But its history can be traced back more than a thousand years.

Fuzzy logic
- handles *degree of truth*
- NOT uncertainty
- e.g. *WetGrass* is true to degree 0.2

Moreover, there is another logical approach that brought us the intelligent washing machines, intelligent ventilation, subway assembly control as early as the 1980s. Here, we assign a number between 0 and 1 to each atomic formula, but this does not mean frequency, but the degree of truth. If, for example, the question is: what is the truth level of the fact that a 170 cm high man is tall, we can answer that with 0.3.




Let's get back to the probability! We use probability when we do not have enough information, or it is too difficult or too expensive to obtain it, or there is so much information that it is costly to process it all. It is very common that we only look at a sample (political opinion, preference on products, etc.) and not the whole population. If the sample is representative, the results obtained are very close to reality (the same test on the whole population).

There are facts and there is chance. In case there is an unknown, we are talking about chance. A card drawn from a regular deck of cards may be an ace of spades or may be something else. If we accept that all cards can be drawn with the same chance, then - before we look at the card - we give a 1/52 chance that this card will be the ace of spades. When will our opinion change? If we turn the card over and we learn what it is. By then, it will be clear whether that card is an ace of spades, or not. This is no longer a chance, but a fact.

Probability

- Probabilistic assertions *summarize* effects of
  - **laziness**: failure to enumerate exceptions, qualifications, etc.
  - **ignorance**: lack of relevant facts, initial conditions, etc.
- **Subjective** or **Bayesian** probability:
  - Probabilities relate propositions to one's own state of knowledge
  - e.g., $P(A_{25}|\text{no reported accidents}) = 0.06$
- These are *not* claims of a "probabilistic tendency" in the current situation
  - but might be learned from past experience of similar situations
- Probabilities of propositions change with new evidence:
  - e.g., $P(A_{25}|\text{no reported accidents, 5 a.m.}) = 0.15$
  - (Analogous to logical entailment status $KB \models \alpha$, not truth.)

Accordingly, the idea/probability of our idea and belief gained from experience is the preliminary/a priori probability. If certain facts are already given and we are talking about the probabilities associated with them, these will be conditional probabilities.

Ideally, we can use the results provided by mathematics, but often we dont have the right tools to calculate the odds, in which case we give them ourselves and possibly modify them from time to time if they are very different from what we experienced.

Making decisions under uncertainty

*Suppose I believe the following:*

$$
\begin{aligned}
P(A_{25} \text{ gets me there on time} | \ldots) &= 0.04 \\
P(A_{90} \text{ gets me there on time} | \ldots) &= 0.70 \\
P(A_{120} \text{ gets me there on time} | \ldots) &= 0.95 \\
P(A_{1440} \text{ gets me there on time} | \ldots) &= 0.9999
\end{aligned}
$$

*Which action to choose?*

*Depends on my* **preferences** *for missing flight vs. airport cuisine, etc.*

**Utility theory** *is used to represent and infer preferences*

**Decision theory** = *utility theory + probability theory*

We can assign a chance, a probability to reach the plane for each time period. However, the question of when to start is still open! The answer depend on the people. It is likely that we will choose a different value for the plan which has no alternatives than when we have a connecting flight as well. Additionally, deciding on the latter will depends on how scared we are of a delayed plane.

Such decisions occur in economic life every day, so the appropriate discipline has also developed. Utility theory explores how the alternatives can be ranked and how this ranking can be used later. We can add chance to the alternatives, which in turn gives decision theory. Here, too, several issues arise, e.g. the issue of successive decisions.

If anyone is interested in this topic, then Chapters 16 and 17 of the AIMA book discuss this in more depth.

Probability basics

- Begin with a set $\Omega$—the **sample space**
  - e.g., 6 possible rolls of a die.
  - $\omega \in \Omega$ is a **sample point**/**possible world**/**atomic event**
- A **probability space** or **probability model** is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ s.t.
  - $0 \leq P(\omega) \leq 1$
  - $\sum_{\omega} P(\omega) = 1$
- e.g., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.
- An **event** $A$ is any subset of $\Omega$

$$P(A) = \sum_{\{\omega \in A\}} P(\omega)$$

- E.g., $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

Let's refresh our knowledge! The first concept will be the atomic event, which are events that are mutually exclusive. The set of these atomic events give us the sample space. If we roll three dice, the atomic events could be their sum, as 3, 4, ...., 18. We have to assign a probability (chance) to each of these options. Calculating the probabilities is not so simple. Therefore choose different atomic events: let us distinguish the three dice their results separately is an atomic event. Then everything becomes very simple, the probabilities of atomic events will be the same. On the other hand, if the question is, what is the chance that we will roll 10 with the three dice, then considering it as one event, and we get its probability by summing the probabilities of all the atomic events that occur in it. Often it is up to us to decide which are the atomic events.

Random variables

- A **random variable** is a function from sample points to some range, e.g., the reals or Booleans
  - e.g., $Odd(1) = true$.
- $P$ induces a **probability distribution** for any r.v. $X$:

$$P(X = x_i) = \sum_{\omega : X(\omega) = x_i} P(\omega)$$

- e.g., $P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$

Let us take a function that assigns a value to elementary events. Later, to exclude the lengthy calculations we usually use logical values. In these cases these functions are characteristic functions. The function odd is one such function. It assigns true and false values to the dice roll results (1-6), assigning true to atomic events 1, 3 and 5.

What is the chance of throwing an even number? We need to take atomic values $\omega$ for which $Even(\omega)$ is true, and sum the corresponding probabilities. For a regular dice, the chance of each number are $1/6$, and since we have 3 good atomic events, $1/2$ is the result. Similarly, there are only two numbers greater than four, so the chance of throwing a number greater than four is $1/3$.

Propositions

- Think of a proposition as the event (set of sample points) where the proposition is true
- Given Boolean random variables $A$ and $B$:
  - event $a$ = set of sample points where $A(\omega) = true$
  - event $\neg a$ = set of sample points where $A(\omega) = false$
  - event $a \wedge b$ = points where $A(\omega) = true$ and $B(\omega) = true$
- Often in AI applications, the sample points are *defined* by the values of a set of random variables, i.e., the sample space is the Cartesian product of the ranges of the variables
- With Boolean variables, sample point = propositional logic model
  - e.g., $A = true$, $B = false$, or $a \wedge \neg b$.
- Proposition = disjunction of atomic events in which it is true
  - e.g., $(a \vee b) = (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$
  - $\Longrightarrow P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$

For logical variables, the proposition coincides with the event. This is no longer true for real-value functions (random variables). The fact that the temperature is pleasant (20-25°C), is a statement/proposition that can include several events.

Events/propositions can be connected with logical connectives you know.

If we have several independent variables (tossing a dice and a coin), the atomic event is given by a pair of values, e.g. $(3, head)$, i.e., the Cartesian product of the sets of outcomes should be considered.

If we have logical variables only, then each combination corresponds to an atomic event. But if we assign a value to each variable, that is we set what is true and what is not, then we are giving an interpretation.

If a proposition is a set of multiple atomic events (which are considered a conjunction), a proposition is very similar to DNF.

# AI #9

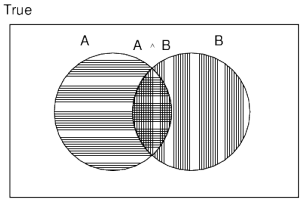
## Why use probability?

So far we looked at the approach from a mathematics point of view, now lets take logic, and some calculus. If we desire at least one of two events to be fulfilled, it is a disjunction. This disjunction is inclusive, it enables both its arguments to be true. If we use probabilities, this common part is used twice, so we need to subtract its probability from the sum of probabilities.

Why use probability?

The definitions imply that certain logically related events must have related probabilities

E.g., $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$




de Finetti (1931): an agent who bets according to probabilities that violate these axioms can be forced to bet so as to lose money regardless of outcome.

Syntax for propositions

**Propositional or Boolean random variables**
- e.g., *Cavity* (do I have a cavity?)
- *Cavity* = *true* is a proposition, also written *cavity*

**Discrete random variables (*finite* or *infinite*)**
- e.g., *Weather* is one of ⟨*sunny*, *rain*, *cloudy*, *snow*⟩
- *Weather* = *rain* is a proposition
- Values must be exhaustive and mutually exclusive

**Continuous random variables (*bounded* or *unbounded*)**
- e.g., *Temp* = 21.6; also allow, e.g., *Temp* < 22.0.

Arbitrary Boolean combinations of basic propositions

What do these propositions look like? Although we have seen an example, let's look at them systematically! *Do I have a cavity?* This question essentially denotes a random variable called *Cavity*. If the value of this variable is true (*Cavity* = *true*), we get a proposition; which is either fulfilled or not.

If for weather we have four options (and snow and sunny together is excluded), then the proposition says the actuals situation, e.g. *weather* = *rain*. For continuous values, we can state that the value of the variable is a specific value or that it falls within a range.

Of course, such statements can be further transformed with the usual Boolean connectives.

Prior probability

- *Prior* or *unconditional* probabilities of propositions
  - e.g., $P(Cavity = true) = 0.1$ and $P(Weather = sunny) = 0.72$
  - correspond to belief prior to arrival of any (new) evidence
- *Probability distribution* gives values for all possible assignments $\mathcal{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (*normalized*, i.e., sums to 1)
- *Joint probability distribution* for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)
  - $\mathcal{P}(Weather, Cavity)$ = a $4 \times 2$ matrix of values:

| *Weather* = | *sunny* | *rain* | *cloudy* | *snow* |
|---|---|---|---|---|
| *Cavity* = *true* | 0.144 | 0.02 | 0.016 | 0.02 |
| *Cavity* = *false* | 0.576 | 0.08 | 0.064 | 0.08 |

- *Every question about a domain can be answered by the joint distribution because every event is a sum of sample points*

If we dont know the facts, we can use experiential (or belief-based) value. Dentists may say that every tenth person has a cavity, or based on weather statistics, we get that the chance of a sunny day is 73 percent. With this, we assigned a specific value (chance, probability) to each proposition. This called ad prior or unconditional probability.
If we give one value to each propositions, we get a probability distribution.
Of course, we must keep in mind that the sum/integral of the chances gives 1 (total probability).
If we have several independent propositions, then the atomic events are given by a Cartesian product. Then a chance must be assigned to each atomic event; and the sum will also be 1.

Question on probabilities of any event can be answered by probabilities of its atomic events.

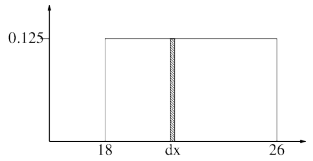
# AI #9

## Probability for continuous variables

Probability for continuous variables

Express distribution as a parameterized function of value:

- $P(X = x) = U[18, 26](x)$ = uniform density between 18 and 26




Here $P$ is a **density**; integrates to 1.
$P(X = 20.5) = 0.125$ really means

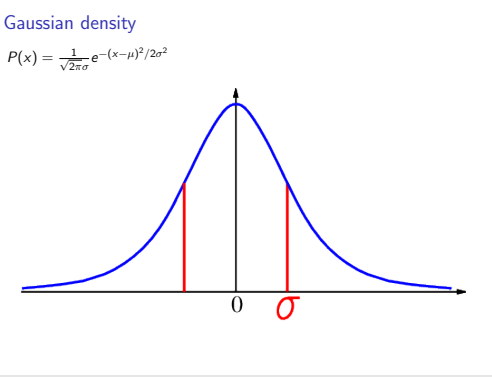
$$\lim_{dx \to 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

Calculating probabilities for discrete events is usually easy. However, in the case of continuous random variables, we need to be careful. You may remember the density and the distribution function. The latter is the definite integrate of the former. The area under the density function will be 1, and therefore the value of the distribution function at the end of the interval is also 1.

In case of the uniform distribution in the example, we will have an 8-wide rectangle. Its height is therefore 1/8. The probability at a specific value is given by the density function. The density function gives the limit of quotient of the probability and the length of the interval, which is now 1/8.

# AI #9

## Gaussian density




In case of phenomena in nature, the normal distribution is common. It has this density function.

Conditional probability

- **Conditional** or **posterior probabilities**
  - e.g., $P(cavity|toothache) = 0.8$
  - i.e., *given that toothache is all I know*
  - *NOT* "if *toothache* then 80% chance of *cavity*"
- Notation for conditional distributions: $\mathcal{P}(Cavity|Toothache) =$ 2-element vector of 2-element vectors
- If we know more, e.g., *cavity* is also given, then we have $P(cavity|toothache, cavity) = 1$
- Note: the less specific belief *remains valid* after more evidence arrives, but is not always *useful*
- New evidence may be irrelevant, allowing simplification, e.g.,
  - $P(cavity|toothache, 49ersWin) = P(cavity|toothache) = 0.8$
  - This kind of inference, sanctioned by domain knowledge, is crucial

Once we know a fact, we cannot rely on our previous beliefs. An excellent example of this is the Monty Hall paradox (`https://hu.wikipedia.org/wiki/Monty_Hall-paradoxon`), I suggest you to have a look and try to understand it!

Returning to the conditional probability, random variables are on both sides of the vertical line. Our (relevant) knowledge – the condition – is after the vertical line.

It is important to address this claim in its place. In case of additional information, the situation will change.

Pay attention to the letter $P$. If a probability variable (capital letter) is included in the expression, it means a conditional distribution, which includes several probabilities. In this case we have calligraphic $P$s, like $\mathcal{P}$. However, if we only have lowercase propositions, it gives a probability/chance.

If we take known facts as a condition, they will surely be fulfilled. Irrelevant propositions ($49ersWin$) are not needed, we can omit them.

# AI #9

## Conditional probability

Conditional probability

- Definition of conditional probability:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0$$

- **Product rule** gives an alternative formulation:
  - $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$
- A general version holds for whole distributions, e.g.
  - $\mathcal{P}(Weather, Cavity) = \mathcal{P}(Weather|Cavity)\mathcal{P}(Cavity)$
- View as a $4 \times 2$ set of equations, *not* matrix mult.
- **Chain rule** is derived by successive application of product rule:
  - $\mathcal{P}(X_1, \ldots, X_n) = \mathcal{P}(X_1, \ldots, X_{n-1})\mathcal{P}(X_n|X_1, \ldots, X_{n-1})\} = \mathcal{P}(X_1, \ldots, X_{n-2})\ \mathcal{P}(X_{n-1}|X_1, \ldots, X_{n-2})\ \mathcal{P}(X_n|X_1, \ldots, X_{n-1}) = \ldots = \prod_{i=1}^{n} \mathcal{P}(X_i|X_1, \ldots, X_{i-1})$

The conditional probability has a formula that can be used if the chance of the condition is not zero. Then we need to divide the common probability with the probability of the condition.

To make our rule more general, we multiply it by the probability of the condition – and this holds even if it is zero – product rule. (These formulae contains propositions, so we have normal $P$s.)

If we write a similar formula for random variables (calligraphic $\mathcal{P}$), it holds for any values of them. Since the variables can have 4 and 2 values respectively, we will get $4 \times 2$ equations.

Applying this product rule over and over again, we can rewrite the combined probability of $n$ variables to the product of $n$ conditional probabilities – chain rule.

# AI #9

## Inference by enumeration

Inference by enumeration

Start with the joint distribution:

| | *toothache* | | $\neg$ *toothache* | |
|---|---|---|---|---|
| | *catch* | $\neg$ *catch* | *catch* | $\neg$ *catch* |
| *cavity* | .108 | .012 | .072 | .008 |
| $\neg$ *cavity* | .016 | .064 | .144 | .576 |

For any proposition $\phi$, sum the atomic events where it is true:

- $P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$

Lets see how probability can be used for inference! We have the dentist example with three boolean random variables. The random variable *Cavity* indicates whether the patient has a perforated tooth. The random variable *Toothache* refers to whether the patient has a toothache. Finally the random variable *Catch* indicates whether the dentist finds a hole with their dental probe or not.

Of course, its interesting if the dentist find a cavity, although there does is none, or in reverse if there is a hole and the dentist does not find it.

There is a probability for each atomic event.

# AI #9

## Inference by enumeration

Inference by enumeration

Start with the joint distribution:

| | *toothache* | | ¬ *toothache* | |
|---|---|---|---|---|
| | *catch* | ¬ *catch* | *catch* | ¬ *catch* |
| *cavity* | .108 | .012 | .072 | .008 |
| ¬ *cavity* | .016 | .064 | .144 | .576 |

For any proposition $\phi$, sum the atomic events where it is true:

- $P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$
- $P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

If we need the probability of a particular proposition, we need to summarize the probabilities associated with it. In the first case, 4 atomic events belong to the proposition *Patient has a toothache*, and we sum their chances.

# AI #9

## Inference by enumeration

Inference by enumeration

Start with the joint distribution:

| | *toothache* | | $\neg$ *toothache* | |
|---|---|---|---|---|
| | *catch* | $\neg$ *catch* | *catch* | $\neg$ *catch* |
| *cavity* | .108 | .012 | .072 | .008 |
| $\neg$ *cavity* | .016 | .064 | .144 | .576 |

For any proposition $\phi$}, sum the atomic events where it is true:

- $P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$
- $P(\mathit{cavity} \vee \mathit{toothache}) =$
  $0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$

When we have a proposition containing disjunction, it has six atomic events. We need to sum the corresponding chances.

# AI #9

## Inference by enumeration

Inference by enumeration

Start with the joint distribution:

| | *toothache* | | ¬ *toothache* | |
|---|---|---|---|---|
| | *catch* | ¬ *catch* | *catch* | ¬ *catch* |
| *cavity* | .108 | .012 | .072 | .008 |
| ¬ *cavity* | .016 | .064 | .144 | .576 |

Can also compute conditional probabilities:

$$P(\neg cavity | toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

If we have conditional probabilities, we need to calculate the numerator and the denominator. The numerator is a joint probability, so we need to sum the corresponding atomic events. The denominator is a known value from a former slide.

Normalization

| | toothache | | ¬ toothache | |
|---|---|---|---|---|
| | catch | ¬ catch | catch | ¬ catch |
| cavity | .108 | .012 | .072 | .008 |
| ¬ cavity | .016 | .064 | .144 | .576 |

Denominator can be viewed as a *normalization constant* $\alpha$

$$\begin{aligned}\mathcal{P}(Cavity|toothache) &= \alpha\,\mathcal{P}(Cavity, toothache)\\ &= \alpha\,[\mathcal{P}(Cavity, toothache, catch) + \mathcal{P}(Cavity, toothache, \neg catch)]\\ &= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle]\\ &= \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle\end{aligned}$$

General idea: compute distribution on query variable by fixing **evidence variables** and summing over **hidden variables**

If were lazy, we can bypass some of the calculations. The denominator will have the same value, so we can denote it by a constant. In fact, to avoid having to write fractions, mark the reciprocal with $\alpha$! Thus, if we calculate a conditional distribution (calligraphic $\mathcal{P}$!), we must multiply the joint distribution by this constant. Since we also have a third variable, we need to sum it up for both values of the *Catch*, so we are essentially adding two vectors. (The vectors come from different values of cavity.) Finally, we need to determine $\alpha$, which will be nothing more than the reciprocal of the sum of the two numbers here ($1/0.2 = 5$) to finally get 1. Then, by performing the multiplication, we get the answer.

# Inference by enumeration, contd.

In the general case, we can classify our variables into three groups:

- There are variables about facts (conditions), their set is denoted by $E$ (as evidence).
- We have the variables in the question (query variables), their distribution function is the question (this was the *Cavity* in the former example).
- Finally, there are all the other variables that werent mentioned, these are the hidden variables – the variable *Catch* before. You need to summarize on them.

If we have $n$ variables and a variable can take $d$ values, then we need to take into account cases with exponential complexity ($d^n$), and we need exponential space to store the partial results.

Inference by enumeration, contd.

Let $\mathbf{X}$ be all the variables. Typically, we want the posterior joint distribution of the **query variables** $\mathbf{Y}$ given specific values $\mathbf{e}$ for the **evidence variables** $\mathbf{E}$
Let the **hidden variables** be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$
Then the required summation of joint entries is done by *summing out* the hidden variables:

$$\mathcal{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) = \alpha\mathcal{P}(\mathbf{Y},\mathbf{E}=\mathbf{e}) = \alpha\sum_{\mathbf{h}}\mathcal{P}(\mathbf{Y},\mathbf{E}=\mathbf{e},\mathbf{H}=\mathbf{h})$$

The terms in the summation are joint entries because $\mathbf{Y}$, $\mathbf{E}$, and $\mathbf{H}$ together exhaust the set of random variables
Obvious problems:

1. Worst-case time complexity $O(d^n)$ where $d$ is the largest arity
2. Space complexity $O(d^n)$ to store the joint distribution
3. How to find the numbers for $O(d^n)$ entries?

Inference by enumeration, contd.

Let $\mathbf{X}$ be all the variables. Typically, we want the posterior joint distribution of the **query variables** $\mathbf{Y}$ given specific values $\mathbf{e}$ for the **evidence variables** $\mathbf{E}$
Let the **hidden variables** be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$
Then the required summation of joint entries is done by *summing out* the hidden variables:

$$\mathcal{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) = \alpha\mathcal{P}(\mathbf{Y},\mathbf{E}=\mathbf{e}) = \alpha\sum_{\mathbf{h}}\mathcal{P}(\mathbf{Y},\mathbf{E}=\mathbf{e},\mathbf{H}=\mathbf{h})$$

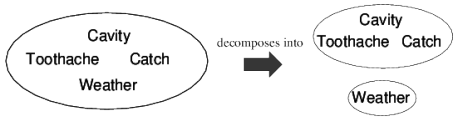
The terms in the summation are joint entries because $\mathbf{Y}$, $\mathbf{E}$, and $\mathbf{H}$ together exhaust the set of random variables
Obvious problems:

1. Worst-case time complexity $O(d^n)$ where $d$ is the largest arity
2. Space complexity $O(d^n)$ to store the joint distribution
3. How to find the numbers for $O(d^n)$ entries?

This doesn't sound too good. If someone has already taken insurance for a car on the internet, in Hungary they had to answer nearly 30 questions, and in each case you had to choose from 5 answers (e.g. Where do you live? Answers: capital, metropolis, town, village, farm).
According to this, you have $5^{30}$ options, which is nice, but it will take years to calculate the value of the insurance!

Insurers would also be in trouble, as they may have no data on accidents committed by a Porse driven by a retired grandmother living in a small village, and there are many other rare cases. How can we get the chances of these options?

Independence
- $A$ and $B$ are **independent** iff
- $\mathcal{P}(A|B) = \mathcal{P}(A)$ or $\mathcal{P}(B|A) = \mathcal{P}(B)$ or $\mathcal{P}(A, B) = \mathcal{P}(A)\mathcal{P}(B)$




- $\mathcal{P}(Toothache, Catch, Cavity, Weather)$
- $= \mathcal{P}(Toothache, Catch, Cavity)\mathcal{P}(Weather)$
- 32 entries reduced to 12; for $n$ independent biased coins, $2^n \rightarrow n$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Fortunately not all the random variables are related to each other, and this can decrease the complexity. If the product rule holds for every value, then the variables are independent.

You can check in a previous chance-table that random variables *Cavity* and *Weather* are really independent.

If we have a richer table with the random variables of weather and dentists, we have 32 atomic events. If we can show the independence between the *Weather* and the others, we get two tables with 12 atomic values.

Moreover if we toss coins (they dont have to be fair), then the joint distribution can be triggered by the individual distributions, i.e., we get linear complexity instead of exponential.

Unfortunately the world is more complicated than this nice mathematical model.

Conditional independence

- $\mathcal{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  1. $P(catch|toothache, cavity) = P(catch|cavity)$
- The same independence holds if I haven't got a cavity:
  1. $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$
- *Catch* is **conditionally independent** of *Toothache* given *Cavity*:
  - $\mathcal{P}(Catch|Toothache, Cavity) = \mathcal{P}(Catch|Cavity)$
- Equivalent statements:
  - $\mathcal{P}(Toothache|Catch, Cavity) = \mathcal{P}(Toothache|Cavity)$
  - $\mathcal{P}(Toothache, Catch|Cavity) = \mathcal{P}(Toothache|Cavity)\mathcal{P}(Catch|Cavity)$

We will use a nicer instrument: conditional independence. The table with the three dental probability variables has 8 fields (slide 18), but since their sum is fixed (=1), we can only change seven of them freely. We can assume that in the case of a cavity the probability of catching the hole is not dependent on a toothache (which is reasonable). We can assume the same in the case of no cavity. If we can leave out a condition from a conditional probability, we say the conditional variable conditionally independent of the missing variable given the other conditions.

Conditional independence contd.

- Write out full joint distribution using chain rule:
  - $\mathcal{P}(Toothache, Catch, Cavity)$
  - $= \mathcal{P}(Toothache|Catch, Cavity)\mathcal{P}(Catch, Cavity)$
  - $= \mathcal{P}(Toothache|Catch, Cavity)\mathcal{P}(Catch|Cavity)\mathcal{P}(Cavity)$
  - $= \mathcal{P}(Toothache|Cavity)\mathcal{P}(Catch|Cavity)\mathcal{P}(Cavity)$
- I.e., $2 + 2 + 1 = 5$ independent numbers (equations 1 and 2 remove 2)
- *In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in $n$ to linear in $n$.*
- *Conditional independence is our most basic and robust form of knowledge about uncertain environments.*

The joined probability of these variables can be written using the chain rule as a triple product, and the first term can be simplified. We have three tables of conditional probabilities with 4, 4 and 4 values, of which 2, 2, and 1 are independent – the others can be calculated from these. Thus, the previous seven independent values are reduced to five. This is not a great result, but it is often possible to reduce the exponential complexity to linear, which allows us calculations in the real life problem.

Furthermore, the more general conditional probabilities are more easier to deal with them, there are already statistics on accident with a sports car, how often villagers crash, and so on.

# AI #9

## Bayes' Rule

Bayes' Rule

Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\Longrightarrow \textbf{Bayes' rule} \quad P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

or in distribution form

$$\mathcal{P}(Y|X) = \frac{\mathcal{P}(X|Y)\mathcal{P}(Y)}{\mathcal{P}(X)} = \alpha\mathcal{P}(X|Y)\mathcal{P}(Y)$$

Useful for assessing **diagnostic** probability from **causal** probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

E.g., let $M$ be meningitis, $S$ be stiff neck:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!

The product rule can be written two ways, and equating them then dividing both sides by a probability we get the Bayes rule for propositions. Similarly, we can get the Bayes rule for probability variables (calligraphic $\mathcal{P}$).
That doesn't sound too exciting when put that way. However, if we take the chances of certain words occurring in our emails and in the spams (e.g. `V1agra`), the new incoming messages can be clustered using these probabilities (Bayesian spam-filter).

In fact, we want to infer the cause (spam) from a characteristic (the presence of a particular word). It can be determined from the empirical conditional probability in the reverse direction and the probability of cause or effect(s).

# AI #9

## Bayes' Rule

Let's take a medical example. If we wake up in the morning with a stiff neck, should we suspect meningitis, or just that we have fallen asleep in the wrong position? The fact is that meningitis is a very common feature of the rigid neck (80 percent). Moreover, a rigid neck is highly likely to occur (10 percent). Which can be reassuring, however, that the chance of meningitis is very low. Substituting all this into the formula also gives a very small probability for the conditional probability.

### Bayes' Rule

Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\Longrightarrow \textbf{Bayes' rule} \quad P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

or in distribution form

$$\mathcal{P}(Y|X) = \frac{\mathcal{P}(X|Y)\mathcal{P}(Y)}{\mathcal{P}(X)} = \alpha\mathcal{P}(X|Y)\mathcal{P}(Y)$$

Useful for assessing **diagnostic** probability from **causal** probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

E.g., let $M$ be meningitis, $S$ be stiff neck:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!

# AI #9

## Bayes' Rule and conditional independence

We can combine the methods. The probability of the cavity in the case of toothache can be calculated from the probability of the cavity and the conditional probabilities with condition Cavity. This gives a linear formula, so the calculation will not be complicated.

Bayes' Rule and conditional independence

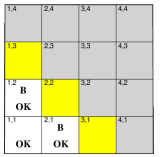
$$\begin{aligned}&\mathcal{P}(Cavity|toothache \wedge catch)\\ &= \alpha\,\mathcal{P}(toothache \wedge catch|Cavity)\mathcal{P}(Cavity)\\ &= \alpha\,\mathcal{P}(toothache|Cavity)\mathcal{P}(catch|Cavity)\mathcal{P}(Cavity)\end{aligned}$$

This is an example of a **naive Bayes** model:

$$\mathcal{P}(Cause, Effect_1, \ldots, Effect_n) = \mathcal{P}(Cause)\prod_i \mathcal{P}(Effect_i|Cause)$$


Total number of parameters is *linear* in $n$

Wumpus World




- $P_{ij} = true$ iff [ $i, j$] contains a pit
- $B_{ij} = true$ iff [ $i, j$] is breezy
- Include only $B_{1,1}, B_{1,2}, B_{2,1}$ in the probability model

Let us take the example where logic does not help! The starting position is safe, but you can feel the breeze in the neighbors. It is not possible to calculate which of the three yellow rooms has a pit and which does not. Maybe there is only one in the middle. But if the others have a pit, we get the same outcome. Even if each of them has a pit, we get the same.

# AI #9

## Specifying the probability model

Specifying the probability model

The full joint distribution is $\mathcal{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$
Apply product rule: $\mathcal{P}(B_{1,1}, B_{1,2}, B_{2,1} \mid P_{1,1}, \ldots, P_{4,4})\mathcal{P}(P_{1,1}, \ldots, P_{4,4})$
(Do it this way to get $P(\mathit{Effect}|\mathit{Cause})$.)
First term: 1 if pits are adjacent to breezes, 0 otherwise
Second term: pits are placed randomly, probability 0.2 per square:

$$\mathcal{P}(P_{1,1}, \ldots, P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathcal{P}(P_{i,j}) = 0.2^n \times 0.8^{16-n}$$

for $n$ pits.

The entire maze consists of 16 rooms and there are three questionable fields where there can be a pit. This means 19 logical variables, i.e. nearly half a million cases.

The joint distribution can be rewritten into conditional distribution. In the first half of this, the rule of the game applies, if the breeze is next to the pit, it is okay; otherwise the probability is 0.

Let us assume that the pits are distributed randomly, i.e. there is a certain probability of a pit in a given room. (this should now be 20 percent). The fact that there are a $n$ pits total in the maze can be written based on the binomial theorem.

Observations and query

- We know the following facts:
  - $b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$
  - $known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$
- Query is $\mathcal{P}(P_{1,3}|known, b)$
- Define $Unknown = P_{ij}$s other than $P_{1,3}$ and $Known$
- For inference by enumeration, we have

$$\mathcal{P}(P_{1,3}|known, b) = \alpha \sum_{unknown} \mathcal{P}(P_{1,3}, unknown, known, b)$$

- Grows exponentially with number of squares!

We are sure that in three rooms there is no pit (*known*), and we have information about breeze in these rooms ($b$).
We want to know what is the chances of a pit in Room $(1, 3)$ in the case of public information (*known* and $b$).
We can call the variables of pits in the other rooms together as *unknown*.
Thus, the former conditional probability can be decomposed, it gives an exponential result which is almost unmanageable.

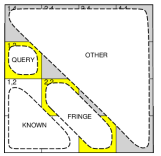
# AI #9

## Using conditional independence

Now conditional independence comes into play because the *known* fields are largely unrelated to the *unknowns*. Therefore, we treat separately those with which they are not really related (*other*)! And we have a *fringe*, which is close to the *known* rooms.

Using conditional independence

Basic insight: observations are conditionally independent of other hidden squares given neighbouring hidden squares




Define $Unknown = Fringe \cup Other$

- $\mathcal{P}(b|P_{1,3}, Known, Unknown) = \mathcal{P}(b|P_{1,3}, Known, Fringe)$

Manipulate query into a form where we can use this!

# AI #9

## Using conditional independence contd.

Using conditional independence contd.

$$\mathcal{P}(P_{1,3}|known, b) = \alpha \sum_{unknown} \mathcal{P}(P_{1,3}, unknown, known, b) =$$

$$\alpha \sum_{unknown} \mathcal{P}(b|P_{1,3}, known, unknown)\mathcal{P}(P_{1,3}, known, unknown) =$$

$$\alpha \sum_{fringe} \sum_{other} \mathcal{P}(b|known, P_{1,3}, fringe, other)\mathcal{P}(P_{1,3}, known, fringe, other) =$$

$$\alpha \sum_{fringe} \sum_{other} \mathcal{P}(b|known, P_{1,3}, fringe)\mathcal{P}(P_{1,3}, known, fringe, other) =$$

$$\alpha \sum_{fringe} \mathcal{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathcal{P}(P_{1,3}, known, fringe, other) =$$

$$\alpha \sum_{fringe} \mathcal{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathcal{P}(P_{1,3})P(known)P(fringe)P(other) =$$

To calculate the conditional probability we need to summarize over the *unknown* variables. The joint distribution behind the summa can rewritten into a product. We can separate the *unknown* variables into two. The *others* are conditionally independent from proposition $b$ in the case *known*, $P_{13}$ and *fringe*, so we can omit them. This enable us to take the first part out from the second sum. The position of pits is independent, so we can break the joint probability into products. (An error in the presentation, we need to use calligraphic $\mathcal{P}$s here.

# AI #9

## Using conditional independence contd.

Using conditional independence contd.

$$\alpha \sum_{fringe} \mathcal{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathcal{P}(P_{1,3}) P(known) P(fringe) P(other) =$$

$$\alpha\, P(known) \mathcal{P}(P_{1,3}) \sum_{fringe} \mathcal{P}(b|known, P_{1,3}, fringe) P(fringe) \sum_{other} P(other) =$$

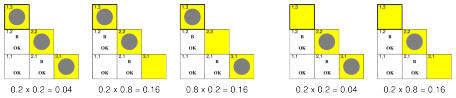
$$\alpha' \mathcal{P}(P_{1,3}) \sum_{fringe} \mathcal{P}(b|known, P_{1,3}, fringe) P(fringe)$$

We can pull out the constants from the sum.

# AI #9

## Using conditional independence contd.

Using conditional independence contd.




$$\mathcal{P}(P_{1,3}|known, b) = \alpha' \langle 0.2(0.04 + 0.16 + 0.16),\ 0.8(0.04 + 0.16) \rangle \approx \langle 0.31, 0.69 \rangle$$

$$\mathcal{P}(P_{2,2}|known, b) \approx \langle 0.86, 0.14 \rangle$$

Let us see the cases corresponding to the perceptions. In the first three cases, there is a pit at $(1, 3)$. Here in the first case there are two more pits (with chance $0.2 \times 0.2$), in the second case there is only one (with chance $0.2 \times 0.8$) and so on. Substituting these into the formula from before gives around 31 percent chance. There is 86 percent chance of pit the middle yellow room based a similar calculation. Therefore, it is worth continuing along the sides of the labyrinth.

Summary

Probability is a rigorous formalism for uncertain knowledge
**Joint probability distribution** specifies probability of every **atomic event**
Queries can be answered by summing over atomic events
For nontrivial domains, we must find a way to reduce the joint size
**Independence** and **conditional independence** provide the tools

Probability theory is part of mathematics with stable and accurate methods. This can be used to determine probabilities for uncertain environments.
If we have several statements, their combined distribution gives the probability of each individual event.
If we need the probability of a particular event, we need to look at the atomic events of it and summarize their probability.

This approach is often exponential, so the size must be reduced for manageability. Through independence and conditional independence, calculations can be simplified and performed separately.