independence and conditional probability

- Events A, B are independent if $P(A \cap B) = P(A)P(B)$
- This is completely different from disjointness. Disjointness means $P(A \cap B) = 0$. Disjoint events are not at all independent, cuz if A occurs then B cannot occur.
- A set of events $A_1, A_2 ... A_n$ are independent if and only if we have all possible k combinations being independent and all combinations follow the product rule $P(X_1..X_k) = P(X_1)..P(X_k)$. So for 3 events A, B, C, we must have $P(A,B) = P(A)P(B), P(B,C) = P(B)P(C), P(A,C) = P(A)P(A), P(A,B,C) = P(A)P(B)P(C)$ Note that the two pairs don't, in general, imply the three pair rule and vice versa. For independence, we need all these to be true.
- conditional probability deals with this problem: we are in an uncertain world, we have a probability distribution for some facts, we are not 100% certain of things. we learn new facts. **how should we update our beliefs about the uncertain world as we learn new facts?** that is the central question conditional probability helps us answer. this is usually a sequential process. we have some probability distribution today, we learn something as time t goes on. at each time of learning, we'd like to update our belief (belief state).
- definition of conditional probability of a given b. $p(a|b) = \frac{p(a \cap b)}{P(b)}$, unless $p(b) = 1$. The intuition for this is pebble world. each pebble is an element in the sample space (set of all possible outcomes). the top circle represents event A. the bottom circle is event B. given B occurs, $P(A|B) = \frac{P(A \cap B)}{P(B)}$. if all are equally likely, the conditional probability is $\frac{1}{4}$.

- some theorems:
  - $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$
  - $P(A_1, A_2 ... A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)...P(A_n|A_1...A_{n-1})$
  - $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ (Bayes' theorem).
- Bayes' theorem has an incredible number of applications and will keep showing up in many many places in machine learning and AI.
- "Conditioning is the soul of statistics" - Joe Blitzstein
- law of total probability:
  $P(B) = \sum_{k=1}^{n} P(B \cap A_k)$ where the A_i's are disjoint and the union of A_i's is the sample space.
    - This is only true when the conditions of disjoint and union are satisfied.

- This will be useful in a lot of places where $P(B)$ is hard to compute directly but each of the terms are easy to compute.
- so we've now seen two reasons why conditional probability is important: to update our beliefs as we learn more and to find absolute probability in certain situations if we know the conditionals using the total law of probability.
- an application of total probability that often comes up: $P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{P(T|D)P(D)}{P(T|D)P(D)+P(T|D^c)P(D^c)}$. notice the denominator is just an application of total probability.
- The coherency property of bayes rule is that you can have a certain distribution, learn something, update your belief, learn something later, update your belief again OR you can learn both things at once and update belief. If you do it correctly, both answer will turn out to be the same. we can use the definitions we have seen till now to show that this is true.
- do not confuse $P(A|B)$ and $P(B|A)$. applied to criminal cases, this is P(innocence | evidence), which is what you want to compute, **not** P(evidence | innocence). this is called prosecutor's fallacy. lot of people have been improperly imprisoned and their lives have been basically fully messed with because of prosecutor's fallacy.
- "Prior" means before we have evidence. "Posterior" means after we have evidence. Prior is $P(A)$, posterior is $P(A|B)$.
- if A and B are independent, then $P(A|B) = P(A)$.
- A and B are conditionally independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$.
  - In general, conditional independence does not imply independence.
  - Also, independence does not imply conditional independence.
- in chance games, if you're playing a game and they ask what should you do, what you should do is whatever makes it most likely for you to win. so whichever set of moves has the most $P(success)$ is what they're really asking for. In other words, they are asking what set of moves has the highest probability of success.
- for probability problems, if we can write up an easy computer simulation, and run it a very large number of times, we can at least start to get an idea of whether our answer is correct or not e.g. in the monty hall car-goat problem. so that's a useful tool to test out answers.
- Simpson's paradox is a general problem an example of which is that 1 doctor has a higher success rate than another doctor for any given surgery type. Even so, the doctor with the lower % of success rates for individual surgery type can have a higher overall success rate. this is because the worse doctor's % of different surgeries performed is vastly different from that of the doctor with better success rates for individual surgery categories.
- In probability terms $P(A|B, C)$ is called probability of A conditioned on B and C.
- Another example of Simpson's paradox: if you have jar A and jar B and jar A has a higher probability of the candy you like and if you have C and D and C has a higher probability of the same, if you combine A and C into 1 jar and B and D into anther, its not necessary, the first jar is the better one to pick from.
- Gambler's ruin: Lets say we have a game where one person has \$$i$ dollars and another has \$$n - i$. Lets say they play a gambling game where on each round, one of them wins \$1. whenever someone gets all \$$n$, that person wins. If the chance that player A wins a round is p and the chance that B wins a round is $q = 1 - p$, what is the chance that A wins the game? we could write a closed form expression for this that takes care of all possibilities. But what's important to point out here is that there is a natural recursive formulation here. Let p_i be the probability that A wins when A starts out with i of the n dollars. We can write the probability p_i as $p_i = p * p_{i+1} + (1 - p) * p_{i-1}$, because A might win or lose the first game. This is just using law of total probability. Notice how problem's like the gambler's ruin lend themselves to a recursive formulation.

- For the above problem, we can show that the probability of having a winner is 1 → the game goes on forever with 0 probability.